Environmental Science & Technology

Article

Cite This: Environ. Sci. Technol. 2019, 53, 1413−1421

pubs.acs.org/est

Performance of Prediction Algorithms for Modeling Outdoor Air Pollution Spatial Surfaces

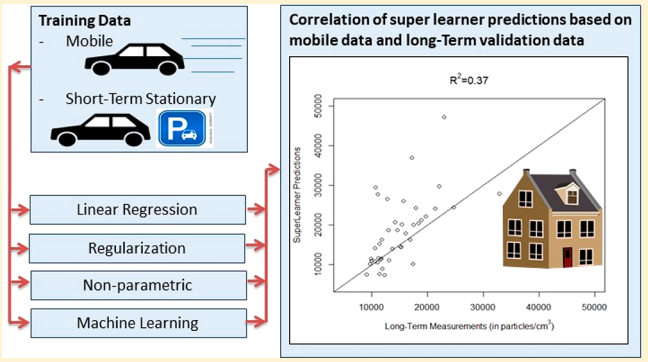
Jules Kerckhoffs,*[,†] Gerard Hoek,[†] Lützen Portengen,[†] Bert Brunekreef,[†,‡] and Roel C. H. Vermeulen[†,‡]

[†]Institute for Risk Assessment Sciences (IRAS), Division of Environmental Epidemiology, Utrecht University, 3584 CK Utrecht, The Netherlands

[‡]Julius Center for Health Sciences and Primary Care, University Medical Center, University of Utrecht, 358 CK Utrecht, The Netherlands

Supporting Information

ABSTRACT: Land use regression (LUR) models for air pollutants are often developed using multiple linear regression techniques. However, in the past decade linear (stepwise) regression methods have been criticized for their lack of flexibility, their ignorance of potential interaction between predictors, and their limited ability to incorporate highly correlated predictors. We used two training sets of ultrafine particles (UFP) data (mobile measurements (8200 segments, 25 s monitoring per segment), and short-term stationary measurements (368 sites, 3 × 30 min per site)) to evaluate different modeling approaches to estimate long-term UFP concentrations by estimating precision and bias based on an independent external data set (42 sites, average of three 24-h measurements). Higher training data R^2 did not equate to higher test R^2 for the external long-term average exposure estimates, making the argument that external validation data are critical to compare model performance. Machine learning algorithms trained on mobile measurements explained only 38−47% of external UFP concentrations, whereas multivariable methods like stepwise regression and elastic net explained 56−62%. Some machine learning algorithms (bagging, random forest) trained on short-term measurements explained modestly more variability of external UFP concentrations compared to multiple linear regression and regularized regression techniques. In conclusion, differences in predictive ability of algorithms depend on the type of training data and are generally modest.



INTRODUCTION

A common approach to predict spatial concentration levels of air pollutants is to use land-use regression (LUR) models. LUR modeling is an empirical technique with the measured concentration of a pollutant as dependent variable and potential predictors such as road type, traffic count, elevation, and land cover as independent variables in a multiple regression model.[1] Model structures and the criteria for variables to be included in the model (variable selection) used in LUR modeling, though, differ between studies.

In a review of Hoek,[2] it was found that most LUR studies use linear regression techniques that rely on least-squares or maximum likelihood fitting to develop prediction models. These models are simple, fast, and often provide interpretable coefficients of predictors.[3] Forward, backward, and best-subsets automatic selection methods are used in most of these settings. Often linear regression is applied with restrictions, typically allowing only slopes that conform with physical reality such as positive traffic intensity slopes. However, in the past decade linear (stepwise) regression methods have been criticized for the assumed linearity of predictor pollution relationships,[3] limited inclusion of potential interactions, and incorporating highly correlated predictors.[4] Next, standard linear regression methods are prone to overfitting, especially when few training sites are used for model development along with a large number of predictor variables.[5,6] Linear regression therefore may not identify the optimal model.

To overcome these issues, new modeling techniques have been introduced into air pollution epidemiology.[7] For example, nonlinear relationships can be obtained with general additive[8,9] or kernel based models.[10] Furthermore, machine learning techniques (such as neural networks[11] and random forests[12]) offer possibilities to create spatial models of air pollutants by learning the underlying relationships in a training data set, without any predefined constrictions. In a review by Bellinger et al.[7] on applications in epidemiology and references therein,

Received: October 26, 2018
Revised: December 20, 2018
Accepted: January 4, 2019
Published: January 4, 2019

Table 1. Overview of UFP Monitoring Data Used for Model Development and Validation

data	reference	cities[a]	duration	sites[b]	instrument	year
MUSiCShort-Term	Montagne et al. (2015)	Amsterdam and Rotterdam	3 × 30 min	128	CPC 3007	Winter, Spring and Summer 2013
MUSiC Mobile	Kerckhoffs et al. (2016)	Amsterdam and Rotterdam	∼20 s	2964	CPC 3007	Winter and Spring 2013
EXPOsOMICS Short-term	Van Nunen et al. (2017)	Amsterdam, Maastricht and Utrecht	3 × 30 min	240	CPC 3007	Winter, Spring and Summer 2014/2015
EXPOsOMICS Mobile	Kerckhoffs et al. (2017)	Amsterdam, Maastricht and Utrecht	∼25 s	5236	CPC 3007	Winter, Spring and Summer 2014/2015
EXPOsOMICS Home Outdoor	Van Nunen et al. (2017)	Amsterdam and Utrecht	3 × 24 h	42	DiscMini	Winter, Spring and Summer 2014/2015

[a]Indication of size (inhabitants) of cities: Amsterdam: 820 000, Maastricht: 120 000, Rotterdam: 620 000, Utrecht: 330 000. [b]For mobile measurements this refers to the number of road segments.

it was concluded that machine learning can be an effective tool for building accurate predictive models,[13,14] especially with increasing size and complexity of data sets.[7]

Comparisons between the performance of modeling techniques are scarce. Reid et al.[15] compared 11 modeling techniques for predicting spatiotemporal variability of $PM_{2,5}$ concentrations during wildfire events in California. General boosting, random forest, and support vector machines performed better than standard linear regression modeling, lasso, and elastic net. Van den Bossche et al.[16] reported no significant differences between linear regression, LASSO, and support vector regression (SVR) to create LUR models for black carbon in a mobile monitoring campaign. Random forest performed better than stepwise selection to develop prediction models for elements in PM measured at 24 sites in Cincinnati, Ohio.[31] Weichenthal et al.[10] found minor differences between a machine learning method (KRLS) and linear stepwise regression for mobile monitoring data in Montreal. KRLS had a higher training model R^2 compared to linear regression, but differences decreased when external data were used to compare predictions.

