Atmospheric Pollution Research 13 (2022) 101611

Contents lists available at ScienceDirect

Atmospheric Pollution Research

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

A comparison of statistical and machine-learning approaches for spatiotemporal modeling of nitrogen dioxide across Switzerland

Tze-Li Liu [a,b], Benjamin Flückiger [a,b], Kees de Hoogh [a,b,*]

[a] Swiss Tropical and Public Health Institute, Allschwil, Switzerland
[b] University of Basel, Basel, Switzerland

ARTICLE INFO

Keywords:
Spatiotemporal models
Air pollution
Satellite
Exposure assessment
Land use regression
Machine learning

ABSTRACT

Land use regression modeling has commonly been used to model ambient air pollutant concentrations in environmental epidemiological studies. Recently, other statistical and machine-learning methods have also been applied to model air pollution, but their relative strengths and limitations have not been extensively investigated. In this study, we developed and compared land-use statistical and machine-learning models at annual, monthly and daily scales estimating ground-level NO_2 concentrations across Switzerland (at high spatial resolution 100 × 100 m). Our study showed that the best model type varies with context, particularly with temporal resolution and training data size. Linear-regression-based models were useful in predicting long-term (annual, monthly) spatial distribution of NO_2 and outperformed machine-learning models. However, linear-regression-based models were limited in representing short-term temporal variation even when predictor variables with temporal variability were provided. Machine-learning models showed high capability in predicting short-term temporal variation and outperformed linear-regression-based models for modeling NO_2 variation at high temporal resolution (daily). However, the best performing models, XGBoost and LightGBM, constantly overfit on training data and may result in erratic patterns in the model-estimated concentration surfaces. Therefore, the temporal and spatial scale of the study is an important factor on which the choice of the suitable model type should be based and validation is required whatever approach is used.

1. Introduction

Nitrogen dioxide (NO_2) is one of the major air pollutants of concern, with the anthropogenic emissions highly related to traffic and combustion. Being a highly reactive gas, nitrogen dioxide is a respiratory tract irritant that is associated with a number of adverse health effects, including both short- and long-term (Chen et al., 2007; Samet and Utell, 1990; World Health Organization, 2006; Yassi et al., 2001).

Epidemiological studies investigating the associations between air pollutants and the adverse health effects rely on a good quality exposure assessment (Röösli and Vienneau, 2014). Modeling is a cost-effective approach able to reflect the spatial variability of air pollution concentrations (Gulliver and de Hoogh, 2015).

One modeling approach, land use regression (LUR), regresses observed concentrations against geographical and environmental features around the point location of monitoring sites (Briggs et al., 1997). LUR assumes an underlying relationship between the variation of the measured concentration and the surrounding environment, for example population density, land use and various traffic-related variables. Geographical features surrounding the monitoring are extracted using geographical information systems (GIS) (Briggs et al., 1997; Eeftens et al., 2012; Gulliver and de Hoogh, 2015; World Health Organization, 2006). Studies have shown that the LUR models are able to explain a large amount of spatial variability (Beelen et al., 2010; Chen et al., 2019a; Eeftens et al., 2012; Lee et al., 2014), and epidemiological studies have increasingly turned to LUR modeling for exposure assessment in air pollution studies (Gulliver and de Hoogh, 2015; Hoek et al., 2008; Montagne et al., 2013).

Although the term land use regression sometimes strictly refers to multiple linear regression models developed with a supervised stepwise variable selection, a number of different models, evolving from typical linear regressions to machine-learning approaches, have been developed and applied for regression tasks with land-use and environmental covariates to model air pollution. Commonly used approaches include

Peer review under responsibility of Turkish National Committee for Air Pollution Research and Control.
* Corresponding author. Swiss Tropical and Public Health Institute, Kreuzstrasse 2, 4123, Allschwil, Switzerland,
E-mail addresses: c.dehoogh@swisstph.ch, c.dehoogh@unibas.ch (K. de Hoogh).

https://doi.org/10.1016/j.apr.2022.101611
Received 15 June 2022; Received in revised form 4 November 2022; Accepted 24 November 2022
Available online 3 December 2022

