Environment International 130 (2019) 104934

Contents lists available at ScienceDirect

Environment International

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

A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide

Jie Chen[a,*], Kees de Hoogh[b,c], John Gulliver[d], Barbara Hoffmann[e], Ole Hertel[f], Matthias Ketzel[f,g], Mariska Bauwelinck[h], Aaron van Donkelaar[i], Ulla A. Hvidtfeldt[j], Klea Katsouyanni[k,l], Nicole A.H. Janssen[m], Randall V. Martin[i,n], Evangelia Samoli[k], Per E. Schwartz[o], Massimo Stafoggia[p,q], Tom Bellander[q], Maciek Strak[a], Kathrin Wolf[r], Danielle Vienneau[b,c], Roel Vermeulen[a,s], Bert Brunekreef[a,s], Gerard Hoek[a]

[a] Institute for Risk Assessment Sciences (IRAS), Utrecht University, Postbus 80125, 3508 TC, Utrecht, the Netherlands
[b] Swiss Tropical and Public Health Institute, Socinstrasse 57, 4051 Basel, Switzerland
[c] University of Basel, Petersplatz 1, Postfach 4001 Basel, Switzerland
[d] Centre for Environmental Health and Sustainability, School of Geography, Geology and the Environment, University of Leicester, University Road, Leicester LE1 7RH, UK
[e] Institute for Occupational, Social and Environmental Medicine, Centre for Health and Society, Medical Faculty, Heinrich Heine University Düsseldorf, Universitätsstraße 1, 40225 Düsseldorf, Germany
[f] Department of Environmental Science, Aarhus University, P.O. Box 358, Frederiksborgvej 399, 4000 Roskilde, Denmark
[g] Global Centre for Clean Air Research (GCARE), Department of Civil and Environmental Engineering, University of Surrey, Guildford GU2 7XH, UK
[h] Interface Demography, Department of Sociology, Vrije Universiteit Brussel, Pleinlaan 2, 1050, Brussels, Belgium
[i] Department of Physics and Atmospheric Science, Dalhousie University, B3H 4R2 Halifax, Nova Scotia, Canada
[j] Danish Cancer Society Research Center, Strandboulevarden 49, 2100 Copenhagen, Denmark
[k] Department of Hygiene, Epidemiology and Medical Statistics, Medical School, National and Kapodistrian University of Athens, 75 Mikras Asias Str, 115 27 Athens, Greece
[l] Department Population Health Sciences and Department of Analytical, Environmental and Forensic Sciences, School of Population Health & Environmental Sciences, King's College Strand, London WC2R 2LS, UK
[m] National Institute for Public Health and the Environment (RIVM), PO Box 1, 3720 BA, Bilthoven, the Netherlands
[n] Atomic and Molecular Physics Division, Harvard-Smithsonian Center for Astrophysics, 60 Garden St, Cambridge, MA 02138, USA
[o] Division of Environmental Medicine, Norwegian Institute of Public Health, PO Box 4404 Nydalen, N-0403 Oslo, Norway
[p] Department of Epidemiology, Lazio Region Health Service/ASL Roma 1, Via Cristoforo Colombo, 112, 00147, Rome, Italy
[q] Institute of Environmental Medicine, Karolinska Institutet, SE-171 77 Stockholm, Sweden
[r] Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Institute of Epidemiology, Ingolstädter Landstr. 1, D-85764 Neuherberg, Germany
[s] Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX, Utrecht, the Netherlands

ARTICLE INFO

Handling Editor: Xavier Querol

Keywords:
Land use regression
Fine particles
Nitrogen dioxide
Machine learning

ABSTRACT

Empirical spatial air pollution models have been applied extensively to assess exposure in epidemiological studies with increasingly sophisticated and complex statistical algorithms beyond ordinary linear regression. However, different algorithms have rarely been compared in terms of their predictive ability.

This study compared 16 algorithms to predict annual average fine particle ($PM_{2.5}$) and nitrogen dioxide (NO_2) concentrations across Europe. The evaluated algorithms included linear stepwise regression, regularization techniques and machine learning methods. Air pollution models were developed based on the 2010 routine monitoring data from the AIRBASE dataset maintained by the European Environmental Agency (543 sites for $PM_{2.5}$ and 2399 sites for NO_2), using satellite observations, dispersion model estimates and land use variables as predictors. We compared the models by performing five-fold cross-validation (CV) and by external validation

* Corresponding author.
E-mail addresses: j.chen1@uu.nl (J. Chen), c.dehoogh@swisstph.ch (K. de Hoogh), jg435@leicester.ac.uk (J. Gulliver), B.Hoffmann@uni-duesseldorf.de (B. Hoffmann), oh@envs.au.dk (O. Hertel), mke@envs.au.dk (M. Ketzel), mariska.bauwelinck@vub.ac.be (M. Bauwelinck), kelaar@Dal.Ca (A. van Donkelaar), ullah@cancer.dk (U.A. Hvidtfeldt), kkatsouy@med.uoa.gr (K. Katsouyanni), nicole.janssen@rivm.nl (N.A.H. Janssen), Randall.Martin@Dal.Ca (R.V. Martin), esamoli@med.uoa.gr (E. Samoli), Per.Schwarze@fhi.no (P.E. Schwartz), m.stafoggia@deplazio.it (M. Stafoggia), Tom.Bellander@ki.se (T. Bellander), M.M.Strak@uu.nl (M. Strak), kathrin.wolf@helmholtz-muenchen.de (K. Wolf), danielle.vienneau@swisstph.ch (D. Vienneau), R.C.H.Vermeulen@uu.nl (R. Vermeulen), B.Brunekreef@uu.nl (B. Brunekreef), G.Hoek@uu.nl (G. Hoek).

https://doi.org/10.1016/j.envint.2019.104934
Received 8 February 2019; Received in revised form 21 May 2019; Accepted 13 June 2019
Available online 20 June 2019

(EV) using annual average concentrations measured at 416 ($PM_{2.5}$) and 1396 sites (NO_2) from the ESCAPE study. We further assessed the correlations between predictions by each pair of algorithms at the ESCAPE sites.

For $PM_{2.5}$, the models performed similarly across algorithms with a mean CV R^2 of 0.59 and a mean EV R^2 of 0.53. Generalized boosted machine, random forest and bagging performed best (CV R^2~0.63; EV R^2 0.58–0.61), while backward stepwise linear regression, support vector regression and artificial neural network performed less well (CV R^2 0.48–0.57; EV R^2 0.39–0.46). Most of the $PM_{2.5}$ model predictions at ESCAPE sites were highly correlated (R^2 > 0.85, with the exception of predictions from the artificial neural network). For NO_2, the models performed even more similarly across different algorithms, with CV R^2s ranging from 0.57 to 0.62, and EV R^2s ranging from 0.49 to 0.51. The predicted concentrations from all algorithms at ESCAPE sites were highly correlated (R^2 > 0.9). For both pollutants, biases were low for all models except the artificial neural network. Dispersion model estimates and satellite observations were two of the most important predictors for $PM_{2.5}$ models whilst dispersion model estimates and traffic variables were most important for NO_2 models in all algorithms that allow assessment of the importance of variables.

Different statistical algorithms performed similarly when modelling spatial variation in annual average air pollution concentrations using a large number of training sites.

Abbreviations

ANN	artificial neural network
BLR	backward stepwise linear regression
CTM	chemical transport models
CV	cross validation
DSA	deletion/substitution/addition
EN25/50/75	elastic net with α = 0.25/0.50/0.75
EV	external validation
FLR	forward stepwise linear regression
GAM	generalized additive model
GBM	generalized boosted machine
KRLS	kernel-based regularized least squares
LASSO	least absolute shrinkage and selection operator
LUR	land use regression
NO_2	nitrogen dioxide
$PM_{2.5}$	particulate matter with an aerodynamic diameter smaller than 2.5 μm
RF	random forest
RMSE	root-mean-square error
SAT	satellite-derived
SLR	supervised linear regression
SVR	support vector regression
WLR	stepwise linear regression

1. Introduction

Research in developed countries is currently focusing on health effects of long-term exposure to ambient air pollution at low concentrations, where the concentration contrast is small (Beelen et al., 2014; Di et al., 2017; Pinault et al., 2017). In order to do so, accurately assessing exposure for study subjects is particularly important.

Land Use Regression (LUR) models are frequently used to assess air pollution exposure in epidemiological studies on long-term health effects of air pollution. These are empirical models derived by combining air pollution concentrations monitored at a limited number of locations and potential predictor variables collected in a geographic information systems (GIS) (Hoek, 2017). In a LUR model, a linear regression with an automatic variable selection algorithm has often been used to maximize the within-sample explained variation of measured air pollution concentrations (Crouse et al., 2009; Hoek et al., 2008). Some LUR models are fitted with a supervised linear regression algorithm to include only predictor variables following the plausible direction of effect, e.g. a positive traffic slope, in order to increase the physical interpretability and potentially transferability of the models (Brauer et al., 2003; Briggs et al., 1997; Henderson et al., 2007).

There are several concerns about the standard linear regression algorithm. One is that the algorithm may overfit the data when there are relatively few monitoring sites to train a model and a large number of potential predictor variables offered (Basagaña et al., 2012; Friedman et al., 2001; Wang et al., 2012). Second, the algorithm may fail to capture potentially complex relationships within the data, since it assumes the relationship between air pollution and a predictor is linear across the whole range of the predictor values, and the impacts of different predictors to be independent (no interaction) (Friedman et al., 2001; Tibshirani et al., 2013). Third, the algorithm may result in unstable and uninterpretable coefficient estimates when highly correlated predictors are included in one model (Crouse et al., 2009; Tibshirani et al., 2013).

A number of different algorithms beyond standard linear regression have increasingly been applied to fit LUR models in air pollution exposure assessment. The Deletion/Substitution/Addition (DSA) algorithm selects the subset of predictors that minimize the cross-validation mean squared errors (Basagaña et al., 2012; Beckerman et al., 2013). The Generalized Additive Model (GAM) algorithm estimates nonlinear relationships between air pollution and predictors (Liu et al., 2009). The LASSO (Least Absolute Shrinkage and Selection Operator) algorithm addresses collinearity by shrinking the coefficients of correlated predictors towards zero (Kim et al., 2016). Machine learning algorithms can detect previously unknown relationships within the data by modelling nonlinearity and interactions. Applications of the machine learning algorithms include Artificial Neural Network (ANN) (Di et al., 2016b; Zou et al., 2015), Random Forests (RF) (Brokamp et al., 2018; Hu et al., 2017; Zhan et al., 2018), Support Vector Regression (SVR) (de Hoogh et al., 2018b; Stafoggia et al., 2017; Van den Bossche et al., 2018), Generalized Boosted Machine (GBM) (Reid et al., 2015; Zhan et al., 2017), and Kernel-based Regularized Least Squares (KRLS) (Weichenthal et al., 2016).