In previous studies we reported LUR models for ultrafine particles (UFP) based on mobile monitoring[17,18] and short-term stationary measurements[19,20] developed with a supervised forward linear regression approach. Now, we use the UFP measurements and GIS predictors from both data sets to evaluate different modeling approaches by estimating precision and bias based on an external data set. Since the main goal of the models is to predict concentrations for use in epidemiological studies of chronic diseases, we used long-term measurements as the test set. These measurements were completely independent of the mobile and short-term measurements. As each modeling algorithm is likely not optimal, we also explored if a combination of models (stacking) could increase predictive performance. Stacking methods are commonly used to gain predictive power and to average out biases as compared to individual models.[4] In additional analysis, we therefore used an algorithm (Super Learner[21]) that works by stacking results from a wide range of different modeling techniques.

MATERIALS AND METHODS

Data Description. We combined data from two measurement campaigns: MUSiC (Measurements of UFP and Soot in Cities) and EXPOsOMICS (Combining Exposure and Omics data). The MUSiC campaign was conducted in 2013 and entailed both mobile and short-term air pollution monitoring. The EXPOsOMICS campaign (2014−2015) had a similar monitoring setup, but also included three repeated 24 h home outdoor measurements in addition to the mobile and short-term stationary measurements. An overview of all data, also showing the spatial and temporal resolution, is given in Table 1. Measurements and models from the mobile MUSiC campaign,[18] the short-term stationary MUSiC campaign,[19] the mobile EXPOSOMICS campaign,[17] and the short-term stationary EXPOSOMICS campaign[20] have been described in detail elsewhere.

In summary, in both campaigns short-term stationary and on-road UFP measurements were made using a condensation particle counter (TSI, CPC 3007) attached to an electric vehicle (REVA, Mahindra Reva Electric Vehicles Pvt. Ltd., Bangalore, India). UFP concentrations were recorded every second alongside a global positioning unit (GPS, Garmin eTrex Vista). For better comparability between sites, we avoided rush hour traffic and started sampling after 9:15 and stopped before 16:00 with about eight short-term sites sampled each day. These sites were monitored for 30 min on 3 different days, spread over three seasons. Mobile measurements were collected when driving from one stationary site to the next and then averaged over the road segment; the measurement was conducted on and all days the road segment was sampled on. A reference site with the same equipment as the electric vehicle was used to temporally correct all measurements. In both campaigns we used the difference method, which first calculates the overall mean concentration at the reference site over the entire campaign. Next, using the data at the reference site, the average reference concentration of 30 min around the time of sampling was subtracted from the overall mean. This difference was used to adjust the measured concentration at the sampling locations. Sites were selected with varying traffic intensities and land use characteristics. About 40% were traffic sites (>10 000 vehicles a day), 40% were urban background sites, and some were industrial sites and sites near urban green or rivers.

In this paper we combined the data sets from both campaigns, creating one mobile data set and one short-term stationary data set. Independent home outdoor measurements were collected during the EXPOsOMICS campaign. Forty-two outdoor home locations where repeatedly sampled for 24 h during three different seasons. Measurements were averaged over all available seasons (at least two) after adjustment for temporal variation at a reference site. Measurements were performed with DiscMinis, which were found to have a ratio of almost 1 compared with colocated CPC 3007 measurements on three occasions in the measurement campaign.[20]

DOI: 10.1021/acs.est.8b06038
Environ. Sci. Technol. 2019, 53, 1413−1421

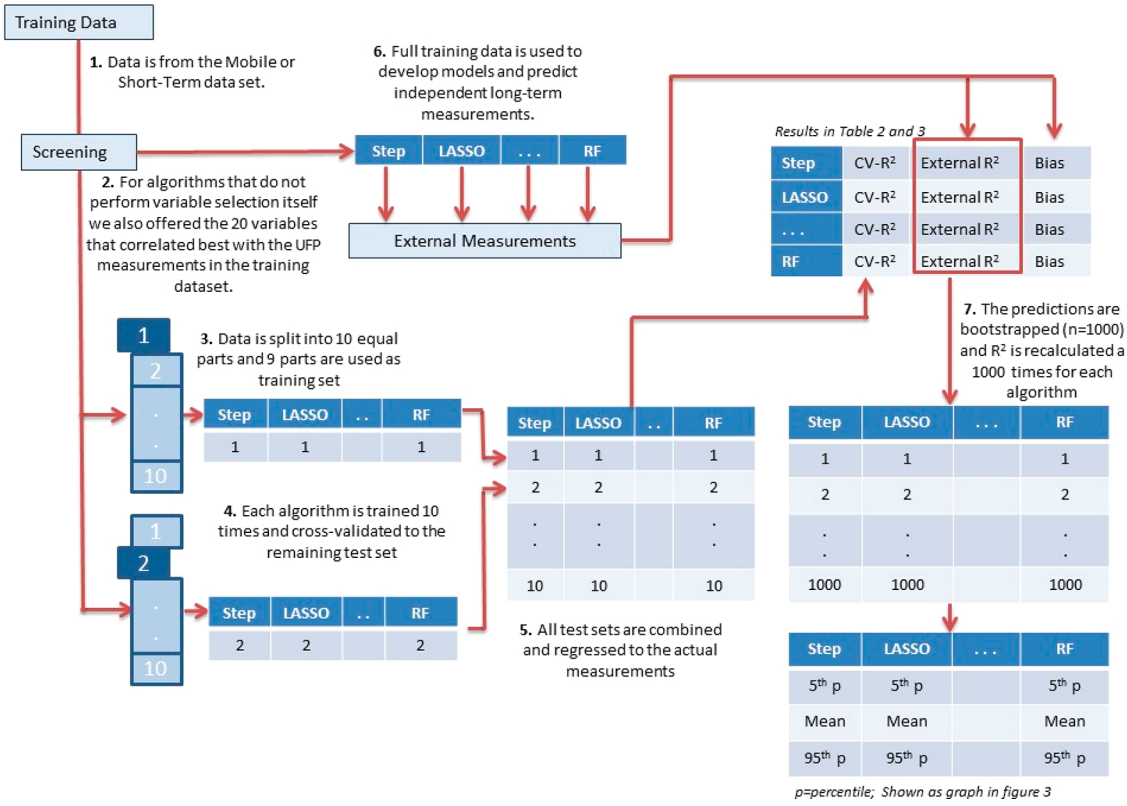


Figure 1. Flow diagram of analysis, with "STEP", "LASSO", and "RF" as example algorithms (stepwise regression, LASSO, and random forest).

Statistical Analysis. Land use regression models were developed for mean UFP concentrations per street segment in the mobile and per site for the short-term stationary data set separately. All model techniques were offered the same set of 72 GIS-derived predictors, shown in Appendix S1. Because of the inoperable predictors related to inverse distance to road in the mobile data set the number of predictors is 70. These include a range of traffic variables, including traffic intensity and road length variables (in 50−1000 m buffers), land use variables (e.g., port, industry, urban green, airports), and population density variables.

