Journal of Exposure Science & Environmental Epidemiology (2018) 28:381–391
https://doi.org/10.1038/s41370-017-0009-6

ARTICLE

Predicting polycyclic aromatic hydrocarbons using a mass fraction approach in a geostatistical framework across North Carolina

Jeanette M. Reyes[1] · Heidi F. Hubbard[2] · Matthew A. Stiegel[3] · Joachim D. Pleil[4,5] · Marc L. Serre[5]

Received: 19 April 2017 / Revised: 6 October 2017 / Accepted: 27 October 2017 / Published online: 9 January 2018

Abstract

Currently in the United States there are no regulatory standards for ambient concentrations of polycyclic aromatic hydrocarbons (PAHs), a class of organic compounds with known carcinogenic species. As such, monitoring data are not routinely collected resulting in limited exposure mapping and epidemiologic studies. This work develops the log-mass fraction (LMF) Bayesian maximum entropy (BME) geostatistical prediction method used to predict the concentration of nine particle-bound PAHs across the US state of North Carolina. The LMF method develops a relationship between a relatively small number of collocated PAH and fine Particulate Matter (PM2.5) samples collected in 2005 and applies that relationship to a larger number of locations where PM2.5 is routinely monitored to more broadly estimate PAH concentrations across the state. Cross validation and mapping results indicate that by incorporating both PAH and PM2.5 data, the LMF BME method reduces mean squared error by 28.4% and produces more realistic spatial gradients compared to the traditional kriging approach based solely on observed PAH data. The LMF BME method efficiently creates PAH predictions in a PAH data sparse and PM2.5 data rich setting, opening the door for more expansive epidemiologic exposure assessments of ambient PAH.

Keywords Ambient exposures · PAHs · Bayesian maximum entropy · Mass fraction · Geostatistics

Introduction

Polycyclic aromatic hydrocarbons (PAHs) are a class of organic compounds containing two or more fused aromatic rings created through incomplete fuel combustion from a variety of sources including biofuel burning, wildfires, coal production, etc. [1, 2]. Several species of PAHs and their metabolites have been designated by the United States Environmental Protection Agency (US EPA) as being probable human carcinogens [3–6]. Currently the EPA only has PAH regulatory standards for drinking water and the National Institute for Occupational Safety and Health (NIOSH) has established occupational exposure limits to coal tar pitch volatiles [7]. International organizations and other countries have established ambient concentration guidelines for one of the more toxic PAHs, benzo(a)pyrene [8]. However, currently in the US there are no regulatory standards for ambient concentrations of PAHs. Compared to regulated ambient air pollutants, there are few epidemiologic studies that have utilized observed data or explored ambient exposures to different PAHs, which can be costly to measure [9, 10]. From a geostatistical perspective, limited ambient observed data have resulted in few studies creating maps of PAHs concentrations [11–14]. Others have used chemical transport models (CTMs) to predict PAH concentrations [8, 15, 16]. However, these studies are also limited in number. As a result, there is a gap in the literature exploring ambient PAH exposures and their associations

Electronic supplementary material The online version of this article (https://doi.org/10.1038/s41370-017-0009-6) contains supplementary material, which is available to authorized users.

✉ Marc L. Serre
marc_serre@unc.edu

[1] Oak Ridge Institute for Science and Education (ORISE) Research Participation Program, hosted at U.S. Environmental Protection Agency, Research Triangle Park, NC 27711, USA

[2] ICF International, Fairfax, VA, USA

[3] Duke University Medical Center, Durham, NC, USA

[4] National Exposure Research Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, NC, USA

[5] Department of Environmental Sciences and Engineering, University of North Carolina – Chapel Hill, 135 Dauer Drive, Chapel Hill, NC 27599-7431, USA

with various health endpoints. Short-term health effects include eye and skin irritation, nausea and vomiting while long-term health effects include increased risk to skin, lung and bladder cancer as well as cardiopulmonary mortality [7]. While many of these health effects are associated with either occupational exposures or drinking water exposures, the relationship between ambient concentrations of PAH to their associated health effects has not been well explored.

Both inside and outside the US there is a lack of consistent PAH observed monitoring outside of monitoring campaigns conducted for specific studies. In contrast to the data sparse environment of PAH observed data, particulate matter ≤2.5 μm in diameter (PM2.5) exists in a data rich environment with a vast, consistent, historical monitoring network across the US [17, 18]. Currently there are 16 EPA designated priority PAHs, nine of which are particle-bound [13]. Thus, a portion of PM2.5 may be particle-bound PAH. Currently, the US state of North Carolina has no maps displaying PAH concentrations using observed data. This study explores the relationship between PM2.5 and PAH and is an extension of previous work done by Allshouse et al. [13] that developed the log-mass fraction (LMF) Bayesian maximum entropy [19, 20] (BME) geostatistical method and applied this method to model the distribution of PAH near the World Trade Center after 11 September. The observational PAH data used in that work came from the analysis of 243 PM2.5 filters from four sites spanning approximately 200 days split between the sites near and around ground zero set up following 11 September.

In this work we analyze the PAH content of PM2.5 filters collected across the US state of North Carolina and implemented the LMF BME method to predict PAH concentration at unmonitored locations, creating the first maps of PAH across North Carolina for 2005 using observed data. Furthermore, we compare the LMF BME method with a simple linear regression (LR) BME method and more traditional geostatistical methods for the first time. Methods are evaluated through cross validation. Predictive maps are used to visualize the probability of exceeding PAH cutoff concentrations. Lastly, a comparison is performed between the LMF BME and other methods to learn how the relationship between PAH concentrations near wildfires may change for different prediction methods. These results provide a way for which a data sparse environment can be exploited in an efficient manner in conjunction with a data rich secondary data (e.g. PM2.5 data) environment in which the resulting relationship between the two can be applied to estimate concentrations of data sparse air pollutants elsewhere in a given domain. This cost-effective method, in terms of analyzing observed data, can be applied to other air pollution parameters that have not yet been previously mapped. This methodology opens the door for greater epidemiologic studies exploring the association between ambient concentrations of PAHs and various health endpoints.

Materials and methods

Observed PM2.5 and PAH data

Daily PM2.5 filters in North Carolina during 2004–2005 were collected as part of the monitoring effort needed to provide the data reported in the EPA's Air Quality Systems (AQS) data base [17]. Of the PM2.5 filters collected during this time period, we selected 84 filters collected in 2005 and analyzed them for the following nine species of PAHs: benz(a)anthracene, chrysene, benzo(b)fluoranthene, benzo(k)fluoranthene, benzo(e)pyrene, benzo(a)pyrene, indeno(1,2,3-cd)pyrene, benzo(g,h,i)perylene, dibenzo(a,h)anthracene and the summation of the nine PAH species called Total PAH. PM2.5 has units of $\mu g/m^3$ and PAH has units of ng/m^3.