Few studies have compared the performance of different algorithms in building LUR models for exposure assessment. The generalized boosted machine algorithm outperformed 10 other algorithms in a study modelling spatiotemporal variation of daily $PM_{2.5}$ concentrations during wildfires (Reid et al., 2015). The random forest algorithm outperformed the linear stepwise regression algorithm in a study modelling spatial variation of $PM_{2.5}$ and its components (Brokamp et al., 2017). In other studies modelling spatial variation of air pollution concentrations, similar performance was found using the Deletion/ Substitution/ Addition algorithm and the linear regression algorithm (Basagaña et al., 2012), the Kernel-based Regularized Least Squares algorithm and the linear regression algorithm (Weichenthal et al., 2016), the linear regression algorithm, the LASSO algorithm and the Support Vector Regression algorithm (Van den Bossche et al., 2018). Only modest differences in performance across algorithms were found in a recent comprehensive algorithm comparison study based on mobile monitoring of ultrafine particles (Kerckhoffs et al., 2019).

Most of the previous air pollution exposure assessment studies evaluated model performance based on cross-validation, which assesses a model's predictive ability within the monitoring domain. For cross-validation to be meaningful, the monitoring sites need to be representative of the locations to which the model is applied (e.g. residential addresses versus routine monitoring sites other than in

residential areas or on-road mobile monitoring). A model that performs well in cross-validation does not necessarily transfer well to application in epidemiological studies (Hystad et al., 2011; Kerckhoffs et al., 2016). Therefore, it is valuable to evaluate models using pollution data collected from monitoring sites which represent the application locations.

We have recently developed spatial air pollution models across Europe, using a Supervised Linear Regression (SLR) algorithm (de Hoogh et al., 2018a). The aim of the current study was to compare 16 different algorithms, including the SLR, in their ability to predict spatial variation of $PM_{2.5}$ and NO_2 concentrations across Europe. To strengthen our comparisons, we used two ground-based monitoring datasets to perform both cross-validation (AIRBASE dataset; EEA) and external validation (ESCAPE (European Study of Cohorts for Air Pollution Effects) dataset) (Cyrys et al., 2012; Eeftens et al., 2012). The selected algorithms follow a recent evaluation of model development of mobile monitoring data by our group (Kerckhoffs et al., 2019).

2. Methods

Measured air pollution concentration data and GIS predictor variables were the same as in our recently published Europe-wide modelling study (de Hoogh et al., 2018a).

2.1. Air pollution monitoring data

To build the spatial empirical models, we used annual mean concentrations for $PM_{2.5}$ (available for 543 sites) and NO_2 (available for 2399 sites) for 2010 from the AIRBASE v8 dataset (EEA, 2015) (Fig. S1). AIRBASE is a database maintained by the European Environmental Agency (EEA) containing monitoring data reported by EU member states and associated countries. Air pollution data are from routine regulatory networks in individual countries, measured by a diversity of methods. The monitoring locations are chosen to check for compliance with the European Union air quality standards (http://ec.europa.eu/environment/air/quality/standards.htm) at background sites, near busy roads or in industrial zones. The annual mean concentrations were aggregated by EEA based on the primary observations uploaded by countries and successfully tested by automated quality control. The primary observations were reported with different frequency (hour, day, or week). Based on the frequency of reported air pollution values, an annual average was calculated only when valid measurements coverage $\geq$75% of a year. AIRBASE monitoring sites were randomly divided into five groups (20% each), stratified by site type and region (de Hoogh et al., 2018a). Main models were built using all measurements (100% sites) in the AIRBASE dataset. Each of the 5 hold-out validation models was developed based on 80% of the monitoring sites, with the remaining 20% used for validation.

Air pollution monitoring data from the ESCAPE study were used for external validation. Three 2-weekly measurement campaigns were held at 416 monitoring sites for $PM_{2.5}$ and 1396 sites for NO_2, using Harvard Impactors and Ogawa badges respectively (Cyrys et al., 2012; Eeftens et al., 2012). The annual mean concentrations reflecting the period 2009–2010 were derived based on measurements in the three seasons with temporal adjustment. Measurement sites in ESCAPE were specifically selected to represent spatial variation of air pollution at home addresses of subjects in the included cohorts, thus the monitoring sites were clustered (Fig. S1).

Summary statistics of the training and validation datasets are presented in Table S1.

2.2. Predictor variables

Potential predictor variables used in this study are described in more detail elsewhere (de Hoogh et al., 2016; de Hoogh et al., 2018a; Vienneau et al., 2013). The predictor variables are integrated into a 100 m gridded GIS database covering Western Europe. All potential predictor variables and summary statistics are shown in Table S2. We offered 150 potential predictor variables.

2.2.1. Satellite-derived air pollution estimates and chemical transport model estimates

Satellite-derived (SAT) estimates of $PM_{2.5}$ were obtained from global datasets (V3.GL.01; Van Donkelaar et al., 2015). A gridded surface of the 2010 annual average $PM_{2.5}$ was produced at a $0.1^\circ \times 0.1^\circ$ ($\sim 10 \times 10$ km) resolution by relating aerosol optical depth (AOD) retrievals from the NASA MODIS (Moderate Resolution Imaging Spectroradiometer), MISR (Multi-angle Imaging Spectroradiometer) and SeaWiFS (Sea-viewing Wide Field-of-view Sensor) instruments to near-surface concentrations using aerosol vertical profiles and scattering properties simulated by the GEOS-Chem chemical transport model. For NO_2, SAT estimates for 2010 were derived from the tropospheric NO_2 columns measured with the OMI (Ozone Monitoring Instrument) on board the Aura satellite. The satellite column-integrated retrievals were related to ground-level concentrations using the global GEOS-Chem model to produce a 10×10 km resolution dataset (Bechle et al., 2013, 2015; Novotny et al., 2011).

Annual $PM_{2.5}$ and NO_2 chemical transport models (CTM) estimates for 2010 were derived from the MACC-II ENSEMBLE model at a $0.1^\circ \times 0.1^\circ$ ($\sim 10 \times 10$ km) resolution (Inness et al., 2013). In the ENSEMBLE model, the median value of seven individual regional CTMs (CHIMERE, EMEP, EURAD, LOTOS-EUROS, MATCH, MOCAGE and SILAM) was provided for each pixel.

2.2.2. Traffic, land use and altitude predictors

Road data were extracted from the 1:10,000 EuroStreets digital road network (version 3.1 based on TeleAtlas MultiNet TM, year 2008), classified into 'all' and 'major' roads. These were then intersected with a 100 m base polygon and the sum of the road lengths was calculated within each grid cell.

The European Corine Land Cover 2006 dataset (ETC-LC, 2009) was used to extract land cover variables for all study areas except for Greece, which has missing data. We used the Corine Land Cover 2000 (ETC-LC, 2013) to extract data for Greece. Six main groups (residential, industry, ports, urban green space, total built up land and natural land) were derived from the initial 44 land classes. A moving window procedure was used to calculate both road and land cover data for selected radii, which ranged from 50 m to 10,000 m (Focalstatistics using sum with a circle).

Elevation was obtained from the SRTM Digital Elevation Database version 4.1 with a resolution of 3 arc sec (approximately 90 m) with vertical error of < 16 m (CGIAR-CSI).

2.3. Model development

We applied 16 statistical algorithms to build the models. These algorithms cover almost all algorithms applied in previous LUR models and have been assessed in a model comparison paper using mobile monitoring ultrafine particle concentration data (Kerckhoffs et al., 2019).

For each algorithm, 6 models (1 main model plus 5 hold-out validation models) were developed for both pollutants (see Section 2.1). We used grid search to optimize hyperparameters (whose values were set before the training process) for each model, based on the minimum mean cross-validated error. This approach helped to minimize the risk of overfitting and ensured that the models we derived had the best predictive power. Hyperparameters for each algorithm were specified in Table S3.

Linear stepwise regression algorithms assume that the relationships between the pollutants and the predictors are linear and additive (Tibshirani et al., 2013). We used 3 automatic variable selection methods to choose the best subset of predictors. **Forward stepwise Linear Regression (FLR)** started with a null model, then the predictor

that generated the highest increase in the adjusted R^2 was added to the model at each subsequent step. This process was repeated until the model adjusted R^2 stopped maximizing. **Backward stepwise Linear Regression (BLR)** began with all variables in the model and deleted the variable with the highest *P*-value one at a time. The procedure stopped when it generated a model that had only significant predictors (significance level of 0.1) with the maximum adjusted R^2. **Stepwise Linear Regression (WLR)** allowed variables to be added or deleted as modelling progresses. The algorithm started off in a forward approach with a null model, and then removed variables if they became statistically insignificant (significance level of 0.1). We also used a **Supervised Linear Regression (SLR)** algorithm that was described previously in de Hoogh et al. (2018a). In this algorithm, a univariate linear regression model was run for each potential predictor to choose the model with the highest adjusted R^2 as the starting point. Additional significant predictor variables were allowed to enter the model if they added to the adjusted R^2 of the previous model step, and only if they adhered to the plausible direction of effect. Variables with variance inflation factor (VIF) larger than 3 were removed from the model to avoid multicollinearity.

Regularization or shrinkage algorithms are used to estimate reliable predictor coefficients when the predictors are highly correlated. By imposing different penalties, **ridge regression** keeps all predictors in the final model, while **LASSO** ensures sparsity of the results by shrinking some coefficients exactly to zero. **Elastic Net** is a hybrid of ridge regression and LASSO by adjusting the values of hyperparameter α (Friedman et al., 2009). Elastic net is the same as lasso when $\alpha = 1$, it approaches ridge regression as α reduces towards 0. In this study, $\alpha = 0.25$ (**EN25**), 0.5 (**EN50**) and 0.75 (**EN75**) were used to build separate elastic net models.

The **Generalized Additive Model (GAM) algorithm** (Wood and Wood, 2015) extends the standard linear regression by introducing non-linear functions for predictors while keeping the additive assumption. We used "gam" function in the "mgcv" R package, which performs automatic smoothing parameter estimation and allows adding an extra penalty to remove redundant variables from the model. A smoothing spline was fit for potential predictors with at least 5 unique values. Variables with < 5 differing values, i.e. the land use variables in the smallest buffers (TBU50, NAT50, IND50, POR50, UGR50, RES50), were deleted because the function could not estimate the smoothing parameters for them. The roughness of the smoothing spline was selected via restricted maximum likelihood method (REML).