Comparing Prediction Algorithms. We compared several different model algorithms for creating LUR models, schematically shown in the abstract graphic. Stepwise regression is by far the most widely used prediction algorithm for developing land use regression models for air pollution.[22,23] We selected different forms of stepwise regression (Forward, Backward, Both, and a Supervised Method).[6,21] Note that "supervised" in this case means a stepwise procedure that is customized and not the fact that the input and output variables are known, a definition that is more common in machine learning terminology. We keep the term supervised as an option for stepwise regression for its extensive use in the literature. Supervised regression approaches usually constrain slopes of predictors to a predefined direction. Related to stepwise regression is the Deletion Substitution Addition (DSA) algorithm.[6,24] The difference with stepwise regression is that model complexity is determined by cross-validation, and the final model is selected by minimizing the residual mean squared error (RMSE) using cross-validation.[24] To deal with possible correlation of predictors, we selected regularization techniques (LASSO, RIDGE, and Elastic Net). They can deal with correlated predictors by imposing a penalty on the absolute size of regression coefficients. Nonlinear methods are represented in our analyses by a log-transformed generalized linear model (GLM),[25] multivariate adaptive regression splines (MARS), and generalized additive modeling (GAM).[8,9,26] More flexibility in the association between predictors and pollution can be obtained by data mining and machine learning methods. We used kernels (Kernel-Based Regularized Least Squares (KRLS)[10]), neural networks,[11] support vectors (support vector machines (SVM)[27]), or combining weak classifiers/trees in an ensemble (random forests,[12] boosting, and bagging). Algorithms are described more extensively in the Supporting Information (SI.2.1−SI.2.11).

In an additional analysis we combined all models in a "stacking" method called Super Learner, which selects and combines different prediction algorithms. Each algorithm is assigned a weight, which is based on the difference between the training model R^2 and cross validated R^2. The super learner then uses a weighted combination of the algorithms which were assigned the highest weight. The supervised stepwise approach, DSA and KRLS, is not supported in the super learner algorithm and is therefore not included in this analysis.

We present the cross-validated training model R^2, and the R^2 value and bias of model predictions compared to the external measurements (Figure 1). We used 10-fold cross-validation with repeated random selection of training and test sets. In each cross-validation set models were redeveloped, creating 10 models for each algorithm. Predictions were combined and then regressed to the actual measurements. Training model R^2 (without cross-validation) can be found in the Supporting Information. To investigate the precision of our external test set regression coefficients, we bootstrapped (n = 1000) all external test set model predictions and show the median, IQR, and 90% probability interval of the estimated R^2 values.

DOI: 10.1021/acs.est.8b06038
Environ. Sci. Technol. 2019, 53, 1413−1421

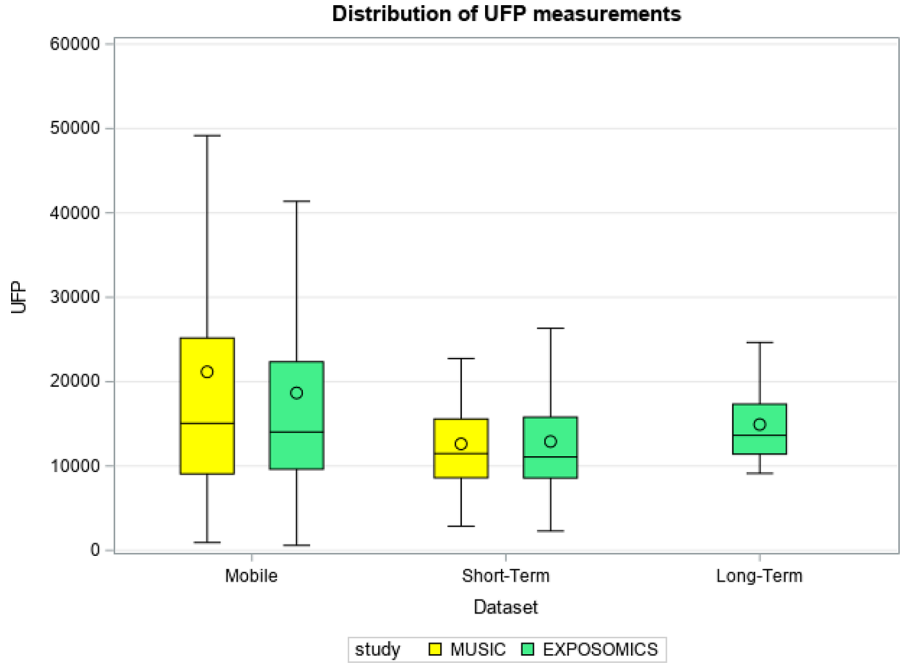


Figure 2. Distribution of site averaged concentrations in the mobile and short-term training data sets and the external long-term validation data set (in particles/cm^3). Distributions extent from 5th to 95th percentile with the IQR in the solid box and the circle representing the mean.

Specifically, we randomly selected 42 sites from the original 42 sites with replacement, thus allowing sites to be included multiple times in the bootstrap sample. This procedure has been shown to give correct estimates of the precision of the R^2.[4] We also ranked all models according to their median R^2 values based on the 1000 bootstrap samples in sensitivity analyses. Bias was calculated as the mean difference between the external measurement and the predicted concentration at the sampling location.

Variable Selection. When a certain algorithm does not perform variable subset selection, exposure assessment studies often use screening of variables. This could be a maximum number of variables to be able to enter the model,[6,24] stepwise regression screening[9,10] or a preselection of variables.[8,28]

For algorithms that do not perform variable selection we also offered the 20 variables that correlated best with the UFP measurements in the training data set. We selected 20 as the cutoff because it restricts the number of variables considerably (by a third), and algorithms that do variable selection never included more than 20 variables in their model (Table S3). For sensitivity analysis (Tables S4 and S5), we also offered other amounts of variables to the models, one based on the absolute value of the Pearson correlation coefficient (variable included when it correlates at least 0.1 to the UFP measurement) and by first using a stepwise regression procedure (variables from the customized stepwise regression (SI.2.1)). The overview of GIS-derived predictors (Table S1) shows which variables pass a specific screening method.

RESULTS

Distribution. Figure 2 and Table S2 show the distribution of UFP measurements collected while driving (n = 8200) and during the short-term stationary (n = 368) and long-term stationary measurements (n = 42). Mean concentration collected while driving is about 50% higher than measurements collected on short-term measurement sites. Especially the frequency at which high concentrations occur is higher while driving (90th percentile of mobile measurements is almost double the 90th percentile of stationary measurements, whereas the 10th percentiles are similar in the mobile and short-term data sets). The external validation set had a similar variation of UFP data as the short-term data, but the average of UFP was slightly higher (about 2000 particles/cm^3).