The LR and LMF method

There are approximately 8,000 space/time locations for the state of North Carolina in 2005 where daily PM2.5 is observed and recorded in AQS. PAH was estimated at these locations using surrounding PAH and PM2.5 information. There were two different PAH estimation methods: (1) a LR method consisting of a regression created from paired PM2.5 and PAH in an estimation neighborhood in which PAH is then predicted at locations where PM2.5 is known and (2) a LMF method that assumes the ratio of PAH/PM2.5 is constant within an optimized estimation neighborhood in which PAH is then predicted by applying the ratio at locations where PM2.5 is known.

Let $\boldsymbol{p}_h$ be the PAH space/time locations where PAH was directly measured from a PM2.5 filter, where the location of the ith PAH measurement is denoted as $\boldsymbol{p}_i = (\boldsymbol{s}_i, t_i) \in \boldsymbol{p}_h$, where $\boldsymbol{s}_i$ is the spatial coordinate and t_i is the time coordinate. Let $\boldsymbol{p}_s$ be the space/time locations where PAH is estimated from PM2.5, in which the location of the jth individual estimate is denoted as $\boldsymbol{p}_j \in \boldsymbol{p}_s$.

The LR method is a simple linear regression of PAH with respect to PM2.5. This linear regression can be expressed at the $\boldsymbol{p}_i$ locations (where both PAH and PM2.5 are measured) as:

$$\ln(\mathrm{PAH}_i) = \beta_0 + \beta_1 \ln(\mathrm{PM2.5}_i) \quad (1)$$

The subsequent sections describe how the parameters β_0 and β_1 are estimated. The relationship in Eq. 1 can then be used to estimate PAH at the $\boldsymbol{p}_j$ locations (where only PM2.5

is measured) with the distribution

$$\ln(\mathrm{PAH}_j) \sim N\left(\hat{\beta}_{0,j} + \hat{\beta}_{1,j}\ln(\mathrm{PM2.5}_j), \sigma^2_{\mathrm{LR},j}\right) \quad (2)$$

where $\sigma^2_{\mathrm{LR},j}$ is the linear regression prediction variance.

The LMF method was defined by Allshouse et al. [13] as expressing the relation between PAH and PM2.5 at the $\boldsymbol{p}_i$ locations as

$$\mathrm{LMF}_i = \ln\left(\frac{\mathrm{PAH}_i}{\mathrm{PM2.5}_i}\right) \quad (3)$$

which can be rewritten as

$$\ln(\mathrm{PAH}_i) = \mathrm{LMF} + \ln(\mathrm{PM2.5}_i) \quad (4)$$

This approach is attractive because it uses only one parameter, namely LMF, to estimate PAH based on PM2.5, making the LMF more parsimonious and more localized around the estimation location of interest. At locations $\boldsymbol{p}_j$ where only PM2.5 is measured, the mean $\mu_{\mathrm{LMF},j}$ and variance $\sigma^2_{\mathrm{LMF},j}$ of LMF can be estimated as

$$\hat{\mu}_{\mathrm{LMF},j} = \sum_{i=1}^{N_{\mathrm{LMF}}(\boldsymbol{p}_j)} \mathrm{LMF}_i / N_{\mathrm{LMF}}(\boldsymbol{p}_j) \quad (5)$$

$$\hat{\sigma}^2_{\mathrm{LMF},j} = \sum_{i=1}^{N_{\mathrm{LMF}}(\boldsymbol{p}_j)} \left(\mathrm{LMF}_i - \mu_{\mathrm{LMF},j}\right)^2 / \left(N_{\mathrm{LMF}}(\boldsymbol{p}_j) - 1\right) \quad (6)$$

$N_{\mathrm{LMF}}(\boldsymbol{p}_j)$ is the number of LMF_i values closest to the space/time location $\boldsymbol{p}_j$. The optimization of $N_{\mathrm{LMF}}(\boldsymbol{p}_j)$ is described next.

The relationship in Eq. 3 can then be used to estimate PAH at locations $\boldsymbol{p}_j$ (where only PM2.5 is measured) as

$$\ln(\mathrm{PAH}_j) \sim N\left(\hat{\mu}_{\mathrm{LMF},j} + \ln(\mathrm{PM2.5}_j), \hat{\sigma}^2_{\mathrm{LMF},j}\right) \quad (7)$$

Equation 7 becomes a component described in the section "BME estimation methodology". In the limiting case, the LMF and LR methods are equivalent when $\beta_0 = \mu_{\mathrm{LMF},j}$ and $\beta_1 = 1$.

Neighborhood optimization

The parameters $N_{\mathrm{LR}}(\boldsymbol{p}_j)$ and $N_{\mathrm{LMF}}(\boldsymbol{p}_j)$ are optimized using the following methodology. For each of the 84 space/time PAH measurements, the measured PAH is excluded and re-estimated based on the collocated PM2.5 using either the LR method (Eq. 2) or the LMF method (Eq. 7) calibrated based on the paired PAH/PM2.5 values located in a local neighborhood of the excluded PAH value. This local neighborhood consists of the n, the number of observed data ranging from 1 to 84, closest pairs, where space/time proximity is defined based on the space/time distance $d = r + STM \times t$, such that r (km) is the spatial distance, t (day) is the time difference and STM (km/day) is the space/time metric. A given choice of the parameters n and STM creates 84 errors between the measured and re-estimated PAH value, from which a mean squared error (MSE) is calculated (Figure S1). This MSE is calculated for 75,600 different combinations of n and STM for each PAH and method (i.e. LR and LMF). For each PAH, the parameters $N_{LR}(\boldsymbol{p}_j)$ and $N_{LMF}(\boldsymbol{p}_j)$ are selected by choosing the n and STM that produced the lowest MSE. Due to the number of parameters of the LMF and LR methods (i.e. one parameter for LMR and two for LR), $N_{\mathrm{LR}}(\boldsymbol{p}_j) \geq 2$ while $N_{\mathrm{LMF}}(\boldsymbol{p}_j) \geq 1$. See Christakos and Serre (1999) for a more detailed explanation of the STM [21].

The values found for $N_{LR}(\boldsymbol{p}_j)$ and $N_{LMF}(\boldsymbol{p}_j)$ for each PAH are then applied to all $\boldsymbol{p}_j$ locations to estimate the corresponding PAH using Eqs. 2 and 7. The optimized n and STM were only used for the optimization of parameters $N_{LR}(p_j)$ and $N_{LMF}(p_j)$. These PAH estimates become input data in the BME estimation framework described next.