Machine learning algorithms are able to model nonlinearity as well as the potentially complex interactions among predictors. One type of machine learning algorithms is the ensemble learning machine based on decision trees. **Bagging**, also known as the bootstrap aggregation, repeatedly draws separate subsets from the full training dataset. The final predictions were calculated by averaging the results from all the decision trees built on bootstrapped training subsets. **Random Forest (RF)** (Breiman et al., 2011) adds an additional layer of randomness to bagging by forcing each split to consider only a randomly chosen subset of candidate predictors, instead of the full set. Instead of building independent trees using bootstrapped samples, **Generalized Boosted Machine (GBM)** (Ridgeway et al., 2013) grows trees sequentially: each tree is fit on the residuals of the given model. Other types of machine learning algorithms include the **Support Vector Regression (SVR)** algorithm (Friedman et al., 2001; Meyer et al., 2017), which uses kernel functions to enlarge the feature space and produces non-linear boundaries by constructing a linear boundary in a transformed high-dimensional feature space; the **Kernel-based Regularized Least Squares (KRLS)** (Ferwerda et al., 2017; Hazlett and Hainmueller, 2017) algorithm, whose kernel function measures the similarity between covariates while the regularization imposes a preference for a smoother function; and the **Artificial Neural Network (ANN)**(Ripley et al., 2016) algorithm, which consist of interconnected "neurons" (represent predictors) in layers that can account for possible nonlinearities and interactions.

We additionally use two approaches to make ensemble predictions that leverage information from all models. In Ensemble 1 model, the median value of 16 model predictions was provided for each site (Inness et al., 2013). In Ensemble 2 model, a weighted average of the 16 model predictions was provided for each site. The weight (w_i) of each model was calculated based on the inverse of average cross-validation absolute bias at all AIRBASE sites. The ensemble prediction at each site ($\hat{y}$) was defined as:

$$\hat{y} = \frac{\sum_{i=1}^{16} w_i y_i}{\sum_{i=1}^{16} w_i}$$

where y_i is the prediction of the individual models.

2.4. Model evaluation and comparison

We evaluated model performance by regression-based R^2 (R^2), mean square error based R^2 (MSE-R^2), and root-mean-square error (RMSE). R^2 was derived from correlations between predicted and observed values. MSE-R^2 can be seen as a rescaling of MSE. It measures fit about the 1:1 line rather than fit about the best fit line in regression-based R^2. The formula was defined as:

$$\text{MSE} - R^2 = 1 - \frac{MSE}{\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2\right)}$$

where $\bar{y}$ is the average of the observed values. MES-R^2 can yield negative values when the average of the observed values performs better than the predictions of the model (Wang et al., 2012). RMSE was computed as the square root of the mean of squared difference between predicted and observed values. **Training** R^2 and RMSE were calculated by comparing the predictions and the observations at all AIRBASE sites. **Cross-validated (CV)** R^2, MSE-R^2, and RMSE were computed by comparing the assembled predictions at 5 held-out sets to the corresponding observations at AIRBASE sites (see Section 2.1). **External-validated (EV)** R^2, MSE-R^2, and RMSE were computed by comparing the predictions, which were derived from the main models, and the observations at all ESCAPE sites. In external validation, bias (mean difference between predictions and measurements) was additionally calculated for each model to evaluate the transferability of models.

We also evaluated model performance in subsets of ESCAPE sites, including areas with low air pollution concentrations (annual average concentration below 10, 12, 15, 20 and 25 $\mu g/m^3$ for $PM_{2.5}$; annual average concentration below 20, 30 and 40 $\mu g/m^3$ for NO_2), sites of different types (street, rural background and urban background), and different regions (north, west, central, and south). External-validated R^2, RMSE and bias were calculated for each evaluation.

The predicted concentrations can correlate poorly between models even if the models have similar explained variance (R^2). Therefore, for each pair of models, scatter plots of predicted concentrations at all ESCAPE sites were made to visually assess the correlations at these independent locations. Pearson correlation coefficients were also calculated. Scatter plots of the predictions versus observations at ESCAPE sites were also made for each of the models.

We compared the structure of models by looking at the number of predictors included in a model and the direction and magnitude of coefficients, where applicable. In linear regression models, regression slopes were multiplied by the difference between the 1st and 99th percentile of each predictor to allow comparison across predictors.

2.5. Sensitivity analysis

2.5.1. NO_2 models based on a reduced number of sites

The main NO_2 models were built on 2399 monitoring sites while the $PM_{2.5}$ models were developed on 543 sites. To separate the impacts of the number of training sites and the differences in pollutant

characteristics, we built additional NO_2 models using a random subset of 543 measurements extracted from all AIRBASE NO_2 monitoring sites (stratified by region and site type). The NO_2 sample models were developed and evaluated by the methods described above.

2.5.2. Models with a reduced number of potential predictors

Three variable selection methods were applied to explore the effect of the number of potential predictors offered. For each set of potential predictors derived from the methods described below, we arbitrarily selected two linear regression-based algorithms (SLR and Elastic Net ($\alpha = 0.75$)) and two machine learning algorithms (RF and ANN) to fit the models. The training R^2, CV R^2 and EV R^2 were calculated for each model.

Firstly, the predictor variables were ranked by their absolute correlation coefficients with pollutant concentrations, based on univariate correlation. In separate models, the first 80, 40 and 20 variables with the highest absolute correlation were used as potential predictor variables.

Secondly, the predictor variables were ranked by their variable importance, calculated as percentage increase in mean squared errors after a random permutation of the values of a variable, derived from the RF algorithm. In separate models, the first 80, 40 and 20 variables were used as potential predictor variables.

Finally, we reduced the number of buffers for road length and land use variables. Only variables with radii of 50 m, 100 m, 300 m, 500 m, 1000 m, 2000 m, 5000 m, 10,000 m were offered as potential predictors, resulting in a total of 64 predictors.

2.5.3. $PM_{2.5}$ models with kriging

In the recently published Europe-wide modelling study (de Hoogh et al., 2018a), kriging proved an efficient technique to explain a part of residual spatial variation for the $PM_{2.5}$ SLR model. To examine whether the residual variation explained by kriging had been captured by a more flexible algorithm, we performed kriging on the residuals from the selected $PM_{2.5}$ models (BLR, SLR, LASSO, GBM and ANN). Ordinary kriging was applied to the residuals of background sites only, and added to the pollution estimates of the models. Models were evaluated by the metrics described in Section 2.4. Scatter plots comparing the predictions at all ESCAPE sites were made, and Pearson correlation coefficients were calculated.

All statistical analyses were conducted in R v 3.4.1(Team, 2013).

3. Results

3.1. $PM_{2.5}$ models

All models had moderate to good performance when evaluated by cross-validation (CV), with CV R^2s ranging from 0.48 to 0.63, and CV RMSEs ranging from 3.1 to 3.9 μg/m^3 (Table 1). CV MSE-R^2s were similar as CV R^2s. Higher CV R^2s and lower CV RMSEs were found for machine learning models based on decision trees (the GBM, the bagging, and the RF). The lowest CV R^2 and the highest CV RMSE were found for the ANN model. Among all linear regression-based models, the BLR model had the lowest CV R^2 and the highest CV RMSE, while it had the highest training R^2 among these linear models.

Model performance measured by external validation (EV) showed good agreement with the results measured by CV, though less of the variation (R^2) in the external data was explained (Table 1). MSE-R^2s were on average 5% lower than the R^2s. The decision tree-based ensemble models performed moderately better than others whilst the BLR, the SVR, and the ANN models performed moderately worse. Biases were lower than 1 μg/m^3 for all models, except the ANN. The better performance of the decision tree-based models disappeared when restricting validation dataset to sites with low $PM_{2.5}$ concentrations (Table S4). For all algorithms, validation R^2 decreased and bias increased when restricting to lower pollution levels. Similar differences in model performance across algorithms were observed for street and urban background sites (Table S5). For all algorithms, bias was higher for background sites than for street sites. The pattern between algorithms was similarly when evaluated at regional scale (Table S6). We

Table 1
Performance of $PM_{2.5}$ spatial models using different model building algorithms.

Algorithm[a]	Training		Cross validation			External validation			
	(N = 543)		(*N* = 543)			(*N* = 416)			
	R^2	RMSE[b] (μg/m^3)	R^2	MSE-R^2	RMSE (μg/m^3)	R^2	MSE-R^2	RMSE (μg/m^3)	Bias (μg/m^3)
FLR	0.657	3.0	0.600	0.598	3.3	0.517	0.481	4.1	0.7
BLR	0.704	2.8	0.506	0.472	3.7	0.463	0.445	4.3	0.8
WLR	0.657	3.0	0.600	0.598	3.3	0.517	0.481	4.1	0.7
SLR	0.622	3.2	0.595	0.594	3.3	0.529	0.478	4.1	0.9
Ridge	0.665	3.0	0.592	0.592	3.3	0.535	0.485	4.1	0.7
EN25	0.643	3.1	0.608	0.607	3.2	0.545	0.483	4.1	0.8
EN50	0.642	3.1	0.609	0.608	3.2	0.546	0.486	4.1	0.8
EN75	0.641	3.1	0.609	0.609	3.2	0.547	0.486	4.1	0.8
LASSO	0.641	3.1	0.610	0.609	3.2	0.547	0.487	4.1	0.8
GAM	0.652	3.0	0.608	0.608	3.2	0.557	0.498	4.1	0.9
Bagging	0.954	1.2	0.627	0.626	3.1	0.575	0.531	3.9	0.4
RF	0.955	1.2	0.626	0.624	3.1	0.583	0.530	3.9	0.4
GBM	0.895	1.8	0.631	0.630	3.1	0.610	0.548	3.9	0.4
SVR	0.799	2.3	0.569	0.568	3.4	0.457	0.432	4.3	0.3
KRLS	0.726	2.7	0.590	0.586	3.3	0.525	0.466	4.2	0.6
ANN	0.723	2.7	0.477	0.428	3.9	0.391	0.286	4.8	1.2
Ensemble1	0.698	2.8	0.618	0.617	3.2	0.553	0.495	4.1	0.7
Ensemble2	0.762	2.6	0.622	0.622	3.2	0.573	0.513	4.0	0.7

[a] FLR = Forward stepwise Linear Regression; BLR = Backward stepwise Linear Regression; WLM = Stepwise Linear Regression; SLR = Supervised Linear Regression; EN25 = Elastic Net with α = 0.25; EN50 = Elastic Net with α = 0.50; EN75 = Elastic Net with α = 0.75; LASSO = Least Absolute Shrinkage and Selection Operator; GAM = Generalized Additive Model; RF = Random Forest; GBM = Generalized Boosted Machine; SVR = Support Vector Regression; KRLS = Kernel-based Regularized Least Squares; ANN = Artificial Neural Network; Ensemble1 = Ensemble model based on median prediction; Ensemble2 = Ensemble model based on weighted average.

[b] RMSE = Root-mean-square error.

noted only small differences of both ensemble models performances compared to the best individual models.

Fig. 1 shows the scatter plots of $PM_{2.5}$ predictions at all ESCAPE sites by each pair of models. Most of the predicted concentrations were highly correlated, with correlation coefficients above 0.85. Almost identical predictions were found for several models, such as the LASSO and three Elastic Net models, the FLR and the WLR models, as well as the Bagging and the RF models. Predictions by the ANN model and other models were less correlated. All models tended to overpredict at low concentrations and underpredict at high concentrations (Fig. S2).