Supervised stepwise linear regression (SLR). SLR selects the "best" set of predictor variables in a sequential manner and models the ground-level concentration by linear regression with the selected predictor variables (Eeftens et al., 2012; Gulliver and de Hoogh, 2015). SLR has been used in many studies to model the air pollution concentration, for example in the European Study of Cohorts for Air Pollution Effects (ESCAPE) project (Beelen et al., 2014; de Hoogh et al., 2016; Eeftens et al., 2012; Lee et al., 2014). Geographically-weighted regression (GWR) relaxes the assumption of the spatial stationarity of model coefficient estimates and allows varying relationships in different areas in space (Thapa and Estoque, 2012). For example, a spatially heterogeneous relationship between AOD and $PM_{2.5}$ was found across the contiguous United States and GWR was used to account for the spatially inconsistent relationship (Hu, 2009). While GWR provides flexibility in the spatial dimension, Linear mixed effect regression models (LMER) can provide flexibility in the temporal dimension by adjusting the estimation with the random effect by time (e.g. by month or by day). Linear mixed effect models have been applied to estimate the daily $PM_{2.5}$ and NO_2 concentration across Switzerland (de Hoogh et al., 2018, 2019). Machine-learning algorithms are theoretically capable of approximating any linear or non-linear function as well as the complex potential interactions among the predictor variables (Bishop, 1995). Applications of the ensemble decision-tree-based algorithms, particularly random forest and gradient boosting machines, have been used to model particulate matter in Italy (Stafoggia et al., 2019), NO_2 and $PM_{2.5}$ at a European scale (Chen et al., 2018, 2019a), and $PM_{2.5}$ across the contiguous United States (Di et al., 2019; Hu et al., 2017). A higher predictive accuracy of the ensemble-tree-based algorithms has been observed compared to the linear regression methods in many studies (Chen et al., 2019a; Lu et al., 2020). Aside from the decision-tree-based algorithms, neural networks have also been applied to model the ground-level concentration of air pollutants. Although a number of studies used a variety of neural networks to model or forecast air pollution (Alimissis et al., 2018; Cabaneros et al., 2019, 2020; Liu et al., 2020; Mlakar and Boznar, 2011; Tsai et al., 2018; Van Roode et al., 2019), most of which explain the spatial variability only by interpolation based on coordinates (Alimissis et al., 2018; Van Roode et al., 2019), relatively few entered environmental covariates into the model as input variables like in the LUR approach. Recent applications in the United States (Di et al., 2016, 2019) modeled the $PM_{2.5}$ concentration with a large number of predictor variables (including land use and meteorological variables) based on neural networks and have shown promising results.

Each model has strengths and limitations. Linear regression makes assumptions like the linearity and spatial-stationarity of the effects and does not assume interactions. Furthermore, the stepwise forward approach is often criticized for preventing identification of relevant interactions and non-linear effects (Bishop, 1995; Guyon and Elisseeff, 2003). Although GWR and LMER provide more flexibilities of the model estimates in both the spatial and/or temporal dimension, the models still make the assumption of the linearity of the effects and the potential overfitting of GWR may be of concern. On the other hand, despite the fact that the machine learning algorithms are able to model non-linear relationships and complex interactions, these algorithms are typically data-hungry, computationally demanding and hard to interpret.

With the increase of the number of available model types comes a need to investigate the strengths and limitations of these models for air pollution modeling. Here we conducted an inter-comparison between the performances of different types of models predicting NO_2 concentration at various temporal resolutions. We incorporated multiple variables, including satellite-derived atmospheric column NO_2, meteorological variables and land-use variables, into land-use statistical and machine-learning models and developed models with six algorithms at annual, monthly, and daily temporal scale for Switzerland at 100 × 100 m spatial resolution. The predictive performance of the models was then compared to investigate the strengths and limitations of each model.

2. Materials and methods

2.1. Study domain

The study area is Switzerland, a landlocked country in Central Europe with a great variety of landscapes and climates in a relatively small area (41,285 km^2). For the aims of the study, we divided the Swiss spatial domain into 100 × 100 m grid cells (projected coordinate system CH 1903+/LV95). The study period is from January 1st, 2019 to December 31st, 2019, a total of 365 days. Using remote-sensing data in air pollution modeling has been a common approach in recent studies (de Hoogh et al., 2016; de Hoogh et al., 2018; de Hoogh et al., 2019; Stafoggia et al., 2019), and the new generation remote-sensing tropospheric NO_2 data product from TROPOMI (online in July 2018, see 2.2.2) is expected to be a main data source for future studies, hence being highly relevant to the selection of the study period to incorporate the availability of the data product.