BME estimation methodology

BME provides a mathematically rigorous geostatistical space/time framework for the estimation of PAH at locations where neither PAH or PM2.5 are monitored [19, 20]. BME can incorporate information from multiple data sources and is implemented using the BMElib suite of functions in MATLAB™ [21, 22]. The buttress of BME has been detailed in other works [21, 23, 24], and can be summarized as performing the following steps: (1) gathering the general knowledge base (G-KB) and site-specific knowledge base (S-KB) characterizing the Space/Time Random Field (S/TRF) $X(\boldsymbol{p})$ representing a process at $\boldsymbol{p}$, (2) using the maximum entropy principle of information theory to process the G-KB in the form of a prior Probability Distribution Function (PDF) f_G, through a mean trend and an isotropic covariance model, (3) integrating S-KB in the form of a PDF f_S with and without measurement error using an epistemic *Bayesian* conditionalization rule (i.e. in this work, a priori information is updated on observed data) to create a posterior PDF f_K and (4) creating space/time predictions based on the analysis. All the code and data used for the analysis presented in this work are available from https://github.com/reyesjmUNC/ReyesEtAl_JESEE_PAH.

In this study, we use an S/TRF to describe the variability of PAH across North Carolina in 2005. In this work $\boldsymbol{x}_h$ are the observed PAH data and $f_S(\boldsymbol{x}_s)$ are obtained at the locations where PAH is estimated from PM2.5 observations through either the LR (Eq. 2) or the LMF (Eq. 7) method. We can then calculate x_k, the predicted daily PAH at the unmonitored location $\boldsymbol{p}_k$. More information about the prediction methodology can be found in the Supplementary Information.

Table 1 Optimized n closest observed data locations (as determined by the space/time metric) corresponding to the minimized mean squared error validation statistic calculated through the linear regression and mass fractions methods across the nine PAHs, with Total PAH being the summation. Bolded numbers indicate the lowest MSE for each PAH across the neighborhood optimization methods

PAH	Linear regression			Log-mass fraction		
	n	S/T metric (km/days)	MSE $(ng/m^3)^2$	n	S/T metric (km/days)	MSE $(ng/m^3)^2$
Benz(a)anthracene	14	0.891	1.128	5	0.839	**0.908**
Chrysene	7	0.600	0.979	5	0.839	**0.799**
Benzo(b)fluoranthene	7	0.863	1.358	5	0.899	**1.180**
Benzo(k)fluoranthene	14	0.895	1.375	5	0.842	**1.046**
Benzo(e)pyrene	14	0.895	1.006	2	0.868	**0.726**
Benzo(a)pyrene	14	0.895	**1.332**	5	0.899	1.417
Indeno(1,2,3-c,d)pyrene	14	0.891	0.892	2	0.868	**0.702**
Benzo(g,h,i)perylene	14	0.895	0.757	2	0.777	**0.742**
Dibenzo(a,h)anthracene	14	0.772	1.532	3	0.820	**1.115**
Total PAH	14	0.895	0.890	3	0.820	**0.675**

Leave-one-out cross validation (LOOCV) accuracy analysis

To assess the prediction accuracy of the LMF and LR methods, a LOOCV accuracy analysis is performed. For each monitoring station where observed PAH data exist, all observed data from a given station are removed one at a time and a BME prediction was conducted (without recalculating f_G) to obtain the BME predictions at that station using all the remaining observed and estimated data.

The difference between each mean predicted PAH $\tilde{x}_i$ and observed PAH value $\hat{x}_i$ is the prediction error, $e_i = \tilde{x}_i - \hat{x}_i$. The prediction accuracy was quantified based on prediction error statistics, which include the mean error (ME, ng/m^3), variance of errors (VE, $(ng/m^3)^2$), root mean squared error (RMSE, ng/m^3), MSE $(ng/m^3)^2$ and the squared of the Pearson correlation coefficient (r^2, unitless) calculated between observed and mean predicted values. LMF BME and LR BME predictions were then compared to kriging (i.e. predictions created only using observed PAH data) and cokriging (i.e. predictions created using both PAH and PM2.5 observed data).

Fire comparisons

Wildfires contribute to a sizable percentage of PAH emissions in the US [2]. The mean difference in PAH concentrations near known wildfire locations were estimated. PAH was estimated on a grid across North Carolina on days with observed PAH data. PAH was estimated using four different prediction methods: (1) kriging, (2) cokriging, (3) LR BME and (4) LMF BME. Fire data were obtained from the Federal Wildfire Fire Occurrence website [25]. All fires greater than or equal to one acre in North Carolina, Virginia, Tennessee and South Carolina were collected in 2005 on days for which PAH observed data were measured where the start and control date of the fires were known ($n = 213$). A two-tailed two-sampled t-test (assuming unequal variances) is calculated on the PAH predictions on a grid at a 5% significance level. The significance test is performed on all grid predictions within 100 km of known fire locations and all grid predictions outside of 100 km.

Results and discussion

Neighborhood optimization

A log-transformation of PM2.5 and PAH were taken due to the skewness of observed values. For each PAH and BME estimation method (i.e. LR and LMF), the optimal values of the parameters n and STM defining the estimation neighborhood were selected such that it minimized the MSE. Across each PAH the n closest observed data that optimized the estimation neighborhood was always smaller for the LMF method compared to the LR method (Table 1). Generally, we expect that the calibration of the LMF method requires less paired PAH/PM2.5 values because it is more parsimonious (i.e. has less parameters) than the LR method. Indeed, we find that the parameter n ranges from 2–5 for the LMF method whereas n ranges from 7–14 for the LR method. Benzo(g,h,i)perylene, indeno(1,2,3-c,d)pyrene and benzo(e)pyrene require $n = 2$ from the LMF method, the least amount of paired PAH/PM2.5 values across all PAHs. Seven out of nine PAHs in the LR method require $n = 14$.

Across each PAH and Total PAH, the minimized MSE was consistently lower for the LMF method than the LR method with the exception of benzo(a)pyrene. With these optimized neighborhoods, estimates were created by each method and each PAH is predicted across North Carolina using BME.

The PAH estimation neighborhood for the LMF method is smaller than the LR method. Out of the