For linear regression-based models (except for the Ridge regression model), the number of predictors included in the main model, the direction and magnitude of regression slope are summarized in Fig. 2 (details shown in Table S7). The SLR model included the lowest number of predictors (7) in the model while the BLR retained the highest number (48). All models included CTM and SAT estimates as well as all roads, natural areas, ports, residential areas, and altitude as predictors. CTM and SAT estimates were positively correlated with $PM_{2.5}$ in all linear regression models, while altitude was always negatively correlated with $PM_{2.5}$. All models except the SLR model included predictors

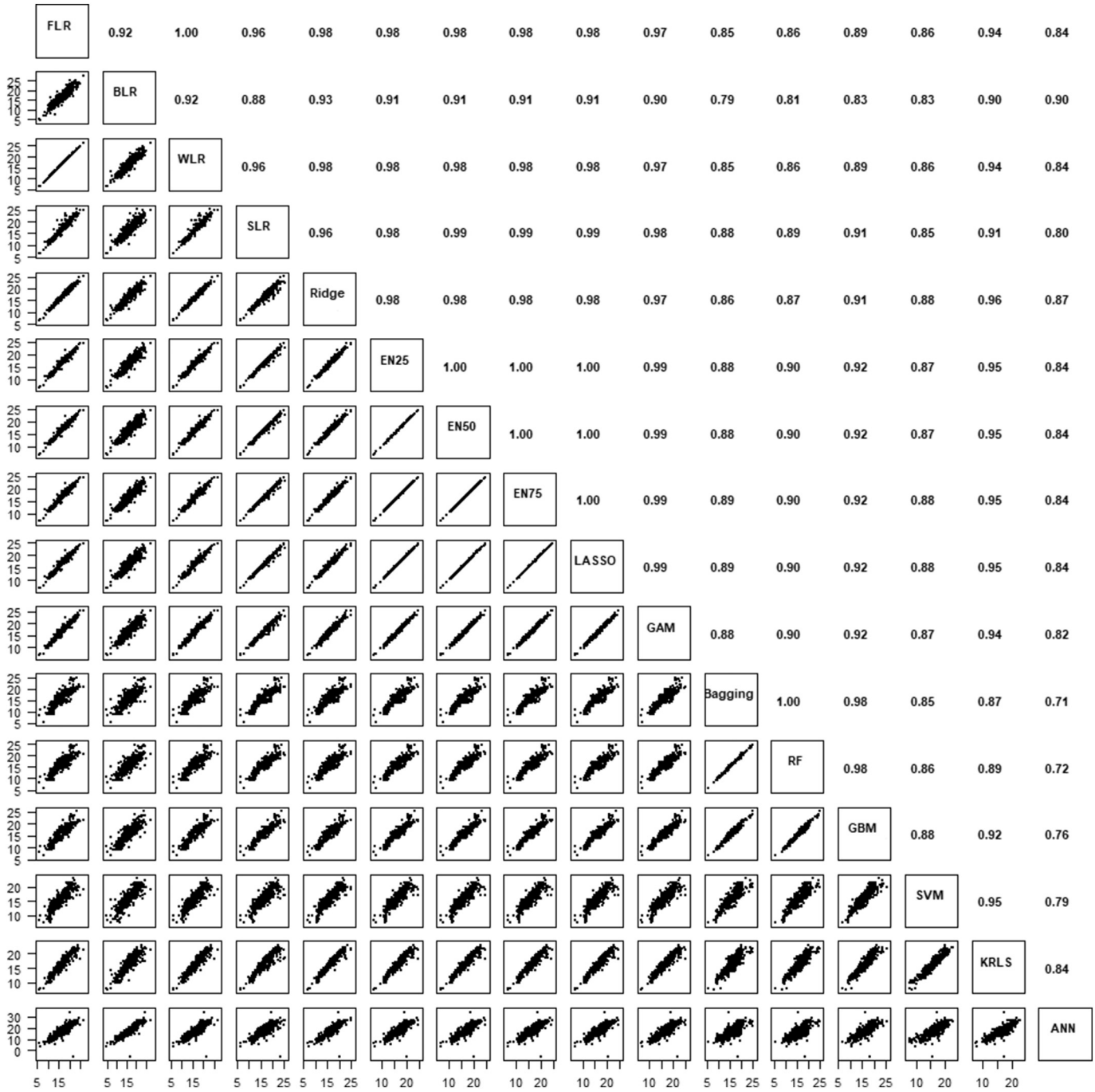

Fig. 1. Correlations between $PM_{2.5}$ predictions at ESCAPE sites. The upper triangle shows the correlation coefficients, the lower triangle shows the scatter plots. FLR = Forward stepwise Linear Regression; BLR = Backward stepwise Linear Regression; WLM = Stepwise Linear Regression; SLR = Supervised Linear Regression; EN25 = Elastic Net with $\alpha = 0.25$; EN50 = Elastic Net with $\alpha = 0.50$; EN75 = Elastic Net with $\alpha = 0.75$; LASSO = Least Absolute Shrinkage and Selection Operator; GAM = Generalized Additive Model; RF = Random Forest; GBM = Generalized Boosted Machine; SVR = Support Vector Regression; KRLS = Kernel-based Regularized Least Squares; ANN = Artificial Neural Network.

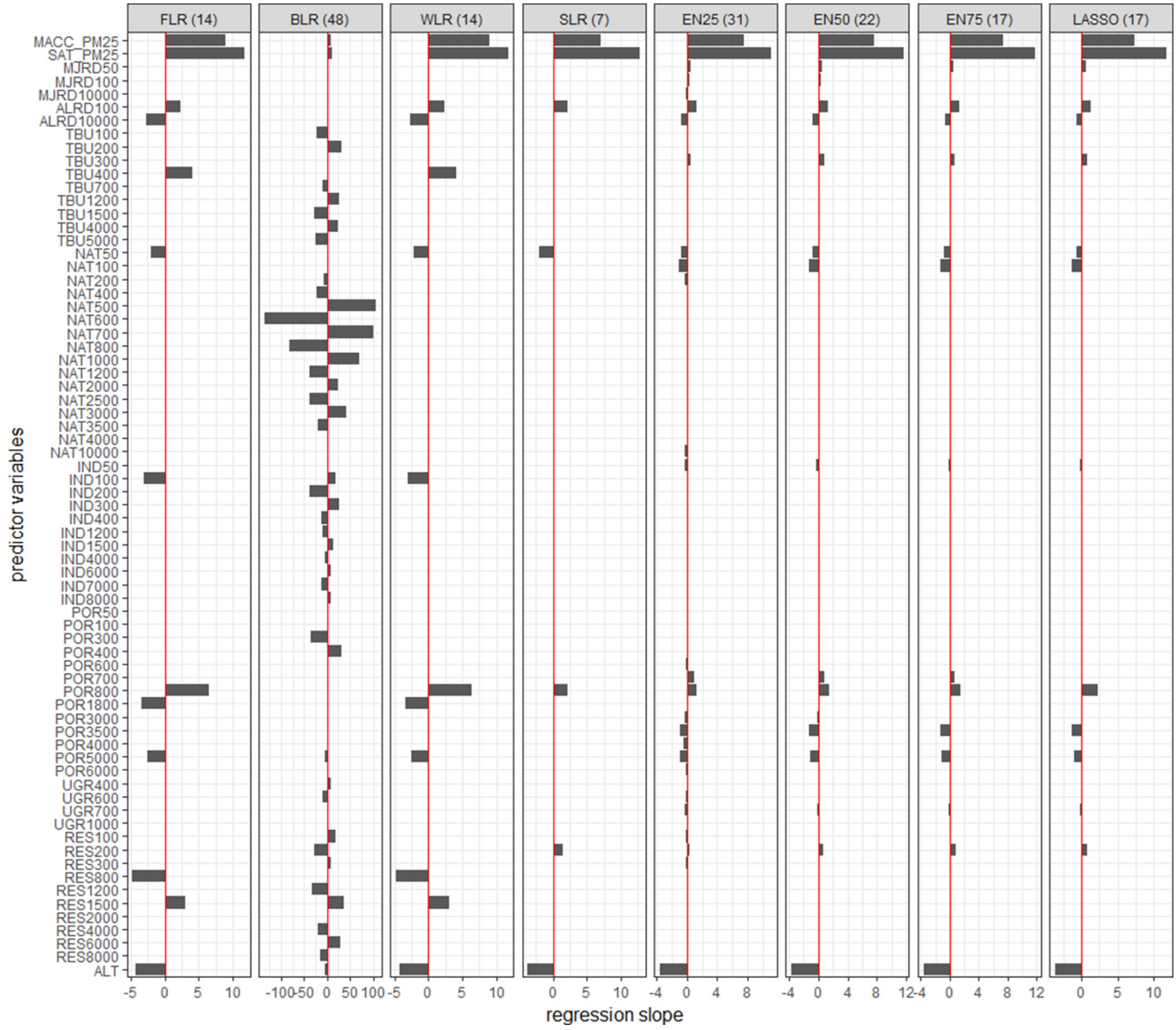

Fig. 2. Predictors included in linear $PM_{2.5}$ models. The figures in the blanket show the number of predictors included in each model.
Regression slopes were multiplied by the difference between the 1st and 99th percentiles of each predictor to allow comparison across predictors
MACC = MACC dispersion model, SAT = Satellite-derived, MJRD = Major Roads, ALRD = All Roads, TBU = Total Build Up, NAT = Natural Land, IND = Industry, POR = Ports, UGR = Urban Green, RES = Residential, ALT = Altitude
FLR = Forward stepwise Linear Regression; BLR = Backward stepwise Linear Regression; WLM = Stepwise Linear Regression; SLR = Supervised Linear Regression; EN25 = Elastic Net with α = 0.25; EN50 = Elastic Net with α = 0.50; EN75 = Elastic Net with α = 0.75; LASSO = Least Absolute Shrinkage and Selection Operator.

with counterintuitive direction of slope, such as negative slopes for road length and ports. CTM and SAT estimates were consistently two of the most important predictors in all models, as shown by the large regression slopes, except in the BLR model. CTM and SAT estimates were also identified as the most important predictors in variable importance plots from the RF model and the GBM model, followed by altitude (Fig. S4). A rapid drop in variable importance was observed after CTM and SAT estimates.

3.2. NO_2 models

Table 2 shows the performance of the different NO_2 models. Though the non-linear models had higher training R^2s than the linear regression-based models, all models had similar performances when measured by cross-validation (CV R^2 0.57 to 0.62, CV RMSE 9.0 to 9.6 μg/m^3), and when measured by external validation (EV R^2 0.49 to 0.51, EV RMSE 11.6 to 14.6 μg/m^3). Biases were low (1.2 to 3.3 μg/m^3) for all models except the ANN (8.8 μg/m^3). Model performance also did not vary much across algorithms when restricting validation subsets to less polluted sites (Table S8) and specific type of sites (Table S9). For all algorithms, validation R^2 decreased and bias increased when restricting to lower pollution levels (Table S8). For all algorithms, validation R^2 was lower for street sites compared to background sites. A small negative bias was observed for street sites and a more substantial positive bias for background sites, again with small differences across algorithms. All algorithms performed similarly when evaluated at regional scale. Only small differences were found for both ensemble models performances compared to the best individual models.