2.2. Data

2.2.1. Nitrogen dioxide monitoring data

Daily NO_2 monitoring data ($\mu g/m^3$) for 2019 were obtained from the Swiss Air Pollution Database "Immissionsdatenbank Luft der Schweiz" (https://www.arias.ch/arbeit/welcome.html). Ninety-six sites with at least 30 daily observations were included in the training data (Figure A1), including 39 traffic sites, 6 industry sites, and 51 background sites (characterized by the monitoring sites and their surrounding areas; Table A1). Most of the sites performed the measurement using chemiluminescent and a small number (4 sites) used optical spectroscopy methods (DOAS).

2.2.2. Satellite-derived tropospheric column NO_2

The spectral absorption characteristics of atmospheric NO_2 allows indirect measurement of the atmospheric column amount NO_2 from satellite-based earth observations (Lamsal et al., 2020). Data collected by the TROPOspheric Monitoring Instrument (TROPOMI) onboard the European Space Agency Sentinel-5 Precursor (S–5P) satellite was used in this study. TROPOMI is a spectrometer that allows observations of key atmospheric constituents including NO_2, O_3, CO, SO_2, CH_4, CH_2O, and aerosols (Veefkind et al., 2012). From July 2018 onwards, the S–5P mission plays a transitional gap-filler role that provides observation time series of tropospheric data products in the timeframe 2017–2023, the period between the current OMI (Ozone Monitoring Instrument), SCIAMACHY (SCanning Imaging Absorption spectroMeter for Atmospheric CartograpHY) (Bovensmann et al., 1999) and the upcoming operational Sentinel-5 observation of air quality and climate. Compared to the previous OMI observation, TROPOMI observes the atmosphere with a higher spectral (extended the wavelength range in the NIR and SWIR) and spatial (7 × 3.5 km) resolution (13 × 25 km for OMI) (Eskes et al., 2020), which in principle allows for a more detailed observation of the spatially inhomogeneous distribution of NO_2. The Level-2 (L2) offline daily tropospheric column NO_2 data product version 1.4.0 (*nitrogendioxide_tropospheric_column*) (Copernicus Sentinel-5P (processed by ESA), 2021) was obtained from the ESA Sentinel-5P Pre-operations Data Hub (https://s5phub.copernicus.eu/dhus/) and used in this study (Table 1).

The availability of satellite-derived atmospheric column NO_2 observations were limited by the meteorological and surface conditions on the sensing date, and a number of missing pixels values exist in the dataset. Before using these satellite-derived observations as input variables to model the ground-level NO_2 concentration, the missing pixels were imputed. The general approach follows approaches used in previous studies (de Hoogh et al., 2019; Stafoggia et al., 2019). A random forest model was used to model the relationship between the TROPOMI observations and some covariates which have no missing values, including meteorological variables, elevation, and CAMS-modeled

Table 1
Summary of input data.