Comparing Prediction Algorithms. Table 2 (mobile data) and Table 3 (short-term stationary data) shows the performance of all individual algorithms (with default settings used in the super learner algorithm). Supporting Information SI.2.1−SI.2.11 provides scatterplots of measured versus predicted concentrations for selected algorithms, as well the settings used in the various algorithms. Variables selected in models that perform variable selection can be found in Table S3.

In general, the highest training model CV-R^2 in the mobile data set were based on some of the data mining approaches, such as support vector machines, boosting and random forest (CV-R^2 range of 0.22−0.24; Table 2). The widely used stepwise regression algorithm predicted (only) 13% of the spatial variance of the mobile measurements.

Data mining algorithms, however, generated lower R^2 values (max R^2 = 0.50) when external long-term measurements were predicted. Regularized regression techniques (LASSO, elastic net and ridge), supervised stepwise regression, and GLM were able to explain external measurements equally well (R^2 0.61−0.62).

In the short term stationary data set most data mining methods, linear regression, and regularization algorithms predicted the short-term measurements equally well (R^2 range of 0.26−0.39, Table 3). In the short term stationary data set, some of the data mining algorithms explained variation in the external validation data by 58−70% (Table 3) compared to 53% for the customized stepwise regression model. Prediction of external measurements improved for most algorithms when only 20 predictors were offered. Short-term stationary predictions were much less affected by bias as the

DOI: 10.1021/acs.est.8b06038
Environ. Sci. Technol. 2019, 53, 1413−1421

Table 2. Comparison of Prediction Algorithms with Mobile UFP Data

algorithm	variables[a]	training model CV-R^2	external prediction R^2	external prediction mean bias[b] (in particles/cm^3)
Linear Regression				
Supervised Forward[c]	all	0.13	**0.62**	6902
Stepwise Both (R^2)	all	0.14	**0.54**	3525
Stepwise Both (AIC)	all	0.13	**0.53**	3307
Stepwise Backward (R^2)	all	0.14	**0.54**	3525
Stepwise Forward (R^2)	all	0.15	**0.60**	3293
Deletion/Substitution/Addition (DSA)[c]	all	0.14	**0.56**	6866
Regularization				
LASSO	all	0.13	**0.62**	3374
Elastic net (alpha = 0.25)	all	0.13	**0.63**	3182
Elastic net (alpha = 0.50)	all	0.13	**0.63**	3163
Elastic net (alpha = 0.75)	all	0.14	**0.62**	3366
Ridge	all	0.14	**0.63**	3144
	Top20	0.12	**0.61**	4367
Nonlinear				
Generalized Linear Model (GLM)	all	0.12	**0.58**	−518
	Top20	0.11	**0.57**	348
Multivariate Adaptive Regression Splines (MARS)	all	0.12	**0.41**	3261
	Top20	0.11	**0.43**	3750
General Additive Model (GAM)	all	0.17	**0.48**	2978
	Top20	0.14	**0.55**	4782
Data Mining				
Kernel Based Regularized Least Squares (KRLS)[c]	Regression Screening	0.18	**0.46**	5993
Neural network	all	0.06	**0.26**	4431
	Top20	0.05	**0.28**	4519
Support Vector Machine	all	0.23	**0.26**	1132
	Top20	0.17	**0.45**	1915
Random forest	all	0.24	**0.40**	5688
	Top20	0.22	**0.45**	5501
Gradient Boosting Machine (GBM)	all	0.15	**0.49**	4268
	Top20	0.14	**0.50**	4561
Extreme Boosting	all	0.23	**0.31**	5533
	Top20	0.18	**0.42**	5947
Bagging	all	0.16	**0.46**	4088
	Top20	0.15	**0.46**	4795
Super Learner	all		**0.37**	3383

[a]For algorithms that do not perform variable selection itself, we also offered the 20 variables that correlated best with the UFP measurements in the training data set. [b]Bias = Predicted concentration − Measured concentration. [c]Algorithm is the only algorithm that considers direction of effect and is not included in the Super Learner method.

Table 3. Comparison of Prediction Algorithms with Short-Term Stationary UFP Data

algorithm	variables[a]	training model CV-R^2	external prediction R^2	external prediction mean bias[b] (in particles/cm^3)
Linear Regression				
Supervised Forward[c]	all	0.35	**0.53**	151
Stepwise Both (R^2)	all	0.36	**0.54**	428
Stepwise Both (AIC)	all	0.29	**0.49**	−88
Stepwise Backward (R^2)	all	0.36	**0.54**	428
Stepwise Forward (R^2)	all	0.36	**0.52**	322
Deletion/Substitution/Addition (DSA)[c]	all	0.35	**0.51**	508
Regularization				
LASSO	all	0.30	**0.55**	−167
Elastic net (alpha = 0.25)	all	0.30	**0.56**	−389
Elastic net (alpha = 0.50)	all	0.30	**0.56**	−250
Elastic net (alpha = 0.75)	all	0.30	**0.55**	−207
Ridge	all	0.28	**0.61**	−142
	Top20	0.31	**0.55**	−30
Nonlinear				
Generalized Linear Model (GLM)	all	0.23	**0.53**	−249
	Top20	0.28	**0.55**	−892
Multivariate Adaptive Regression Splines (MARS)	all	0.04	**0.38**	740
	Top20	0.33	**0.45**	101
General Additive Model (GAM)	all	0.03	**0.37**	475
	Top20	0.02	**0.37**	973
Data Mining				
Kernel Based Regularized Least Squares (KRLS)[c]	Regression Screening	0.39	**0.60**	−625
Neural network	all	0.06	**0.22**	−1283
	Top20	0.10	**0.27**	−1166
Support Vector Machine	all	0.32	**0.61**	−796
	Top20	0.33	**0.60**	−731
Random forest	all	0.30	**0.66**	−229
	Top20	0.33	**0.70**	−54
Gradient Boosting Machine	all	0.31	**0.66**	−1225
	Top20	0.32	**0.64**	−1142
Extreme Boosting	all	0.26	**0.51**	−978
	Top20	0.29	**0.56**	−775
Bagging	all	0.31	**0.65**	−566
	Top20	0.34	**0.70**	−347
Super Learner	all		**0.60**	63

[a]For algorithms that do not perform variable selection itself we also offered the 20 variables that correlated best with the UFP measurements in the training data set. [b]Bias = Predicted concentration − Measured concentration. [c]Algorithm is the only algorithm that considers direction of effect and is not included in the Super Learner method.

mobile predictions. Mobile predictions are on average 3900 particles/cm^3 higher than actual measurements, whereas the bias in the short term prediction is close to zero (−260 particles/cm^3).