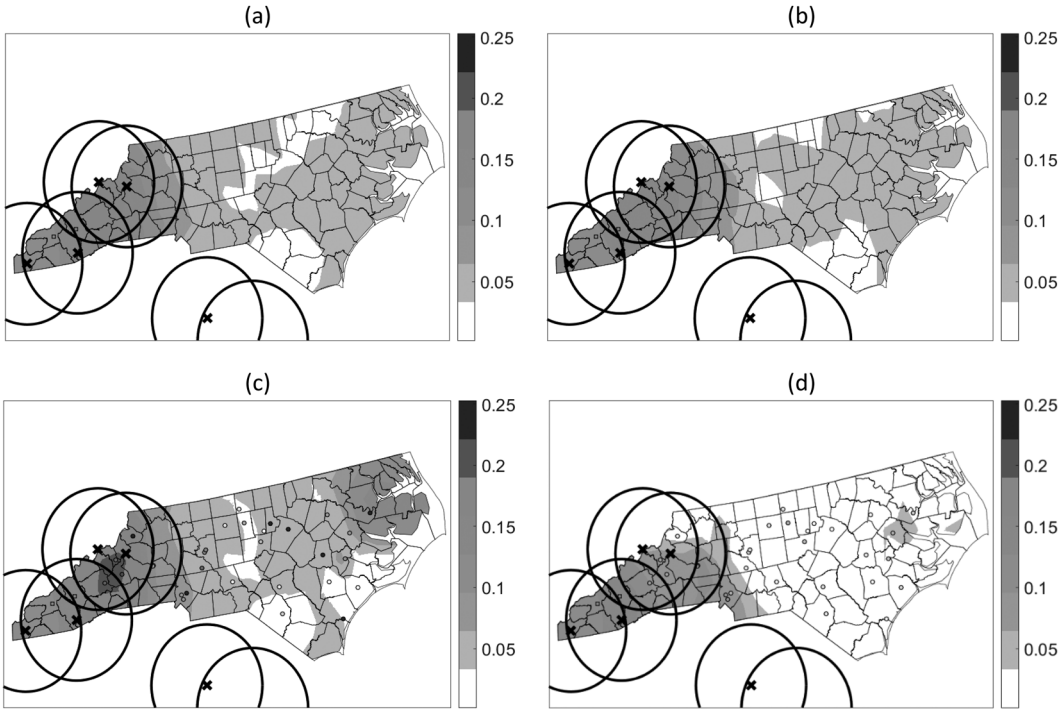

Fig. 1 Maps of mean benzo(b)fluoranthene concentration for North Carolina on April 16, 2005 across the four prediction methods: **(a)** kriging, **(b)** cokriging, **(c)** linear regression BME, **(d)** log-mass fraction BME (ng/m^3). Square markers indicate observed data, circle markers indicate PAH estimates, Xs mark known fires for that day with a 100 km buffer

previously mentioned studies, very little use observed PAH data and of those studies that do, most observed data come from short-lived monitoring campaigns. The results presented in this work utilize long-term, established PM2.5 regulatory monitoring sites. PM2.5 data is comparatively plentiful. Previously, data fusion methods have blended together multiple air pollutants that have different spatial supports [26, 27]. By developing a relationship between a few PAH observations and several PM2.5 observations, the door is opened to applying this relationship to a network with a large amount of publicly available data. Data sparse environments (e.g. PAH) can benefit from data rich secondary environments (e.g. PM2.5). However, for this relationship to be fully exploited, it must be constructed in such a manner that best utilizes the limited data set. That is, the relationship between PAH and PM2.5 must be parsimonious. The LMF method has only one parameter to be estimated, namely, LMF (Eq. 6). The minimum number of observed data needed to construct a PAH estimation is low with $N_{\mathrm{LMF}}(\boldsymbol{p}_j) \geq 1$. The PAH estimates created from the LMF method required less observed data than the LR method. We hypothesize that this increase in the number of parameters makes the LR model less parsimonious, requiring more paired PAH and PM2.5 to optimize the estimation neighborhood. The PAH paired data is then outside of the relevant air shed of estimation.

PAH prediction maps

This work created the first maps of predicted PAH in space/time across the US state of North Carolina for 2005 using observed data. Each of the nine PAHs was predicted on a grid across North Carolina every day observed PAH data were collected (41 days) in 2005 for the four prediction methods: kriging, cokriging, LR BME and LMF BME. Estimation method parameters can be found in Supplementary Information (Tables S1 and S2). Mean prediction maps of benzo(b)fluoranthene for the four methods are displayed across North Carolina on April 16, 2005 with observed and PAH estimates pictured (Fig. 1). The kriging map consistently predicts the highest PAH concentrations across the four methods at unmonitored locations with the least realistic gradient. Kriging has difficulty distinguishing between multiple PAH fronts and plumes. The minimal gradation is influenced by the sparse data. Predictions made far from observed data therefore had a large associated variance. The sparse data was only able to pick up the coarsest of PAH gradients. The cokriging map is visually similar to the kriging map. The cross-covariance relationship between PAH and PM2.5 contributed little to the cokriging predictions (Table S2). The prediction map becomes visibly different for the LR BME method. The gradient for the LR BME method falls more in line with a geographical pattern across the state. There is an increase in concentration in Eastern

North Carolina compared with the kriging and cokriging maps. The LMF BME prediction method produces the lowest PAH concentrations across large sections of the state. Across all four methods, the relatively highest concentrations were found in Western North Carolina and concentrations become increasingly more refined across methods. The LMF map is the only map to show two different PAH concentration fronts: one in the western part of the state and another separate front in Eastern North Carolina.

Few other studies have created maps of ambient PAH concentrations across a given area using geostatistical methods from observed data. These limited studies are due in part to the lack of observed data, much like the mapping scenario presented in this work and previous works [13]. One previous study fit a temporal trend comparing a few long-running PAH stations from the Great Lakes region of the US and a few stations across Europe. However, only a temporal trend was fit through a regression and a spatial interpolation was not conducted [28]. One of the few studies that created maps over a large area, displayed benzo(a) pyrene across Europe for 1990, 2001 and 2005 using a transport model [8]. Another study creating maps of PAH across Europe utilized kriging to estimated benzo(a)pyrene for 2012 using two different CTMs as data [15]. A study in Portugal used observed PAH data extracted from lichen and created maps using kriging [11, 12]. Land use regression models have also been used to estimate PAH [29, 30]. Few studies have investigated PAH bound to PM [31]. The closest study to the LR BME method presented in this work used a monitoring campaign along with personal monitors to analyze PAH from PM2.5 in which predictions were made at unmonitored locations using kriging in Kaohsiung city, Taiwan [14]. A regression model with a variety of explanatory variables was then applied to PM2.5 data to predict PAH. With only a handful of observed PAH data taken throughout the year, the LR BME and LMF BME method can create estimates with a corresponding uncertainty that was incorporated into the BME framework. Incorporating the PAH estimates added to the set of available data ultimately used for prediction allowing for increased spatial variation. Of the BME methods, LMF BME was superior in terms of visually distinguishing spatial variations of mean predicted PAH concentrations across the state.

Cross-validation

A LOOCV analysis was performed across 2005 using the four prediction methods. Summary statistics were calculated showing performance for Total PAH (Table 2). Cross validations statistics for all 9 PAHs can be found in Supplementary Information (Table S3). For Total PAH, ME decreases from kriging to LMF BME. ME is negative across each prediction method meaning that overall, the methods under-predicts observed Total PAH concentrations. ME is highest in magnitude for kriging and closest to zero for LMF BME. There is a 58.8% reduction in ME from LR BME to LMF BME. There is less variation in error from kriging to LMF BME as seen through a 26.7% reduction in VE from kriging to LMF BME. There is a consistent reduction in MSE across the four prediction methods. There is a 28.4% reduction in MSE from kriging to LMF BME. The correlation coefficient increases across methods. There is a 10.3% increase in r^2 from LR BME to LMF BME. The performance statistics from kriging are similar to cokriging. This echoes the results seen in the prediction maps. Traditional incorporation of PM2.5 as a co-pollutant through cokriging adds little to the predictive capacity of PAH. Incorporating PM2.5 with the BME methods showed more substantial improvements in the cross-validation statistics, with the best performance obtained through the LMF BME method.