The predictions at the ESCAPE sites by each pair of models were highly correlated, with Pearson correlation coefficients ranging from

Table 2
Performance of NO_2 spatial models using different model building algorithms.

Algorithm[a]	Training		Cross validation			External validation			
	(N = 2399)		(*N* = 2399)			(*N* = 1396)			
	R^2	RMSE[b]	R^2	MSE-R^2	RMSE	R^2	MSE-R^2	RMSE	Bias
		(μg/m^3)			(μg/m^3)			(μg/m^3)	(μg/m^3)
FLR	0.596	9.3	0.584	0.583	9.4	0.499	0.485	11.6	1.3
BLR	0.614	9.1	0.573	0.571	9.5	0.496	0.481	11.6	1.8
WLR	0.596	9.3	0.584	0.583	9.4	0.499	0.485	11.6	1.3
SLR	0.588	9.4	0.575	0.575	9.5	0.495	0.468	11.8	2.5
Ridge	0.606	9.2	0.586	0.586	9.4	0.500	0.471	11.7	2.5
EN25	0.605	9.2	0.588	0.588	9.4	0.504	0.483	11.6	2.0
EN50	0.606	9.2	0.588	0.588	9.4	0.505	0.485	11.6	1.9
EN75	0.606	9.2	0.588	0.588	9.4	0.505	0.485	11.6	1.9
LASSO	0.606	9.2	0.588	0.588	9.4	0.505	0.485	11.6	1.9
GAM	0.639	8.8	0.609	0.609	9.1	0.506	0.486	11.6	2.3
Bagging	0.950	3.6	0.612	0.612	9.1	0.490	0.449	12.0	3.2
RF	0.951	3.6	0.613	0.612	9.1	0.487	0.444	12.0	3.3
GBM	0.807	6.5	0.621	0.621	9.0	0.499	0.471	11.7	2.7
SVR	0.708	8.0	0.607	0.601	9.2	0.492	0.481	11.6	1.2
KRLS	0.687	8.2	0.613	0.613	9.1	0.505	0.480	11.6	2.4
ANN	0.623	9.0	0.570	0.568	9.6	0.488	0.181	14.6	8.8
Ensemble1	0.628	8.9	0.597	0.60	9.3	0.509	0.49	11.6	2.1
Ensemble2	0.706	8.0	0.611	0.61	9.1	0.518	0.49	11.5	2.5

[a] FLR = Forward stepwise Linear Regression; BLR = Backward stepwise Linear Regression; WLM = Stepwise Linear Regression; SLR = Supervised Linear Regression; EN25 = Elastic Net with α = 0.25; EN50 = Elastic Net with α = 0.50; EN75 = Elastic Net with α = 0.75; LASSO = Least Absolute Shrinkage and Selection Operator; GAM = Generalized Additive Model; RF = Random Forest; GBM = Generalized Boosted Machine; SVR = Support Vector Regression; KRLS = Kernel-based Regularized Least Squares; ANN = Artificial Neural Network; Ensemble1 = Ensemble model based on median prediction; Ensemble2 = Ensemble model based on weighted average.
[b] RMSE = Root-mean-square error.

0.91 to 1.00 (Fig. 3). All models tended to overpredict at low concentrations and underpredict at high concentrations (Fig. S3).

Even though the NO_2 model predictions were similar, their structures were different (Fig. 4 and Table S11). The SLR model included the lowest number of predictors (8), while the EN25 model included the highest number (55). CTM estimates were positively associated with NO_2 in all models. All models included SAT estimates with a counterintuitive negative slope, except the SLR which did not include SAT at all. Counterintuitive slopes were also found for road length, ports, natural areas and residential areas in BLR model and regularization models. The variable importance plots derived from the RF and the GBM indicate that the CTM estimates and road variables were strong predictors (Fig. S5).

3.3. Sensitivity analysis

3.3.1. NO_2 models based on a reduced number of sites

The performances of NO_2 models built on a subset of 543 sites (the number of $PM_{2.5}$ sites) are summarized in Table S12. Compared to NO_2 models built on all 2399 sites, the CV R^2s were virtually the same, while the EV R^2s were about 0.05 lower. More variation in the model performances across different algorithms was found compared to the original NO_2 models. The ANN model performed the most poorly when evaluated by CV. The BLR and the ANN performed moderately worse than other models when evaluated by EV. Other models performed similarly when comparing CV and EV results.

3.3.2. Models with a reduced number of potential predictors

For $PM_{2.5}$ models, the training R^2s, CV R^2s and EV R^2s were relatively stable in relation to the number of potential predictors offered to fit the SLR, the EN75 and the RF algorithms (Fig. S6). The ANN model had lower training R^2 but higher CV R^2 and EV R^2 when fitted with fewer potential predictors.

NO_2 models fitted with different algorithms show consistent patterns in relation to the number of potential predictors offered (Fig. S7). The plots on the left show the relatively poor performance of NO_2 models built with only 20 predictors selected based on univariate correlation, where NO_2 CTM estimates ranked 30 and were not included in the first 20 predictors. The model training R^2s, CV R^2s and EV R^2s all increased rapidly when the number of potential predictors offered increased from 20 to 40. The training R^2s, CV R^2s and EV R^2s further increased mildly when 80 potential predictors were offered. In contrast, the training R^2s, CV R^2s and EV R^2s were not affected by the number of potential predictors offered when variables were selected based on RF or a priori reduction of the number of buffers.

3.3.3. $PM_{2.5}$ models with kriging

The kriging technique performed on residuals further increased the CV R^2s and EV R^2s of the linear models (BLR, SLR, LASSO) by 4.0% and 12.4% on average, while it increased less for the GBM and ANN models (the CV R^2 and EV R^2 of the GBM model increased by 0.8% and 3.8% respectively, the CV R^2 and EV R^2 of the ANN model increased by 2.3% and 6.1% respectively) (Table S13). The correlation coefficients of model predictions at ESCAPE sites were both 0.90 for SLM + kriging and GBM + kriging, and LASSO + kriging and GBM + kriging (Fig. S8).

4. Discussion

We compared 16 algorithms to develop Europe-wide models predicting annual average $PM_{2.5}$ and NO_2 concentrations in 2010. For both validation methods, $PM_{2.5}$ models developed on 543 sites performed similarly across algorithms, though models developed with the generalized boosted machine, random forest and bagging performed slightly better than others in the full datasets. The $PM_{2.5}$ predictions at all ESCAPE sites derived from different models were highly correlated, except for predictions from the artificial neural network. For both validation methods, NO_2 models developed on 2399 sites performed even more similarly across different algorithms. The NO_2 predictions at external sites were all highly correlated. For both pollutants, low biases

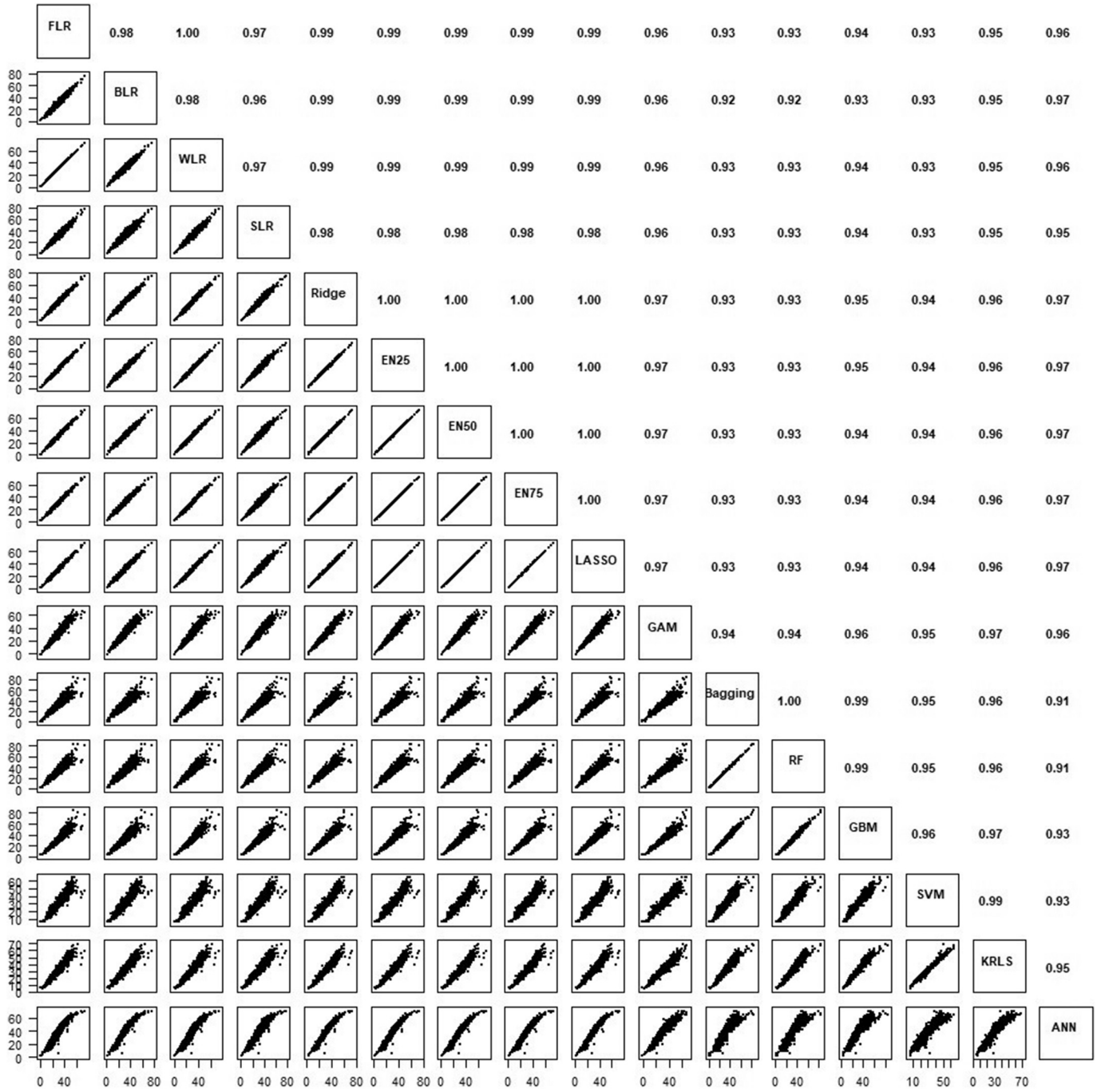

Fig. 3. Correlations between NO_2 predictions at ESCAPE sites. The upper triangle shows the correlation coefficients, the lower triangle shows the scatter plots. FLR = Forward stepwise Linear Regression; BLR = Backward stepwise Linear Regression; WLM = Stepwise Linear Regression; SLR = Supervised Linear Regression; EN25 = Elastic Net with $\alpha = 0.25$; EN50 = Elastic Net with $\alpha = 0.50$; EN75 = Elastic Net with $\alpha = 0.75$; LASSO = Least Absolute Shrinkage and Selection Operator; GAM = Generalized Additive Model; RF = Random Forest; GBM = Generalized Boosted Machine; SVR = Support Vector Regression; KRLS = Kernel-based Regularized Least Squares; ANN = Artificial Neural Network.

were found when different models were applied on all ESCAPE sites, except for the ANN models.