Figure 3 and Figure S1 illustrate the precision of predictive performance at the 42 external sites in 1000 bootstrap samples. Most algorithms showed an IQR of about 10−15% difference in R^2 (solid boxes in Figure 3) with increasing variability with more flexible model algorithms, meaning that such algorithms are more dependent on the sites used for validation. The precision illustrates that the performance of algorithms based on regularization, (supervised) linear regression, DSA, and some machine learning methods does not differ consistently. These algorithms ranked in the top 10 (out of 31 algorithms) in almost all bootstrapped samples (Figure S1). Figures SI.2.1−SI.2.10 illustrate that some of the modest differences in

DOI: 10.1021/acs.est.8b06038
Environ. Sci. Technol. 2019, 53, 1413−1421

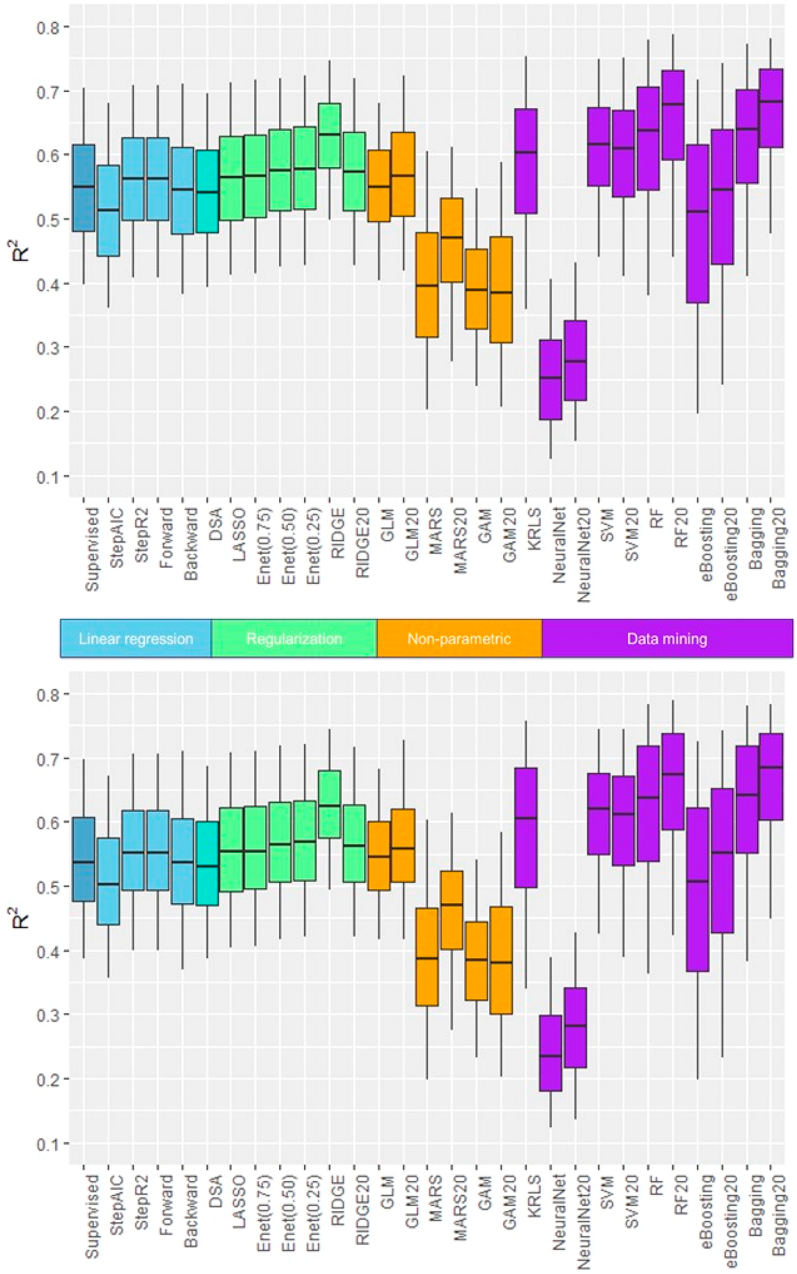


Figure 3. Distribution of *R*-squares when predictions of mobile (top) and short-term stationary (bottom) models are compared to external measurements, based on 1000 bootstrapped samples.

R^2 between models is partially due to how well the highest measured concentration is explained. Deletion of this point from the external data set diminished the difference in predictive performance between the algorithms (Figure S2). The super learner "stacking" algorithm did not improve upon the individual algorithms in our study. In the mobile monitoring data, model techniques that exploited training data with great detail (such as SVM, random forest, and extreme boosting) were chosen in the Super Learner (Table S4). These algorithms were individually able to generate a high training model R^2 but are unable to predict external long-term average concentrations. For example, the R^2 was 0.37 for the super learner algorithm, while regularization methods were able to predict external measurements by 62−63%. For the short-term stationary data, different prediction algorithms were chosen in the super learner algorithm: stepwise backward regression, regression splines (with screening), and extreme boosting (Table S5).

DISCUSSION

We compared different algorithms for developing land use regression models of spatial concentration variations of UFP based on mobile and short-term stationary monitoring campaigns. In general, we found modest differences in performance when external measurements were used for validation. In our mobile data set algorithms based on regularization (LASSO, elastic net, and ridge), (supervised) linear regression and DSA explained variance in the long term measurements almost equally and slightly better than data mining approaches. In the short term data set, however, data mining approaches, such as random forest, boosting, and bagging, explained variance in external measurements slightly

DOI: 10.1021/acs.est.8b06038
Environ. Sci. Technol. 2019, 53, 1413−1421

better than stepwise regression and regularization techniques. Algorithms based on regularization (LASSO, elastic net, and ridge), (customized) linear regression, and DSA explained variance in the long term measurements almost equally. The super learner algorithm did not improve external prediction in both data sets but performed reasonably well when short-term stationary data was used and had marginally lower bias estimates.