Table 2 Leave-one-out cross validation statistics for Total PAH (summation of the nine PAHs) comparing observed and predicted concentrations across the four prediction methods for North Carolina in 2005

Statistic	Kriging	Cokriging	Linear regression BME	Log-mass fraction BME
ME (ng/m^3)	−0.145	−0.137	−0.102	−0.042
VE (ng/m^3)^2	0.806	0.782	0.764	0.591
RMSE (ng/m^3)	0.904	0.890	0.875	0.765
MSE (ng/m^3)^2	0.818	0.792	0.766	0.586
r^2 (unitless)	0.747	0.752	0.744	0.821

ME mean error, *VE* variance of error, *RMSE* root mean squared error, *MSE* mean squared error, r^2 Pearson correlation coefficient squared

The LMF BME method consistently outperformed the other comparison methods as seen visually through maps and through the LOOCV statistics. Of the four prediction methods, kriging performed the worst. Kriging predictions were driven exclusively by the observed data. Cokriging performed similarly to kriging. Cokriging is an intuitive choice for collocated, ambient, environmental parameters in a geostatistical setting. In the literature, to the best of our knowledge, cokriging has not been used to predict ambient PAH concentrations, making it an ideal candidate method to explore. In this work the cokriging cross-covariance is able to capture the relationship between PAH and PM2.5. However, as seen through predictive maps and through cross validation, the cokriging incorporation of PM2.5 contributes little in terms of predictive capacity. Linear regression is another intuitive choice with collocated data. The LR BME method shows a marked improvement

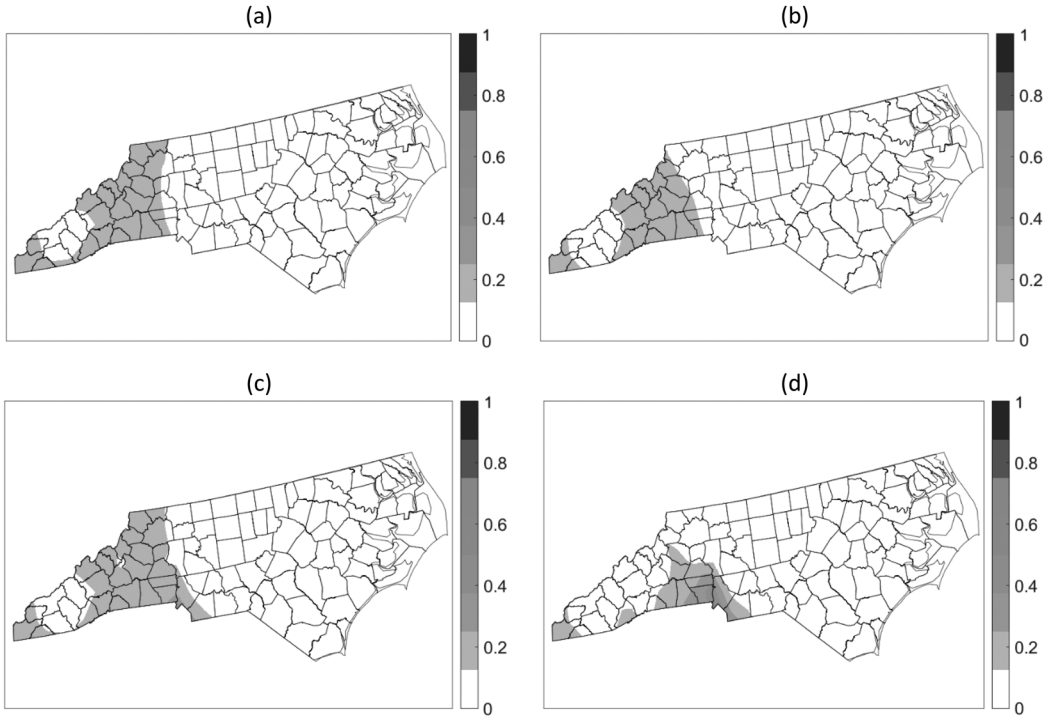

Fig. 2 Probability of annual benzo(a)pyrene exceeding 0.25 ng/m^3 across North Carolina in 2005 as predicted by **(a)** kriging, **(b)** cokriging, **(c)** linear regression BME and **(d)** log-mass fraction BME

visually and through estimation accuracy. The LR method is able to estimate PAH at PM2.5 space/time locations using an optimized neighborhood customized for each PAH. However, LR performed consistently worse than the LMF method. The LR method uses two parameters (i.e. β_0 and β_1) while the LMF method uses only one. We hypothesize that this difference in the number of parameters influences cross validation performance.

Probability of exceedance

In a geostatistical framework, predictions come in the form of a PDF with a corresponding mean and variance. With this PDF, the probability of exceeding a given value can be calculated. An annual benzo(a)pyrene concentration of 0.25 ng/m^3 has been suggested in the United Kingdom [8]. With this standard in mind, the probability of exceeding this cutoff was calculated for annual benzo(a)pyrene concentrations on a grid in North Carolina in 2005 for each prediction method by taking the mean and variance of daily benzo(a) pyrene predictions (Fig. 2). Overall PAH concentration decrease across methods, thus the probability of exceeding the 0.25 ng/m^3 cutoff in turn decreases from kriging to LMF BME. Across methods, the region of the state with the relatively highest probability of exceedance is maintained as Western North Carolina as well as the border with the US state of South Carolina. Across all prediction methods, the probability of exceedance remained relatively low with the maximum probability of exceedance remaining below 0.50. The area covered from increasing probabilities of exceedance increases across the four prediction methods (Table 3). The cokriging method had the lowest maximum probability of exceedance (i.e. 0.16) across the annual prediction locations with 54,432 km^2 having a probability of exceedance ≥0.15. We see the BME methods were better able to differentiate areas of high and low probabilities of exceedance. The LMF BME was best able to distinguish the maximum probability of exceedance. Neither the kriging nor the cokriging maps contain any area with a probability of exceedance ≥0.30. The LMF BME method has 2.5 times the area with ≥0.30 probability of exceedance compared to the LR BME method (i.e. 6,480 km^2 and 2,592 km^2, respectively). Through having more realistic ambient predictive gradients, the LMF BME method becomes an effective tool to identify areas of exceedance of different PAH concentrations.