Data type	Name	Source	Details
Ground NO_2 monitoring	Monitored NO_2	Immissionsmesswerte Luft der Schweiz	Unit: μg/m^3
Satellite-derived tropospheric column NO_2	OMI Tropospheric column NO_2	NASA EarthData GES DISC	- Spatial resolution: 0.25° × 0.25° (~13 × 25 km) - Temporal resolution: 1 day - Unit: molec/cm^2
	TROPOMI Tropospheric column NO_2	ESA Sentinel-5P Pre-operations data hub	- Spatial resolution: 7 × 3.5 km - Temporal resolution: 1 day - Unit: molec/cm^2 (preprocesseed)
Modeled NO_2 (for the imputation models)	Modeled total column amount NO2	CAMS global reanalysis (EAC4)	- Spatial resolution: 0.75° × 0.75° (~80 × 80 km) - Temporal resolution: 1 day - Unit: kg/m^2
Spatial predictor variables	CORINE land cover 2018	Copernicus Land Monitoring Service	Spatial resolution: 100 × 100 m
	EU-DEM v1.1 elevation	Copernicus Land Monitoring Service	Spatial resolution: 25 × 25 m
	GHS population 2016	EU Open Data	Spatial resolution: 100 × 100m
	VIIRS light at night	Earth Observation Group Annual VNL V2	Spatial resolution: 15 arc second
	NOx emissions 2015 (by source types)	Meteotest	Spatial resolution: 200 × 200m
	NDVI	MODIS	Spatial resolution: 30 × 30 m resampled to 100 × 100m
	Number of intersections Traffic intensity Road density Distance to nearest major road	Calculated from Sonbase database road network	Spatial resolution: 100 × 100m
Spatialtemporal predictor variables	ERA5 meteorological variables	ERA5 (ECMWF)	- Variables (units): 2m temperature (K), 10m u-component of wind (m/s), 10m v-component of wind (m/s), Surface pressure (Pa), Total precipitation (m), Total cloud cover (0–1), Boundary layer height (m) - Spatial resolution: 0.25° × 0.25° - Temporal resolution: 3 h
	NDVI	MODIS vegetation MYD13Q1	- Spatial resolution: 250 × 250 m - Temporal resolution: 1 day

atmospheric NO_2 (see Appendix A.1 for details). The final imputation model was used to fill in the missing pixel values in the TROPOMI tropospheric column NO_2 observations, where missing pixels were filled in with the model-predicted values in the final imputed data product and the pixels that were not missing took the original observed values.

2.2.3. Spatial-temporal predictor variables

Daily meteorological variables (2m air temperature, 10m u-component of wind, 10m v-component of wind, surface pressure, total precipitation, total cloud cover, and boundary layer height) of the ERA5-Reanalysis dataset (Hersbach et al., 2018) modeled for 12:00:00UTC were obtained from European Centre for Medium-Range Weather Forecasts (ECMWF) (Table 1). Vegetation changes and distribution may associate with the variation of air pollutants as a pollution sink, and the 16-day normalized difference vegetation index (NDVI) time series data was obtained from the Aqua Moderate Resolution Imaging Spectroradiometer (MODIS) Vegetation Indices Version 6 (MYD13Q1 v006) data product (Didan, 2015) from the NASA EarthData platform (Table 1). Detailed information can be found in Appendix A.2. Pixel-wise temporal aggregation was applied to the spatial-temporal predictor variable raster time-series to obtain the annual and monthly mean raster from the daily time-series.

2.2.4. Spatial predictor variables

The following spatial predictor variables were used in the study to model ground level concentration: land cover type (percentage of residential, industrial or commercial, total built-up, urban green areas, agricultural, and semi-natural and forest area), elevation, population, nighttime light, NOx emission inventories, 30m-NDVI, and traffic variables (number of intersections, traffic intensity, distance to nearest major road, major road density, and all road density) (Table 1). Detailed descriptions of the variables can be found in Appendix A.3. The spatial predictor variables (except "distance from nearest major road" and "30m-NDVI") were summarized using circular buffer moving windows with various radii (100, 200, 500, 1000, 2000, 5000, 10000m) for the neighborhood of the focal cells to account for the information of the surrounding pixels (Eeftens et al., 2012; Gulliver and de Hoogh, 2015).

2.3. Modeling ground-level concentration

Different algorithms were applied to model the relationship between the ground-level concentration and the environmental and land use covariates at 100 × 100 m spatial resolution.

Prior to modeling, logarithm transformation $x'=ln(x+1)$ was applied to the following predictor variables (and their spatially-buffed products) that were most skewed: emissions, nighttime light, population, traffic intensity, the density of major roads and all roads, and number of intersections. The scalar 1 was added to the predictor variables before the logarithm transformation to avoid $ln(0)$ because some original values were zero. Following the transformation, the predictor variables as well as their spatially buffered products were standardized to mean 0 and unity standard deviation ($x' = \frac{x-\bar{x}}{\sigma_x}$) to allow comparison between model coefficients and better model performance (except land cover and total cloud cover, whose unit is "percent" and value range is between 0 and 1).