Comparing Prediction Algorithms. All algorithms in both the mobile and short-term stationary data set showed higher external prediction performance than the (cross-validated) performance in the training data (Tables 2 and 3). Training data were based on mobile or short-term data, with a resolution of ∼25 s and 3 × 30 min of sampling per site, respectively. Hence, training data contains more noise (with mobile measurements more so) than the long-term measurements of 3 times 24 h used for validation. We previously discussed and explained the higher R^2 in validation samples.[17,19,20]

For algorithms with a higher cross-validated training model R^2, we did not always find a higher R^2 for the external long-term average concentrations, stressing the importance of an external validation data set reflecting the temporal resolution of what one wants to predict. For example, the random forest algorithm, which predicted mobile measurements best (24%), predicted 40% of the variance in the external measurements (Table 2), whereas the customized stepwise regression procedure, regularization methods, and DSA predicted 13−15% of the mobile measurements and 61−62% of the external measurements. Data mining methods train models in great detail, possibly assigning too much value to individual mobile measurements, and create models that reflect patterns that are not present at the external sites using longer averaged measurements. We note that our study developed models based on training data with a different time scale than the validation data. The data further differ in their spatial features: on-road mobile monitoring versus residential outdoor monitoring, typically near the façade of homes. The validation data are important because they represent the locations where the models are mostly applied for epidemiological studies. Our study setting is therefore relevant for other studies based on mobile monitoring as well. We note that the data mining methods were developed using internal cross-validation using randomly selected sets. Autocorrelation present in the mobile monitoring set may have contributed to some overfitting of the models. We therefore performed an additional cross-validation analysis based on clusters of driving days and found that the CV-R^2 dropped for all methods (shown as clustered CV-R^2 in Table S4). Models that are used for external prediction are based on the full data set;, it did not alter external prediction performance of individual algorithms.

In the short term data, several machine learning algorithms (bagging, random forest) trained on short-term measurements explained modestly more variability of external UFP concentrations compared to linear regression and regularization techniques. We did not observe higher training model R^2 for the data mining methods. The short-term stationary measurements are an average of 3 times 30 min, and hence averages are more stable (less noise) than those based upon the mobile measurements. This may explain that certain features picked up by data mining algorithms were also helpful to predict long-term measurements.

This is also reflected in the algorithms that are selected in the super learner algorithms. In the mobile data three machine learning algorithms are chosen, whereas stepwise backward regression is given the most weight in the short term training data. These algorithms are chosen because the CV-R^2 shows that they do not overfit the training data. In other words, the predictive ability of that particular algorithm in the training data is high, and the risk of choosing that algorithm is low.

In both data sets, (supervised) linear regression model and regularization approaches provided similar and relatively high external prediction R^2. These algorithms were also found to describe data equally well in a study by Van de Bossche et al.,[16] evaluating LUR models for black carbon based on mobile measurements. LASSO, ridge, and elastic net can (better) deal with correlated predictors as they do not use a custom variable selection method but a regularization penalty.[16] Compared to data mining algorithms, an advantage of the regularization algorithms is that they keep their interpretability, similar to linear stepwise regression procedures, and in addition can perform variable selection not necessitating any a priori selection of variables. We note that a screening step based on simple correlation between predictors and pollution takes away some of the attractiveness of applying sophisticated data mining methods.

We have focused extensively on explained variance (R^2) of the models, as generalization of exposure contrast in an epidemiological study is a key goal of application of these models. We additionally assess bias. Models based upon the mobile monitoring but not the short-term monitoring overestimated long-term outdoor home façade measurements for most model algorithms. In previous studies we observed an average overestimation of 5000 particles/cm^3(30%) when models were developed with mobile monitoring data[17] and concluded that this is because mobile measurements were taken on the road itself, while short-term and long-term stationary measurements were taken on the sidewalk near the façade of homes. This is similar to a study by Simon et al.,[29] who compared UFP concentrations in Chelsea (Massachusetts, USA) based on residence and mobile monitoring. They found average differences of 5300 particles/cm^3, which is about the same as found in our studies.

Although no strong conclusions can be drawn regarding which algorithm is best, some recommendations can be given. The fact that most algorithms generate similar predictions, above all, is encouraging. Predictors that are selected or given the most weight in our exposure models are also similar. One could argue that with more robust training data (i.e., longer-term average), machine learning algorithms tend to operate better. Correlation structures in the data are probably more important. When the relationships between UFP and predictor variables are not complex (no major interactions and no major deviations from linear associations), there is not much to gain for machine learning algorithms as opposed to "simple" linear regression. More detailed predictor data could benefit certain algorithms as well. Computational runtime is not a limiting factor. Individual algorithms never took more than 15 min to run on a standard desktop computer. Total runtime of the super learner algorithm was about 12 h for the mobile data and 2 h for the short-term data. We therefore recommend using multiple approaches (that differ in model structure) in future studies. An external data set to test predictions may be valuable when training data differ in temporal resolution or when spatial features differ from the locations to which the models are

DOI: 10.1021/acs.est.8b06038
Environ. Sci. Technol. 2019, 53, 1413−1421

applied (short-term on road mobile monitoring versus long-term average residential address exposure). Such long-term UFP measurements as a test set of sufficient size is however difficult to obtain.

Limitations. Although the availability of long-term external residential level validation data at 42 sites is a major strength not typically available in other studies on UFP, we acknowledge that the limited number of sites requires that differences in external R^2 values between algorithms must be interpreted with caution. We therefore used a bootstrapping approach to quantify the precision of the R^2 and interpret the differences as modest. Most algorithms showed an IQR of about 10−15% difference in R^2 (solid boxes in Figure 3) with increasing variability with more flexible model algorithms, meaning that such algorithms are more dependent on the sites used for validation.

On top of that, it is possible that different settings (instead of default parameters) for certain algorithms can change the predictive performance. If a random forest or boosting model would be trained with parameters that are constrained in such a way that the model does not try to aggressively explain every single observation in the training data, it could improve external prediction performance. In the extreme case of extreme boosting, training model R^2 (without cross-validation) can go all the way to 1 in the short term stationary data (Table S5). Limiting the number of steps in a boosting model or restricting the number of trees or nodes in a random forest could alleviate this issue. Farrell et al.[25] and Patton et al.[9] restricted the number variables in their models (GLM and GAM respectively) by regression screening, which in our study increased R^2 values for most algorithms when external measurements were predicted. For GLM, the external prediction R^2 increased from 0.57 to 0.58 to 0.62 in the mobile data and remained the same in the short term data set. Predictions from the GAM models were 0.48 and 0.55 in the mobile data and 0.37 in the short term data. Using regression screening increased performance to 0.60 and 0.53 in the mobile and short-term data set, respectively. As there usually is no external long-term data available, it is difficult to choose parameters, especially when models are trained on data sets with different time-scales.