Association with wildfires

The mean difference in PAH predictions (for the nine PAHs and Total PAH) as calculated through the four prediction methods was found through a two-sampled *t*-test comparing areas near (≤100 km) and far (>100 km) from known wildfire locations predicted across all days with observed data in

Table 3 Area (in km^2) covered corresponding to increasing probabilities of exceeding the average annual benzo(a)pyrene standard of 0.25 ng/m^3 among the prediction locations in and around North Carolina in 2005

Probability of exceedance	≥0.10	≥0.15	≥0.20	≥0.25	≥0.30
Kriging	162,000	119,232	5,184	0	0
Cokriging	143,856	54,432	0	0	0
Linear regression BME	164,592	139,968	36,288	10,368	2,592
Log-mass fraction BME	156,816	116,640	28,512	15,552	6,480

2005 (Table 4). For the LMF method, all nine PAHs and Total PAH showed a statistically significant difference between predictions near versus far from fires. For the LR method six PAHs and Total PAH showed a significant difference. For both kriging and cokriging four PAHs and Total PAH showed a significant difference greater than zero. Of those PAHs that showed a significant difference greater than zero, the LMF method had the largest differences across eight PAHs (i.e. benzo(g,h,i)perylene being the exception) and Total PAH. Known fire locations for April 16, 2005 are marked along with a 100 km radial buffer surrounding each location (Fig. 1). Across prediction methods, PAH concentrations are higher within/near these buffers. Indeed, benzo(b)fluoranthene (depicted in Fig. 1) was one of the four PAHs (along with Total PAH) that showed both a significant, positive difference across all four prediction methods.

This work investigates ambient concentrations of a set of particle-bound PAHs. Ambient concentrations alone cannot distinguish sources. However, there are PAH ratios associated with certain sources. The diagnostic ratio of indeno (1,2,3-c,d)pyrene/(indeno(1,2,3-c,d)pyrene + benzo(g,h,i) perylene) = 0.62 is associated with wood burning [8]. This ratio was calculated for March 5, 2005 data across all four prediction methods (Fig. 3). This day was chosen as one of the highest fire activity day for 2005, and thus, most likely to show an impact from fires. The ratio for kriging and cokriging remained under 0.62 across all prediction locations of the day. There is little variation of this ratio across the state for kriging and cokriging. This ratio increases and becomes closer in magnitude to 0.62 for the BME methods. There is more variation of this ratio for the LR BME method. We hypothesize that the LR BME method has better differentiation between PAH sources. LMF BME has the largest variation of the PAH diagnostic ratio, with the largest number of predictions near the 0.62 value. Both kriging and cokriging have ≤3.5% of prediction ratios for the day around 0.62 (i.e. 0.62 ± 0.05), LR BME has 5.6% of prediction ratios around 0.62 and LMF BME has 12% of prediction ratios around 0.62.

Gathering information about wildfire smoke has become increasingly important as the number of large wildfires have increased in recent years [32]. The chronic health effects of wildfire smoke for firefighters and the general population is currently lacking or sparse in the literature [33–35]. The LMF BME method was better able to distinguish higher significant differences in PAH concentrations near known fire locations compared with other prediction methods. Of the four prediction methods, the LMF BME method the only method that showed statistically significant, positive differences around areas with fires across all nine PAHs and Total PAH. Although each fire may have a different acreage burned even though the same buffer size was used for all the fires, the significance implies an association. Depending on the acreage burned from a fire, the type of vegetation burned and the duration of the fire, the smoke produced may be long lasting and may have long range transport. Smoke may have lingering effects past the control date of a fire. When the control date of fires is extended by one day, the kriging and cokriging methods have more PAHs with a statistically

Table 4 95% Confidence intervals comparing the mean difference (ng/m^3) in predicted PAH near (within 100 km) versus far (>100 km) from fires for each of the nine PAHs and Total PAH across the four prediction methods.

PAH	Kriging	Cokriging	Linear Regression BME	log-Mass Fraction BME
Benz(a)anthracene	(−4.94E-03, −2.17E-03)*	(−2.61E-03, −1.07E-05)*	(−1.26E-03, 1.07E-03)	(1.57E-03, 4.38E-03)*,#
Chrysene	(−6.67E-03, −3.53E-03)*	(−3.74E-03, −8.28E-04)*	(−9.40E-04, 1.67E-03)	(2.07E-03, 5.39E-03)*,#
Benzo(b)fluoranthene	(3.98E-03, 1.11E-02)*,#	(3.78E-03, 1.10E-02)*,#	(1.09E-02, 2.23E-02)*,#	(2.36E-02, 3.02E-02)*,#
Benzo(k)fluoranthene	(3.14E-03, 6.47E-03)*,#	(2.32E-03, 5.01E-03)*,#	(5.27E-03, 7.94E-03)*,#	(7.80E-03, 1.08E-02)*,#
Benzo(e)pyrene	(−2.92E-03, 2.23E-03)	(−3.17E-03, 1.71E-03)	(5.22E-03, 9.80E-03)*,#	(1.83E-02, 2.49E-02)*,#
Benzo(a)pyrene	(−3.83E-03, 1.84E-03)	(−6.24E-03, −8.42E-04)*	(2.23E-03, 1.37E-02)*,#	(5.14E-03, 1.02E-02)*,#
Indeno(1,2,3-c,d)pyrene	(1.87E-02, 3.05E-02)*,#	(1.73E-02, 2.87E-02)*,#	(2.04E-02, 3.24E-02)*,#	(4.79E-02, 6.11E-02)*,#
Benzo(g,h,i)perylene	(3.04E-02, 4.27E-02)*,#	(2.54E-02, 3.66E-02)*,#	(2.72E-02, 4.06E-02)*,#	(3.13E-02 4.06E-02)*,#
Dibenzo(a,h)anthracene	(−2.07E-02, −1.36E-02)*	(−1.71E-02, −1.07E-02)*	(−4.80E-03, 2.16E-03)	(1.90E-03, 8.77E-03)*,#
Total PAH	(2.28E-02, 6.75E-02)*,#	(2.13E-02, 6.57E-02)*,#	(6.29E-02, 1.03E-01)*,#	(1.72E-01, 2.30E-01)*,#

*Mean difference is statistically significant (p-value ≤ 0.05), #mean difference > 0