4.1. Predictive ability

Our study found small differences in performance and in predictions at all external sites derived from different algorithms. The algorithms identified the same key predictor variables. The small differences between algorithms may be the result of the large number of training sites, the use of relatively stable annual average concentrations to develop models and the lack of complex relationships between predictors and annual average concentrations.

Previous algorithm comparison studies were based on either a smaller number of sites (Brokamp et al., 2017) or on mobile monitoring with much more variation in the measured concentration data (Kerckhoffs et al., 2019; Van den Bossche et al., 2018; Weichenthal et al., 2016). In our study, all algorithms may have the advantage of a lower risk of overfitting because of the combination of a large number of training sites and stable annual average concentrations. The NO_2 models built on 2399 sites performed more similarly across algorithms

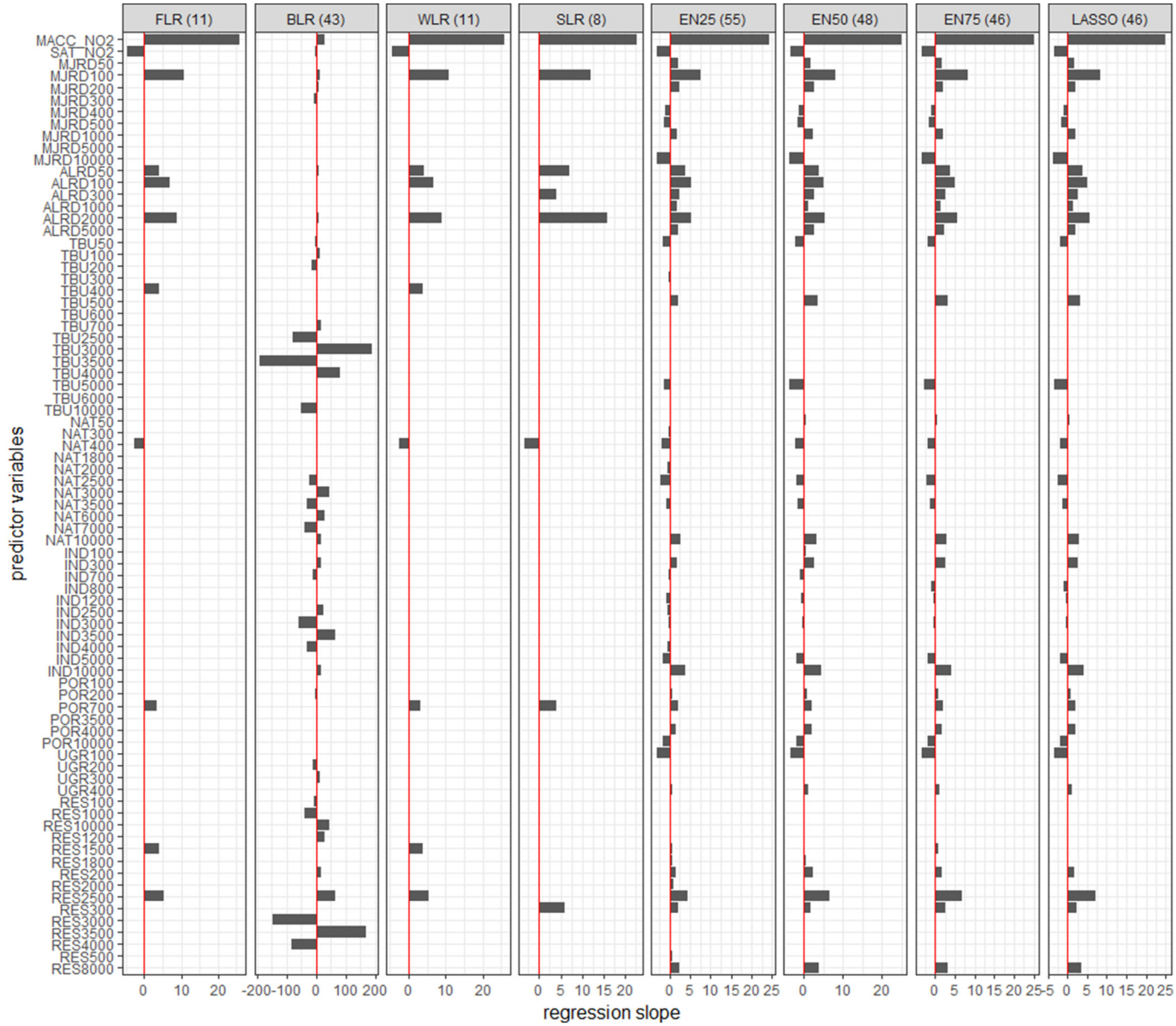

Fig. 4. Predictors included in linear NO_2 models. The figures in the blanket show the number of predictors included in each model.
Regression slopes were multiplied by the difference between the 1st and 99th percentiles of each predictor to allow comparison across predictors.
MACC = MACC dispersion model, SAT = Satellite-derived, MJRD = Major Roads, ALRD = All Roads, TBU = Total Build Up, NAT = Natural Land, IND = Industry, POR = Ports, UGR = Urban Green, RES = Residential, ALT = Altitude. FLR = Forward stepwise Linear Regression; BLR = Backward stepwise Linear Regression; WLM = Stepwise Linear Regression; SLR = Supervised Linear Regression; EN25 = Elastic Net with $\alpha = 0.25$; EN50 = Elastic Net with $\alpha = 0.50$; EN75 = Elastic Net with $\alpha = 0.75$; LASSO = Least Absolute Shrinkage and Selection Operator.

than the $PM_{2.5}$ models built on 543 sites, which could be due to the fact that the NO_2 models were built on four times more training sites. This is supported by the sensitivity analysis where the performance of NO_2 models built on 543 sites (the same number of training sites available for $PM_{2.5}$) varied more across algorithms than the original NO_2 models built on 2399 sites.

The machine learning algorithms did not perform better in our study. However, their ability to model complex relationships among the data is a clear benefit in studies modelling spatiotemporal variations of air pollution, where the variability in concentration is often larger and the relationships between pollution concentration and predictors are more complicated (e.g. nonlinear relationships between pollution and satellite data exist under different meteorological conditions and emission features (Liu et al. 2009; Schaap et al., 2009)). The RF algorithm was fitted to assess spatiotemporal patterns of air pollution in Japan (Araki et al., 2018), the United States (Hu et al., 2017) and China (Zhan et al., 2018). The neural network algorithm was trained to predict daily $PM_{2.5}$ concentrations over the continental United States from 2000 to 2012 (Di et al., 2016a). The SVR algorithm was used in recent European studies to estimate daily $PM_{2.5}$ concentrations across Switzerland (de Hoogh et al., 2018b) and daily PM_{10} concentrations in Italy (Stafoggia et al., 2017). The GBM algorithm outperformed 10 other algorithms, including linear-regression based algorithms, to model spatiotemporal variation of $PM_{2.5}$ concentrations during a wildfire (Reid et al., 2015). In the Reid et al. algorithm comparison study, the larger variability in pollution concentrations and the more complicated relationships between predictors and pollution may explain why the more sophisticated algorithms, which are able to model nonlinearity and complex interactions, outperformed the simple algorithms. However, if there are no strong nonlinear relationships or complex

interactions in the data, as we have good reasons to assume is the case in our data based on annual average spatial variation, the more sophisticated algorithms do not add to the simple linear regression-based algorithms. In a recent algorithm comparison study trained on mobile and short-term measurements (Kerckhoffs et al., 2019), differences in performance evaluated by external long-term exposure estimates were also small. The investigators used spatial average concentrations for 368 short-term sites and over 8000 road segments, resulting in much less stable estimates of site-specific averages probably due to the shorter sampling time.

We noted only small differences of both ensemble models performances compared to the best individual models. The fact that ensemble models did not improve upon the individual algorithms in our setting could be due to the similar performances and highly correlated predictions across algorithms. Ensemble models are attractive because the researcher does not have to make (arbitrary) choices of what model to choose for final exposure assignment. We used global weights for the different algorithms. If there is evidence for spatially different performance of the different algorithm, spatially varying weights could be used.

Comparing $PM_{2.5}$ and NO_2 models built on the same number of training sites, the ensemble learning algorithms (Bagging, RF and GBM) performed slightly better than other algorithms for $PM_{2.5}$ but not for NO_2. We speculate that this might be due to the different characteristics of $PM_{2.5}$ and NO_2. $PM_{2.5}$ concentrations vary at large regional scales (Eeftens et al., 2012), whereas NO_2 concentrations, strongly influenced by local traffic emissions, vary more widely at smaller scales (Cyrys et al., 2012). The ensemble learning algorithms modelled detailed fluctuations of the measurements – as indicated by the extremely high training R^2s, which may not transfer to the validation dataset for NO_2.

The results of the GAM models might not be comparable with other models because of the slightly different input variables. However, we would not expect much deviation from the current results as none of the deleted small buffer land use variables was identified as important predictors in other models. The ANN models as specified in the current study did not perform well among the algorithms. One possible reason is that the large number of predictors and relatively small number of observations in the training dataset required more careful training. Our sensitivity analysis also supported that with less potential predictors or more training data, the ANN algorithm tended to perform better. In this study, we used one hidden layer to build ANN models because, as suggested by Schalkoff (1997), one hidden layer is sufficient for avoiding overfitting in most applications. However, we cannot rule out the possibility of better performance by using more than one hidden layer.

Kriging is a technique which can be used to explain spatial variation within the data. In the sensitivity analysis, kriging on residual variation did not significantly improve the performance for $PM_{2.5}$ GBM and ANN models, indicating the machine learning algorithms have some ability to address spatial autocorrelation in air pollution concentrations. The CV-R^2 of SLR, LASSO and GBM models became closer after adding kriging surfaces, indicating part of the residual variance of the SLR and LASSO models explained by kriging had been accounted for by a more flexible GBM algorithm. For NO_2, kriging did not explain the residual spatial variation (de Hoogh et al., 2018a).

4.2. Model structure and interpretation

Although our main interest is in the predictive performance of the models, it is informative to interpret the structure of the models. The importance of specific determinants such as traffic in the model may be helpful to compare risks across epidemiological studies in different areas.

The machine learning algorithms are often considered “black boxes” (Zhang and Ding, 2017) since the models derived from these algorithms are difficult to interpret. Even though some algorithms provide variable importance measures, such as the RF and the GBM (Breiman et al., 2011; Ridgeway et al., 2013), the magnitude and direction of the predictor effects are unknown. Models built with linear stepwise regression and regularization algorithms are easier to interpret, both in terms of included predictors and the magnitude and direction of predictor effects. An overview of the applied algorithms is shown in Table 3.