Previous Model Algorithm Comparisons. A few studies compared linear regression techniques to typically one or a few other algorithms. Weichenthal et al.[10] and Zou et al.[30] respectively compared KRLS and GAM to linear regression and found minor differences between methods, especially when tested on an external data set. Basagaña et al.[6] compared DSA to linear regression and found that predictive power was more related to the number of measurements sites than the applied algorithm. Brokamp et al.[31] compared linear regression to random forest and found that regression models showed higher prediction error variance with cross validation in 24 fixed sites. A difference with the settings in our study is the much smaller number of training sites, the more stable pollution data, and the fact that the authors used a variable selection method based on the variable importance in a random forest approach. This makes it more difficult to compare these results to algorithms that use other variable selection methods.[31]

In our study, we compared a much larger number of algorithms. The different findings in our study for the mobile and short-term training data and the somewhat inconsistent findings in previous studies preclude drawing strong conclusions based on empirical performance of models. This was also found by van den Bossche et al.,[16] who compared LASSO and Support Vector Regression (SVR) to linear regression in a mobile monitoring campaign of black carbon. To minimize overfitting of their models, they performed cross-validation with a fully rebuilt model (including variable selection) on every training set and found only small differences between the techniques. Worse predictive performance was found for algorithms that did not limit the number of predictor variables.[16] Performance of our SVM model was also better in predicting external measurements when only the variables were used that correlated best with the outcome in the training data set (0.45 opposed 0.36 in the mobile data set and 0.60 opposed to 0.50 in the short term data set). In general, data mining algorithms tended to benefit from variable reduction in the mobile data set (Table S4).

Cross validation and a subset of variables were also used by Reid et al.[15] predicting spatial variation of PM concentrations during wildfire events. They compared 11 algorithms, each with their optimal number of variables. Data mining methods (boosting, random forest, and SVM) were better at explaining $PM_{2.5}$ related to wildfire events (based on cross validated R^2 values) than regularization methods and additive models,[15] but differences were small and no external data set was used to assess their performance. The relative performance of these algorithms may depend on number of sites, noise in air pollution measurement, spatial contrast, extent of study area, number of and correlation among predictors.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.est.8b06038.

Part 1 is general tables and figures listing the GIS predictors and showing distributions of UFP measurements and sensitivity analyses. Part 2 elaborates on the different algorithms separately (in text, scatterplots, and parameter settings) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*Address: Institutes for Risk Assessment Sciences (IRAS) 3508 TD, Utrecht, The Netherlands. E-mail: j.kerckhoffs@uu.nl.

ORCID

Jules Kerckhoffs: 0000-0001-9065-6916

Notes

The authors declare no competing financial interest.

■ REFERENCES

(1) Ryan, P. H.; LeMasters, G. K. A Review of Land-Use Regression Models for Characterizing Intraurban Air Pollution Exposure. *Inhalation Toxicol.* **2007**, *19* (sup1), 127−133.

(2) Hoek, G. Methods for Assessing Long-Term Exposures to Outdoor Air Pollutants. *Curr. Environ. Heal. Reports* **2017**, *4* (4), 450−462.

(3) James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning*, 2000; DOI: 10.1007/978-1-4614-7138-7.

(4) Hastie, T.; Tibshirani, R.; Friedman, J. The Elements of Statistical Learning. *Elements* **2009**, *1*, 337−387.

(5) Wang, M.; Beelen, R.; Basagana, X.; Becker, T.; Cesaroni, G.; de Hoogh, K.; Dedele, A.; Declercq, C.; Dimakopoulou, K.; Eeftens, M.; Forastiere, F.; Galassi, C.; Gražulevičiené, R.; Hoffmann, B.; Heinrich,

J.; Iakovides, M.; Künzli, N.; Korek, M.; Lindley, S.; Mölter, A.; Mosler, G.; Madsen, C.; Nieuwenhuijsen, M.; Phuleria, H.; Pedeli, X.; Raaschou-Nielsen, O.; Ranzi, A.; Stephanou, E.; Sugiri, D.; Stempfelet, M.; Tsai, M.-Y.; Lanki, T.; Udvardy, O.; Varró, M. J.; Wolf, K.; Weinmayr, G.; Yli-Tuomi, T.; Hoek, G.; Brunekreef, B. Evaluation of Land Use Regression Models for NO 2 and Particulate Matter in 20 European Study Areas: The ESCAPE Project. *Environ. Sci. Technol.* **2013**, *47* (9), 4357−4364.

(6) Basagaña, X.; Rivera, M.; Aguilera, I.; Agis, D.; Bouso, L.; Elosua, R.; Foraster, M.; de Nazelle, A.; Nieuwenhuijsen, M.; Vila, J.; Künzli, N. Effect of the Number of Measurement Sites on Land Use Regression Models in Estimating Local Air Pollution. *Atmos. Environ.* **2012**, *54*, 634−642.

(7) Bellinger, C.; Mohomed Jabbar, M. S.; Zaïane, O.; Osornio-Vargas, A. A Systematic Review of Data Mining and Machine Learning for Air Pollution Epidemiology. *BMC Public Health* **2017**, *17* (1), 907.

(8) Hasenfratz, D.; Saukh, O.; Walser, C.; Hueglin, C.; Fierz, M.; Arn, T.; Beutel, J.; Thiele, L. Deriving High-Resolution Urban Air Pollution Maps Using Mobile Sensor Nodes. *Pervasive and Mobile Computing* **2015**, *16*, 268−285.

(9) Patton, A. P.; Collins, C.; Naumova, E. N.; Zamore, W.; Brugge, D.; Durant, J. L. An Hourly Regression Model for Ultrafine Particles in a Near-Highway Urban Area. *Environ. Sci. Technol.* **2014**, *48* (6), 3272−3280.

(10) Weichenthal, S.; Van Ryswyk, K.; Goldstein, A.; Bagg, S.; Shekkarizfard, M.; Hatzopoulou, M. A Land Use Regression Model for Ambient Ultrafine Particles in Montreal, Canada: A Comparison of Linear Regression and a Machine Learning Approach. *Environ. Res.* **2016**, *146*, 65−72.

(11) Liu, W.; Li, X.; Chen, Z.; Zeng, G.; León, T.; Liang, J.; Huang, G.; Gao, Z.; Jiao, S.; He, X.; Lai, M. Land Use Regression Models Coupled with Meteorology to Model Spatial and Temporal Variability of NO_2 and PM10 in Changsha, China. *Atmos. Environ.* **2015**, *116*, 272−280.

(12) Song, B.; Wu, J.; Zhou, Y.; Hu, K. Fine-Scale Prediction of Roadside CO and NOx Concentration Based on a Random Forest Model. *J. Residuals Sci. Technol.* **2014**, *11*, 83−89.

(13) Lima, A. R.; Cannon, A. J.; Hsieh, W. W. Nonlinear Regression in Environmental Sciences Using Extreme Learning Machines: A Comparative Evaluation. *Environ. Model. Softw.* **2015**, *73*, 175−188.

(14) Lary, D. J.; Alavi, A. H.; Gandomi, A. H.; Walker, A. L. Machine Learning in Geosciences and Remote Sensing. *Geosci. Front.* **2016**, *7* (1), 3−10.

(15) Reid, C. E.; Jerrett, M.; Petersen, M. L.; Pfister, G. G.; Morefield, P. E.; Tager, I. B.; Raffuse, S. M.; Balmes, J. R. Spatiotemporal Prediction of Fine Particulate Matter during the 2008 Northern California Wildfires Using Machine Learning. *Environ. Sci. Technol.* **2015**, *49* (6), 3887−3896.