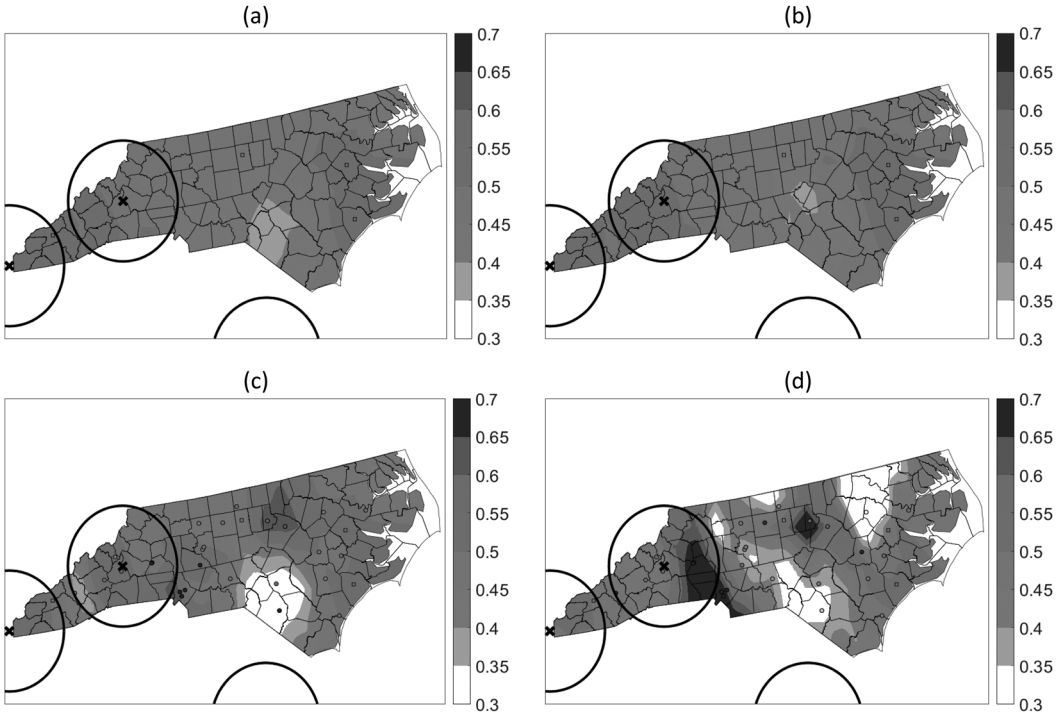

Fig. 3 Ratio of indeno(1,2,3-c,d)pyrene/(indeno(1,2,3-c,d)pyrene + benzo(g,h,i)perylene) on 5th March 2005 in North Carolina across the four prediction methods: **(a)** kriging, **(b)** cokriging, **(c)** linear regression BME, **(d)** log-Mass Fraction BME. Square markers indicate the ratio of observed data, circle markers indicate the ratio of PAH estimates, Xs mark known fires for that day with a 100 km buffer

significant increase in concentrations near versus far from fires (Tables S4-S7). Diagnostic ratios should not be used in isolation. However, when used along known fire locations, it can strengthen the association between PAH concentration and its known sources.

Overall contributions

The LMF BME method allows for straightforward predictions of PAHs to be used for exposure assessments. There are a plethora of studies exploring the association between ambient PM2.5 and various health endpoints [36–38]. However, there are far less studies that explore ambient PAH exposures and associated health effects. Occupational inhalation exposures and associated health outcomes including lung cancer have been more thoroughly investigated in comparison to ambient exposures [7]. Few studies have investigated chronic ambient concentrations of PAHs. Many of the epidemiologic studies that have been explored investigate respiratory illnesses such as lung cancer and pulmonary function [7, 15, 39]. However, these studies are small. The lack of long-term ambient concentrations to PAHs may be related to inadequate exposure data. Analyzing PM filters for specific PAHs can be very costly, making it difficult of obtain larger amounts of observed data needed for exposure assessment [9]. The LMF BME method allows for an efficient and cost-effective way to utilize minimal PAHs observed data. The LMF BME method can be easily utilized to fill in this clear gap in the literature. Tied with corresponding health data, ambient predictions calculated through the LMF BME method could be used to assign exposure. Health metrics can then be calculated from the exposures. This opens the door to investigate possible health endpoints as well as assigning risk.

This work created the first maps of ambient PAH concentration across the US state of North Carolina using observed data through the LMF BME geostatistical method. This method developed a relationship between paired PAH and PM2.5 data in a manner that is a parsimonious and cost-effective that can be utilized in a data sparse environment. The LMF BME method outperforms more traditionally used geostatistical methods and has the ability to elucidate a significant association between PAH predictions and known fire locations. The LMF BME method has the potential to be used to assign exposure in epidemiologic analyses to fill in the significant knowledge gap currently existing in the literature between ambient PAH exposures and potential health outcomes.

Disclaimer

The views expressed in this document are solely those of the authors and do not necessarily reflect those of the NIA,

NIOSH, NIEHS or the EPA. This article was reviewed in accordance with the policy of the National Exposure Research Laboratory, U.S. Environmental Protection Agency, and approved for publication.