For both pollutants, the SLR models by definition included only predictors following the plausible direction of effect, resulting in a substantially smaller number of predictors than other models. Models developed with all other algorithms included predictors with counter-intuitive directions of effect, though in most cases not for the key predictor variables. For example in the NO_2 models, satellite NO_2 was

Table 3
Overview of algorithms as applied in this study.

Algorithm[a]	Group	Model possible nonlinear relationships between pollutant and predictors	Model possible interactions among predictors	Variable selection	Computation time (mins)[b]	Model structure
FLR	Linear stepwise regression algorithms	No; a priori transformations can be offered e.g. inverse distance	No; selected product terms can be added in principle	Yes	< 1	Showed magnitude and direction of predictor effects; biased coefficient estimates when predictors are highly correlated (except SLR, which excluded highly correlated predictors in a model)
BLR					4	
WLR					< 1	
SLR					< 1	
Ridge	Regularization or shrinkage algorithms	No; a priori transformations can be offered e.g. inverse distance	No; selected product terms can be added in principle	Yes, except for Ridge regression	< 1	Showed magnitude and direction of predictor effects; reliable coefficient estimates even when predictors are highly correlated
EN25					< 1	
EN50					< 1	
EN75					< 1	
LASSO					< 1	
GAM	Generalized Additive Model algorithm	Yes	No	Yes, not in default method	2313	Difficult to interpret with multiple predictors
Bagging	Machine learning algorithms	Yes	Yes	No	41	Difficult to interpret, though RF and GBM provide variable importance measures
RF					96	
GBM					66	
SVR					8	
KRLS					5	
ANN					6	

[a] FLR = Forward stepwise Linear Regression; BLR = Backward stepwise Linear Regression; WLM = Stepwise Linear Regression; SLR = Supervised Linear Regression; EN25 = Elastic Net with $\alpha = 0.25$; EN50 = Elastic Net with $\alpha = 0.50$; EN75 = Elastic Net with $\alpha = 0.75$; LASSO = Least Absolute Shrinkage and Selection Operator; GAM = Generalized Additive Model; RF = Random Forest; GBM = Generalized Boosted Machine; SVR = Support Vector Regression; KRLS = Kernel-based Regularized Least Squares; ANN = Artificial Neural Network.

[b] Computation time was recorded for $PM_{2.5}$ models developed on a standard office computer.

included with a negative slope possibly to compensate for over-prediction by the other large spatial scale predictor variable (CTM estimates). Restricting the inclusion of predictors did not affect model performance in our study and may be considered more a philosophical choice (Brauer et al., 2003; Vienneau et al., 2010). However, model prediction may deviate when the models are applied in another domain than the Europe-wide training domain such as a smaller area within Europe. This is supported by our subset validation where the NO_2 SLR model outperformed automatic variable selection linear models at rural background sites.

In our study, correlated potential predictors were offered to build linear-regression based models. These predictors are usually the same land use feature/ road length in buffers with different radii. Offering highly correlated predictors would lead to incorrect selection of predictors in a model (Agier et al., 2016), and including highly correlated predictors in an ordinary least squares-based model would lead to biased coefficient estimates (Tibshirani et al., 2013). Clear evidence for this is found in the backward stepwise selection algorithm, which included very different predictor variables than all other algorithms, and had a lower explained variance in the validation dataset. The SLR algorithm deals with collinearity by deleting predictors with a variance inflation factor larger than 3, at the expense of including fewer predictor variables. The regularization algorithms impose a penalty to shrink the coefficients of the least informative predictors towards zero, which have been shown to be more efficient in identifying correct predictors than the ordinary least squares-based algorithms (Agier et al., 2016). Compared to the SLR algorithm, regularization algorithms included more buffers of the same variable, which is consistent with the notion of smooth changes of pollution with increasing buffer size: if a road length variable contributes to the pollution estimate, one would expect road length variables with other buffer sizes also to add to the pollution estimate. However, interpretation is hampered because the same land use feature/road length with different radii entered the model with both plausible and implausible directions of effect.

4.3. Strengths and limitations

One strength of our study is that we compared multiple algorithms with very different assumptions. Most of the previous algorithm comparison studies only compared two or three algorithms (Basagaña et al., 2012; Brokamp et al., 2017; Van den Bossche et al., 2018; Weichenthal et al., 2016), therefore results between studies are difficult to compare. Studies have assessed more algorithms in different settings such as assessing spatiotemporal variation of $PM_{2.5}$ during a wildfire (Reid et al., 2015), or assessing spatial variation of ultrafine particles using mobile monitoring data (Kerckhoffs et al., 2019). Our study gives new insight into the predictive ability of these algorithms because the number of training sites, variation in the monitoring data and complexity of relationships within the data can all affect relative performance of the algorithms.

Secondly, we used both cross-validation and external validation to strengthen our comparison. CV is commonly used to evaluate model performance in air pollution exposure assessment (Kim et al., 2016; Liu et al., 2009). Because CV is restricted to the monitoring domain, a good performance evaluated by CV does not necessarily mean the model can accurately predict residential exposure, which is often used as surrogates for long-term exposure of subjects in epidemiological studies (Beelen et al., 2014; Di et al., 2017). Concerns about overestimating exposures at residential addresses have been raised for models based upon on-road mobile monitoring (Hoek, 2017; Kerckhoffs et al., 2016). In this situation, CV alone cannot reflect true transferability because the mobile monitoring sites are located on roads and the measurements usually have significant variations. Validation of such models using an external dataset reflecting residential exposure is important (Kerckhoffs et al., 2019). In our study, monitoring data from the regulatory AIRBASE network were used to develop air pollution models. With the aim to check for compliance with the European Union air quality standards, the monitors in the AIRBASE network have not primarily been located with the goal of characterizing residential exposure in mind. On the other hand, the location of ESCAPE sites was purposely selected to be representative for air pollution exposure at home addresses of study subjects. In this sense, results of our EV using pollution data from the ESCAPE study reflects the transferability of models in application. Besides the choice of monitoring locations, there are also a number of differences between the two datasets that could impact the evaluation. Firstly, ESCAPE monitoring sites do not cover the same geographical area as the AIRBASE monitoring sites. Secondly, the two datasets differed in measurement techniques. A comprehensive comparison showed limited systematic differences between the ESCAPE and AIRBASE methods for NO_2 (Cyrys et al., 2012). NO_2 was mainly measured by chemiluminescence in AIRBASE versus Ogawa badges in ESCAPE. Chemiluminescence is subject to interference from other reactive nitrogen species that can vary spatially (Suzuki et al., 2011), which could, at least partially, explain the higher bias at low concentration and background ESCAPE sites. For PM2.5 no methods comparison was made. Thirdly, the ESCAPE measurements are based on 2-week sampling in three seasons with temporal adjustment versus continuous measurements in AIRBASE. Lastly, ESCAPE measurements were conducted following the same standard measurement methods and strict quality control procedures, while the AIRBASE were measured by inconsistent methods and reported with different frequency across countries. Nevertheless, the agreement in results showed by CV and EV help us to strengthen the comparison.

Thirdly, we used grid search to optimize hyperparameters for some algorithms before fitting the models, based on the best performance in 5-fold cross-validation. This approach helped to minimize the risk of overfitting and avoid overly optimistic performance estimates (Van den Bossche et al., 2018). This adds to a recent algorithm comparison performed by Kerckhoffs et al. (2019) which used default parameters for algorithms. Though the grid search approach added to the computation load in model development for the GAM and machine learning algorithms, the computation time is generally short for all algorithms, and was not a decisive factor for choosing one algorithm over another in this moderately small dataset (Table 3).

Despite similarities across algorithms, the model performances remained moderate in our study for both pollutants. This is probably a result of missing explanatory variables. A previous European $PM_{2.5}$ model developed using supervised linear regression algorithm obtained a CV R^2 of 0.80 (Wang et al., 2014), which was driven by the inclusion of local traffic intensity and measured regional background concentrations. Such variables are often only available at local level, and were not available in the current study. The relevant influence of missing explanatory variables is also supported by our sensitivity analysis where performance of different NO_2 models all reduced dramatically when the CTM estimates were not offered as a potential predictor. Additionally, the overall model performance we evaluated in the study could be dominated by regional variation and might not reflect within-city variations well. Decreased and varying R^2s were reported when the SLR model was validated at individual study areas (de Hoogh et al., 2018a). However, the overall model performance was our primary interest as the models developed in this study were aimed to estimate air pollution exposure for participants across Europe. Our sensitivity analyses showed moderate performances when models were validated at regional scale (Tables S6 and S10). Another limitation is that none of the algorithms explicitly include handling spatial autocorrelation in air pollution concentrations in model building. We addressed this with a 2-step kriging sensitivity analysis. Our results showed that kriging significantly improved model performance of linear regression algorithms, and suggest that the machine learning algorithms have some ability to handle autocorrelation within data. Our 2-step kriging approach likely underestimates the value of including spatial autocorrelation (Mercer et al., 2011). The predictive power of these models might be further

improved by applying universal kriging, which can be seen as a LUR with addition of correlated residuals.

4.4. Conclusion

Different statistical algorithms performed similarly when modelling spatial variation of annual average air pollution concentrations using a large number of training sites.

The results of our study and the previous algorithm comparisons suggest that the relative performance of algorithms may differ with the study setting, therefore generic recommendations for one algorithm cannot be made. To take appropriate decisions for a particular study, future studies may opt for models developed using more than one algorithm.

Acknowledgements

Research described in this article was conducted under contract to the Health Effects Institute (HEI), an organization jointly funded by the United States Environmental Protection Agency (EPA) (Assistance Award No. R-82811201) and certain motor vehicle and engine manufacturers. The contents of this article do not necessarily reflect the views of HEI, or its sponsors, nor do they necessarily reflect the views and policies of the EPA or motor vehicle and engine manufacturers. This work was also supported by a scholarship under the State Scholarship Fund by the China Scholarship Council (File No. 201606010329).

Declaration of Competing Interest

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envint.2019.104934.