(16) Van den Bossche, J.; De Baets, B.; Verwaeren, J.; Botteldooren, D.; Theunis, J. Development and Evaluation of Land Use Regression Models for Black Carbon Based on Bicycle and Pedestrian Measurements in the Urban Environment. *Environ. Model. Softw.* **2018**, *99*, 58−69.

(17) Kerckhoffs, J.; Hoek, G.; Vlaanderen, J.; van Nunen, E.; Messier, K.; Brunekreef, B.; Gulliver, J.; Vermeulen, R. Robustness of Intra Urban Land-Use Regression Models for Ultrafine Particles and Black Carbon Based on Mobile Monitoring. *Environ. Res.* **2017**, *159*, 500−508.

(18) Kerckhoffs, J.; Hoek, G.; Messier, K. P.; Brunekreef, B.; Meliefste, K.; Klompmaker, J. O.; Vermeulen, R. Comparison of Ultrafine Particles and Black Carbon Concentration Predictions from a Mobile and Short-Term Stationary Land-Use Regression Model. *Environ. Sci. Technol.* **2016**, *50* (23), 12894−12902.

(19) Montagne, D. R.; Hoek, G.; Klompmaker, J. O.; Wang, M.; Meliefste, K.; Brunekreef, B. Land Use Regression Models for Ultrafine Particles and Black Carbon Based on Short-Term Monitoring Predict Past Spatial Variation. *Environ. Sci. Technol.* **2015**, *49* (14), 8712−8720.

(20) van Nunen, E.; Vermeulen, R.; Tsai, M.-Y.; Probst-Hensch, N.; Ineichen, A.; Davey, M. E.; Imboden, M.; Ducret-Stich, R.; Naccarati, A.; Raffaele, D.; Ranzi, A.; Ivaldi, C.; Galassi, C.; Nieuwenhuijsen, M. J.; Curto, A.; Donaire-Gonzalez, D.; Cirach, M.; Chatzi, L.; Kampouri, M.; Vlaanderen, J.; Meliefste, K.; Buijtenhuijs, D.; Brunekreef, B.; Morley, D.; Vineis, P.; Gulliver, J.; Hoek, G. Land Use Regression Models for Ultrafine Particles in Six European Areas. *Environ. Sci. Technol.* **2017**, *51* (6), 3336−3345.

(21) van der Laan, M. J.; Polley, E. C.; Hubbard, A. E. Super Learner. *Stat. Appl. Genet. Mol. Biol.* **2007**, *6* (1), Article 25, DOI: 10.2202/1544-6115.1309.

(22) Hoek, G.; Beelen, R.; de Hoogh, K.; Vienneau, D.; Gulliver, J.; Fischer, P.; Briggs, D. A Review of Land-Use Regression Models to Assess Spatial Variation of Outdoor Air Pollution. *Atmos. Environ.* **2008**, *42* (33), 7561−7578.

(23) Eeftens, M.; Tsai, M. Y.; Ampe, C.; Anwander, B.; Beelen, R.; Bellander, T.; Cesaroni, G.; Cirach, M.; Cyrys, J.; de Hoogh, K.; De Nazelle, A.; de Vocht, F.; Declercq, C.; Dedele, A.; Eriksen, K.; Galassi, C.; Gražulevičiene, R.; Grivas, G.; Heinrich, J.; Hoffmann, B.; Iakovides, M.; Ineichen, A.; Katsouyanni, K.; Korek, M.; Krämer, U.; Kuhlbusch, T.; Lanki, T.; Madsen, C.; Meliefste, K.; Mölter, A.; Mosler, G.; Nieuwenhuijsen, M.; Oldenwening, M.; Pennanen, A.; Probst-Hensch, N.; Quass, U.; Raaschou-Nielsen, O.; Ranzi, A.; Stephanou, E.; Sugiri, D.; Udvardy, O.; Vaskövi, É.; Weinmayr, G.; Brunekreef, B.; Hoek, G. Spatial Variation of PM2.5, PM10, PM2.5 Absorbance and PMcoarse Concentrations between and within 20 European Study Areas and the Relationship with NO2 - Results of the ESCAPE Project. *Atmos. Environ.* **2012**, *62*, 303−317.

(24) Beckerman, B. S.; Jerrett, M.; Martin, R. V.; Van Donkelaar, A.; Ross, Z.; Burnett, R. T. Application of the Deletion/Substitution/Addition Algorithm to Selecting Land Use Regression Models for Interpolating Air Pollution Measurements in California. *Atmos. Environ.* **2013**, *77*, 172−177.

(25) Farrell, W.; Weichenthal, S.; Goldberg, M.; Valois, M. F.; Shekarrizfard, M.; Hatzopoulou, M. Near Roadway Air Pollution across a Spatially Extensive Road and Cycling Network. *Environ. Pollut.* **2016**, *212*, 498−507.

(26) Zwack, L. M.; Paciorek, C. J.; Spengler, J. D.; Levy, J. I. Modeling Spatial Patterns of Traffic-Related Air Pollutants in Complex Urban Terrain. *Environ. Health Perspect.* **2011**, *119* (6), 852−859.

(27) Lu, W.-Z.; Wang, W.-J. Potential Assessment of the "Support Vector Machine" Method in Forecasting Ambient Air Pollutant Trends. *Chemosphere* **2005**, *59* (5), 693−701.

(28) Luna, A. S.; Paredes, M. L. L.; de Oliveira, G. C. G.; Corrêa, S. M. Prediction of Ozone Concentration in Tropospheric Levels Using Artificial Neural Networks and Support Vector Machine at Rio de Janeiro, Brazil. *Atmos. Environ.* **2014**, *98*, 98−104.

(29) Simon, M. C.; Hudda, N.; Naumova, E. N.; Levy, J. I.; Brugge, D.; Durant, J. L. Comparisons of Traffic-Related Ultrafine Particle Number Concentrations Measured in Two Urban Areas by Central, Residential, and Mobile Monitoring. *Atmos. Environ.* **2017**, *169*, 113−127.

(30) Zou, B.; Chen, J.; Zhai, L.; Fang, X.; Zheng, Z. Satellite Based Mapping of Ground PM2.5 Concentration Using Generalized Additive Modeling. *Remote Sens.* **2017**, *9* (1), 1.

(31) Brokamp, C.; Jandarov, R.; Rao, M. B.; LeMasters, G.; Ryan, P. Exposure Assessment Models for Elemental Components of Particulate Matter in an Urban Environment: A Comparison of Regression and Random Forest Approaches. *Atmos. Environ.* **2017**, *151*, 1−11.

DOI: 10.1021/acs.est.8b06038
Environ. Sci. Technol. 2019, 53, 1413−1421