Acknowledgements This research was supported in part by the National Institute on Aging (NIA) under award number R01AG033078, the National Institute of Occupational Safety and Health (NIOSH) under grant 2T42/OH-008673, the National Institute of Environmental Health Sciences (NIEHS) under grant T32ES007018 and an appointment of Jeanette Reyes to the Postdoctoral Research Program at the National Center for Environmental Assessment, Office of Research and Development, USEPA administered by the Oak Ridge Institute for Science and Education through Interagency Agreement No. DW-89-92298301 be- tween the U.S. Department of Energy and the USEPA.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Di-Toro DM, McGrath JA, Hansen DJ. Technical basis for narcotic chemicals and polycyclic aromatic hydrocarbon criteria.I. Water and tissue. Environ Toxicol Chem. 2000;19: 1951–70.
2. Zhang Y, Tao S. Global atmospheric emission inventory of polycyclic aromatic hydrocarbons (PAHs) for 2004. Atmos Environ. 2009;43:812–9.
3. Bocskay KA, Tang D, Orjuela MA, Liu X, Warburton DP, Perera FP. Chromosomal aberrations in cord blood are associated with prenatal exposure to carcinogenic polycyclic aromatic hydrocarbons. Caner Epidemiol Biomark Prev. 2005;14:506–11.
4. Wolff MS, Teitelbaum SL, Lioy PJ, Santella RM, Wang RY, Jones RL, et al. Exposures among pregnant women near the World Trade Center Site on 11 September 2001. Environ Health Perspect. 2005;113:739–48.
5. Menzie CA, Potocki BB, Santodonato J. Exposure to carinogenic PAHs in the environment. Environ Sci Technol. 1992;26:1278–84.
6. EPA U. Polycyclic aromatic hydrocarbons (PAHs) — EPA fact sheet. 2013 https://www.epa.gov/sites/production/files/2014-03/documents/pahs_factsheet_cdc_2013.pdf.
7. Kim K-H, Jahan SA, Kabir E, Brown RJC. A review of airborne polycyclic aromatic hydrocarbons (PAHs) and their human health effects. Environ Int. 2013;60:71–80.
8. Ravindra K, Sokhi R, van Grieken R. Atmospheric polycyclic aromatic hydrocarbons: source attribution, emission factors and regulation. Atmos Environ. 2008;42:2895–921.
9. Pleil JD, Vette AF, Rappaport SM. Assaying particle-bound polycyclic aromatic hydrocarbons from archived PM2.5 filters. J Chromatogr A. 2004;1033:9–17.
10. Abdel-Shafy HI, Mansour MSM. A review on polycyclic aromatic hydrocarbons: source, environmental impact, effect on human health and remediation. Egypt J Pet. 2015;25:107–23.
11. Augusto S, Máguas C, Matos J, Pereira MJ, Soares A, Branquinho C. Spatial modeling of PAHs in lichens for fingerprinting of multisource atmospheric pollution. Environ Sci Technol. 2009;43:7762–9.
12. Ribeiro MC, Pinho P, Llop E, Branquinho C, Pereira MJ. Geostatistical uncertainty of assessing air quality using high-spatial-resolution lichen data: a health study in the urban area of Sines, Portugal. Sci Total Environ. 2015;562:740–50.
13. Allshouse WB, Pleil JD, Rappaport SM, Serre ML. Mass fraction spatiotemporal geostatistics and its application to map atmospheric polycyclic aromatic hydrocarbons after 9/11. Stoch Environ Res Risk Assess. 2009;23:1213–23.
14. Lee C-L, Huang H-C, Wang C-C, Sheu C-C, Wu C-C, Leung S-Y, et al. A new grid-scale model simulating the spatiotemporal distribution of PM2.5-PAHs for exposure assessment. J Hazard Mater. 2016;314:286–94.
15. Guerreiro CBB, Horálek J, de Leeuw F, Couvidat F. Benzo(a) pyrene in Europe: ambient air concentrations, population exposure and health effects. Environ Pollut. 2016;214:657–67.
16. Zhang J, Wang P, Li J, Mendola P, Sherman S, Ying Q. Estimating population exposure to ambient polycyclic aromatic hydrocarbon in the United States - Part II: Source apportionment and cancer risk assessment. Environ Int. 2017;99:263–74.
17. US EPA. Air Quality System (AQS). 2011. http://www.epa.gov/ttn/airs/airsaqs/ (Accessed 11 Sep 2010).
18. Liu B, Xue Z, Zhu X, Jia C. Long-term trends (1990–2014), health risks, and sources of atmospheric polycyclic aromatic hydrocarbons (PAHs) in the U.S. Environ Pollut. 2017;220:1171–9.
19. Christakos G. M*odern spatiotemporal geostatistics*. New York, NY: Oxford University Press; 2000.
20. Christakos G, Serre ML, Kovitz JL. BME representation of particulate matter distributions in the state of California on the basis of uncertain measurements. J Geophys Res. 2001;106:9717–31.
21. Serre ML, Christakos G. Modern geostatistics: computational BME analysis in the light of uncertain physical knowledge - the Equus Beds study. Stoch Environ Res Risk Assess. 1999;13:1–26.
22. Christakos G, Bogaert P, Serre ML. Temporal GIS: Advanced functions for field-based applications. New York, NY: Springer; 2002.
23. Berntsen J, Espelid TO, Genz A. An adaptive algorithm for the approximate calculation of multiple integrals. ACM Trans Math Softw. 1991;17:437–51.
24. Genz A. Statistics applications of subregion adaptive multiple numerical integration. In: Espelid T.O., Genz A. (eds) Numerical Integration. NATO ASI Series (Series C: Mathematical and Physical Sciences), Netherlands: Springer. 1992;357:267–80.
25. United States Geological Survey. Federal Wildland Fire Occurrence Data. 2016. http://wildfire.cr.usgs.gov/firehistory/data.html.
26. Fuentes M, Raftery AE. Model evaluation and spatial interpolation by Bayesian combination of observations with outputs from numerical models. Biometrics. 2005;61:36–45.
27. Xu Y, Serre ML, Reyes JM, Vizuete W. Bayesian Maximum Entropy integration of ozone observations and model predictions: A national application. Environ Sci Technol. 2016;50:4393–400.
28. Liu LY, Kukucka P, Venier M, Salamova A, Klanova J, Hites RA. Differences in spatiotemporal variations of atmospheric PAH levels between North America and Europe: data from two air monitoring projects. Environ Int. 2013;64:48–55.
29. Jedynska A, Hoek G, Wang M, Eeftens M, Cyrys J, Keuken M, et al. Development of land use regression models for elemental, organic carbon, PAH, and hopanes/steranes in 10 ESCAPE/TRANSPHORM European study areas. Environ Sci Technol. 2014;48:14435–44.
30. Noth EM, Hammond SK, Biging GS, Tager IB. A spatial-temporal regression model to predict daily outdoor residential PAH concentrations in an epidemiologic study in Fresno, CA. Atmos Environ. 2011;45:2394–403.
31. Sosa BS, Porta A, Colman Lerner JE, Banda Noriega R, Massolo L. Human health risk due to variations in PM10-PM2.5 and associated PAHs levels. Atmos Environ. 2017;160:27–35.

32. Dennison PE, Brewer SC, Arnold JD, Moritz MA. Large wildfire trends in the western United States, 1984–2011. Geophys Res Lett. 2014;41:2928–33.
33. Adetona O, Reinhardt TE, Domitrovich J, Adetona AM, Kleinman MT, Ottmar RD, et al. Review of the health effects of wildland fire smoke on wildland firefighters and the public. Inhal Toxicol. 2016;28:95–139.
34. Reid CE, Brauer M, Johnston FH, Jerrett M, Balmes JR, Elliott CT. Critical review of health impacts of wildfire smoke exposure. Environ Health Perspect. 2016;124:1334–43.
35. Reid CE, Jerrett M, Tager IB, Petersen ML, Mann JK, Balmes JR. Differential respiratory health effects from the 2008 northern California wildfires: A spatiotemporal approach. Environ Res. 2016;150:227–35.
36. Pope CA, Burnett RT, Thurston GD, Thun MJ, Calle EE, Krewski D, et al. Cardiovascular mortality and long-term exposure to particulate air pollution: epidemiological evidence of general pathophysiological pathways of disease. Circulation. 2004;109:71–7.
37. Beelen R, Hoek G, Fischer P, van den Brandt PA, Brunekreef B. Estimated long-term outdoor air pollution concentrations in a cohort study. Atmos Environ. 2007;41:1343–58.
38. Krewski D, Jerrett M, Burnett RT, Ma R, Hughes E, Shi Y, et al. Extended follow-up and spatial analysis of the American Cancer Society study linking particulate air pollution and mortality. Respir Rep Heal Eff Inst. 2009;140:5–114.
39. Padula AM, Balmes JR, Eisen EA, Mann J, Noth EM, Lurmann FW, et al. Ambient polycyclic aromatic hydrocarbons and pulmonary function in children. J Expo Sci Environ Epidemiol. 2015;25:295–302.