References

Agier, L., Portengen, L., Chadeau-Hyam, M., et al., 2016. A systematic comparison of linear regression–based statistical methods to assess exposome-health associations. Environ. Health Perspect. 124, 1848.
Araki, S., Shima, M., Yamamoto, K., 2018. Spatiotemporal land use random forest model for estimating metropolitan no 2 exposure in Japan. Sci. Total Environ. 634, 1269–1277.
Basagaña, X., Rivera, M., Aguilera, I., et al., 2012. Effect of the number of measurement sites on land use regression models in estimating local air pollution. Atmos. Environ. 54, 634–642.
Bechle, M.J., Millet, D.B., Marshall, J.D., 2013. Remote sensing of exposure to no 2: satellite versus ground-based measurement in a large urban area. Atmos. Environ. 69, 345–353.
Bechle, M.J., Millet, D.B., Marshall, J.D., 2015. National spatiotemporal exposure surface for no2: monthly scaling of a satellite-derived land-use regression, 2000–2010. Environ. Sci. Technol. 49, 12297–12305.
Beckerman, B.S., Jerrett, M., Martin, R.V., et al., 2013. Application of the deletion/substitution/addition algorithm to selecting land use regression models for interpolating air pollution measurements in California. Atmos. Environ. 77, 172–177.
Beelen, R., Raaschou-Nielsen, O., Stafoggia, M., et al., 2014. Effects of long-term exposure to air pollution on natural-cause mortality: an analysis of 22 European cohorts within the multicentre escape project. Lancet (London, England) 383, 785–795.
Brauer, M., Hoek, G., van Vliet, P., et al., 2003. Estimating long-term average particulate air pollution concentrations: application of traffic indicators and geographic information systems. Epidemiology (Cambridge, Mass) 228–239.
Breiman L, Cutler A, Liaw A, et al. 2011. Package'randomforest'. software available at URL: http://stat-www berkeley edu/users/breiman/RandomForests.
Briggs, D.J., Collins, S., Elliott, P., et al., 1997. Mapping urban air pollution using gis: a regression-based approach. Int. J. Geogr. Inf. Sci. 11, 699–718.
Brokamp, C., Jandarov, R., Rao, M.B., et al., 2017. Exposure assessment models for elemental components of particulate matter in an urban environment: a comparison of regression and random forest approaches. Atmos Environ (1994) 151, 1–11.
Brokamp, C., Jandarov, R., Hossain, M., et al., 2018. Predicting daily urban fine particulate matter concentrations using a random forest model. Environ. Sci. Technol. 52, 4173–4179.
CGIAR-CSI. Srtm 90m digital elevation data.
Crouse, D.L., Goldberg, M.S., Ross, N.A., 2009. A prediction-based approach to modelling temporal and spatial variability of traffic-related air pollution in Montreal, Canada. Atmos. Environ. 43, 5075–5084.
Cyrys, J., Eeftens, M., Heinrich, J., et al., 2012. Variation of no2 and nox concentrations between and within 36 European study areas: results from the escape study. Atmos. Environ. 62, 374–390.
Di, Q., Kloog, I., Koutrakis, P., et al., 2016a. Assessing pm2. 5 exposures with high spatiotemporal resolution across the continental United States. Environ. Sci. Technol. 50, 4712–4721.
Di, Q., Koutrakis, P., Schwartz, J., 2016b. A hybrid prediction model for pm2. 5 mass and components using a chemical transport model and land use regression. Atmos. Environ. 131, 390–399.
Di, Q., Wang, Y., Zanobetti, A., et al., 2017. Air pollution and mortality in the medicare population. N. Engl. J. Med. 376, 2513–2522.
EEA, 2015. Airbase - The European Air Quality Database, Version 8 (Available). http://www.eea.europa.eu/data-and-maps/data/airbase-the-european-air-quality-database-8 (Accessed date: 13 January 2015).
Eeftens, M., Tsai, M.-Y., Ampe, C., et al., 2012. 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. 62, 303–317.
ETC-LC, 2013. Corine Land Cover (clc2006), Raster Database (Version 12/2013).
ETC-LC, (2009). Corine Land Cover (clc2000), Raster database (version 12/2009).
Ferwerda, J., Hainmueller, J., Hazlett, C.J., 2017. Kernel-based regularized least squares in r (krls) and stata (krls). J. Stat. Softw. 79.
Friedman, J., Hastie, T., Tibshirani, R., 2001. The Elements of Statistical Learning:Springer Series in Statistics New York.
Friedman, J., Hastie, T., Tibshirani, R., 2009. Glmnet: Lasso and elastic-net regularized generalized linear models. R package version 1.
Hazlett, J.H.S.C., Hainmueller, M.J., 2017. Package 'krls'.
Henderson, S.B., Beckerman, B., Jerrett, M., et al., 2007. Application of land use regression to estimate long-term concentrations of traffic-related nitrogen oxides and fine particulate matter. Environ. Sci. Technol. 41, 2422–2428.
Hoek, G., 2017. Methods for assessing long-term exposures to outdoor air pollutants. Environmental Science 4, 450–462.
Hoek, G., Beelen, R., de Hoogh, K., et al., 2008. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmos. Environ. 42, 7561–7578.
de Hoogh, K., Gulliver, J., van Donkelaar, A., et al., 2016. Development of west-european pm 2.5 and no 2 land use regression models incorporating satellite-derived and chemical transport modelling data. Environ. Res. 151, 1–10.
de Hoogh, K., Chen, J., Gulliver, J., et al., 2018a. Spatial pm2.5, no2, o3 and bc models for western europe - evaluation of spatiotemporal stability. Environ. Int. 120, 81–92.
de Hoogh, K., Heritier, H., Stafoggia, M., et al., 2018b. Modelling daily pm2.5 concentrations at high spatio-temporal resolution across Switzerland. Environ. Pollut. 233, 1147–1154.
Hu, X., Belle, J.H., Meng, X., et al., 2017. Estimating pm2.5 concentrations in the conterminous United States using the random forest approach. Environ. Sci. Technol. 51 (12), 6936–6944.
Hystad, P., Setton, E., Cervantes, A., et al., 2011. Creating national air pollution models for population exposure assessment in Canada. Environ. Health Perspect. 119, 1123.
Inness, A., Baier, F., Benedetti, A., et al., 2013. The macc reanalysis: an 8 yr data set of atmospheric composition. Atmos. Chem. Phys. 13, 4073–4109.
Kerckhoffs, J., Hoek, G., Messier, K.P., et al., 2016. Comparison of ultrafine particle and black carbon concentration predictions from a mobile and short-term stationary land-use regression model. Environ. Sci. Technol. 50, 12894–12902.
Kerckhoffs, J., Hoek, G., Portengen, L., et al., 2019. Performance of prediction algorithms for modeling outdoor air pollution spatial surfaces. Environmental Science Technology.
Kim, S.Y., Sheppard, L., Bergen, S., et al., 2016. Prediction of fine particulate matter chemical components with a spatio-temporal model for the multi-ethnic study of atherosclerosis cohort. J Expo Sci Environ Epidemiol 26, 520–528.
Liu, Y., Paciorek, C.J., Koutrakis, P., 2009. Estimating regional spatial and temporal variability of pm(2.5) concentrations using satellite data, meteorology, and land use information. Environ. Health Perspect. 117, 886–892.
Mercer, L.D., Szpiro, A.A., Sheppard, L., et al., 2011. Comparing universal kriging and land-use regression for predicting concentrations of gaseous oxides of nitrogen (NOx) for the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). Atmos. Environ. 45 (26), 4412–4420.
Meyer, D., Dimitriadou, E., Hornik, K., et al., 2017. Package 'e1071'.
Novotny, E.V., Bechle, M.J., Millet, D.B., et al., 2011. National satellite-based land-use regression: No2 in the United States. Environ. Sci. Technol. 45, 4407–4414.
Pinault, L.L., Weichenthal, S., Crouse, D.L., et al., 2017. Associations between fine particulate matter and mortality in the 2001 Canadian census health and environment cohort. Environ. Res. 159, 406–415.
Reid, C.E., Jerrett, M., Petersen, M.L., et al., 2015. Spatiotemporal prediction of fine particulate matter during the 2008 northern California wildfires using machine learning. Environ. Sci. Technol. 49, 3887–3896.
Ridgeway, G., Southworth, M.H., RUnit, S., 2013. Package 'gbm'. Viitattu 10, 40.
Ripley, B., Venables, W., Ripley, M.B., 2016. Package 'nnet'. R package version 7, 3–12.
Schaap, M., Apituley, A., Timmermans, R., et al., 2009. Exploring the relation between aerosol optical depth and pm 2.5 at cabauw, the Netherlands. Atmos. Chem. Phys. 9, 909–925.
Schalkoff, R.J., 1997. Artificial neural networks:McGraw-Hill. New York.
Stafoggia, M., Schwartz, J., Badaloni, C., et al., 2017. Estimation of daily pm10 concentrations in Italy (2006-2012) using finely resolved satellite data, land use variables and meteorology. Environ. Int. 99, 234–244.
Suzuki, H., Miyao, Y., Nakayama, T., et al., 2011. Comparison of laser-induced

fluorescence and chemiluminescence measurements of no2 at an urban site. Atmos. Environ. 45, 6233–6240.
Team, R.C., 2013. R: A Language and Environment for Statistical Computing.
Tibshirani, R., James, G., Witten, D., et al., 2013. An Introduction to Statistical Learning-with Applications in r. Springer, New York, NY.
Van den Bossche, J., De Baets, B., Verwaeren, J., et al., 2018. Development and evaluation of land use regression models for black carbon based on bicycle and pedestrian measurements in the urban environment. Environ. Model. Softw. 99, 58–69.
Van Donkelaar, A., Martin, R.V., Brauer, M., et al., 2015. Use of satellite observations for long-term exposure assessment of global concentrations of fine particulate matter. Environ. Health Perspect. 123, 135.
Vienneau, D., de Hoogh, K., Beelen, R., et al., 2010. Comparison of land-use regression models between great britain and the Netherlands. Atmos. Environ. 44, 688–696.
Vienneau, D., De Hoogh, K., Bechle, M.J., et al., 2013. Western european land use regression incorporating satellite-and ground-based measurements of no2 and pm10. Environ. Sci. Technol. 47, 13555–13564.
Wang, M., Beelen, R., Eeftens, M., et al., 2012. Systematic evaluation of land use regression models for no2. Environ. Sci. Technol. 46, 4481–4489.
Wang, M., Beelen, R., Bellander, T., et al., 2014. Performance of multi-city land use regression models for nitrogen dioxide and fine particles. Environ. Health Perspect. 122, 843–849.
Weichenthal, S., Ryswyk, K.V., Goldstein, A., et al., 2016. A land use regression model for ambient ultrafine particles in Montreal, Canada: a comparison of linear regression and a machine learning approach. Environ. Res. 146, 65–72.
Wood, S., Wood, M.S., 2015. Package 'mgcv'. R package version 1–7.
Zhan, Y., Luo, Y., Deng, X., et al., 2017. Spatiotemporal prediction of continuous daily pm 2.5 concentrations across China using a spatially explicit machine learning algorithm. Atmos. Environ. 155, 129–139.
Zhan, Y., Luo, Y., Deng, X., et al., 2018. Satellite-based estimates of daily no2 exposure in China using hybrid random forest and spatiotemporal kriging model. Environ. Sci. Technol. 52, 4180–4189.
Zhang, J., Ding, W., 2017. Prediction of air pollutants concentration based on an extreme learning machine: the case of Hong Kong. Int. J. Environ. Res. Public Health 14.
Zou, B., Wang, M., Wan, N., et al., 2015. Spatial modeling of pm 2.5 concentrations with a multifactoral radial basis function neural network. Environ. Sci. Pollut. Res. 22, 10395–10404.