2.3.1. Supervised forward stepwise linear regression (SLR)

We applied the supervised stepwise model selection algorithm and criteria following the protocol of the ESCAPE study (de Hoogh et al., 2016; Eeftens et al., 2012). In short, the expected sign of the regression coefficients of all predictor variables were defined a priori. Univariate linear regressions were first fitted for the respective predictor variables, and the one with the highest adjusted-R^2 was included as the initial predictor variable. Then, the remaining predictor variables were sequentially added in to the linear regression model with the initially included variable. The variable that (1) further gave the highest gain in the model adjusted-R^2, (2) with the right direction of effect, and (3) did not change the direct of effect of the existing predictor variable was included in the model. This step was repeated iteratively until there were no more variables (1) with the right direction of effect and (2) which added at least 0.01 to the adjusted-R^2 of the previous model. Finally, variables with (1) a p-value >0.10 (insignificant variables) or (2) Variable Inflation Factor (VIF) > 3 (variables with collinearity) were removed from the model.

For the monthly and daily model, we engineered additional predictor variables for the SLR models to take into account the information of temporal changes. Seasonal cyclic patterns exist in the NO_2 concentration, but the relationship between the time step ("month" or "day of year") and the concentration is non-linear. We applied cosine transformation to the "month" variable (1–12) of the monthly model and the "DOY" variable (1–365) of the daily model to linearize the relationship (Appendix A.4). The transformed variable ranges between −1 and 1 and has a more linear relationship to the response variable.

2.3.2. Geographically weighted regression

Whereas conventional regression models $y_i = \beta_0 + \sum_k \beta_k x_{ik} + \varepsilon_i$ assume the regression coefficients β_k to be spatially stationary, geographically weighted regression models permit the coefficient estimates to vary locally [16]: $y_i = \beta_0(u_i, v_i) + \sum_k \beta_k(u_i, v_i) x_{ik} + \varepsilon_i$ where (u_i, v_i) were the coordinates of point *i*. Data closer to the point i were weighted more than data from observations that were far away [16]. GWR was used for annual prediction to take into account the spatial variability of long-term NO_2 concentration.

GWR was used to fit the annual average concentration based on the predictor variables selected by SLR. An adaptive bandwidth for the distance decay function was adopted for the GWR model. The GWR model was implemented with the R package "GWmodel" (Gollini et al., 2015).

2.3.3. Supervised stepwise linear mixed effect regression model (SLMER)

Linear mixed effect models were applied to the monthly and daily data to adjust the regression coefficient of the pooled linear regression as the random effect, which allows precise prediction for each individual time step (best linear unbiased predictors; BLUPs). In this study, the estimation of random effect was based on a preliminary variable selection with the SLR algorithm (without the presence of the cosine-transformed temporal variables). For the monthly model, a random intercept (an adjustment to the intercepts for each month) estimation was added to the model random effect structure. For the daily model, a random slope on the TROPOMI satellite NO_2 observation was further added to the random effect structure for each individual day along with the random intercept. The mixed effect models were implemented with the "lme4" package (Bates et al., 2015).

Note the limitation for applying the temporally-blocked cross validation (see the following section 2.4 Model evaluation) for the mixed effect models. The estimated values of the test set in the temporally-blocked cross validation were only population-level predictions (no random effect adjustments) since the random effect grouping of test set observations (an entire month) was not present in the data for model fitting. Therefore, the temporally-blocked CV predictions only represent the estimation based on the fixed effects.

2.3.4. Random forest (RF)

Random forest (Breiman, 2001) uses bootstrap aggregation of classification and regression trees (CARTs) by randomly drawing of only a subset (instead of all) of the predictor variables at each split (Breiman, 2001; Liaw and Wiener, 2002). We applied random forest to fit the annual, monthly, and daily land-use random forest models with the "ranger" package (Wright et al., 2015). An embedded variable selection (Guyon and Elisseeff, 2003) based on variable importance was applied prior to the fitting of the final random forest model (Appendix A.5 for details). The hyperparameters (n.trees, mtry; see Table A2-A4) were selected based on a grid search over a number of hyperparameter combinations that minimize the model cross-validated RMSE loss in a parameter space.

2.3.5. Gradient boosting machines (XGBoost, LightGBM)

Whereas random forest grows and aggregates parallel bootstrapped decision trees, "boosting" grows trees sequentially where the later trees learn the pattern based on the errors of previous trees. Two most widely-used variants of boosting decision-tree-based algorithms, Extreme Gradient Boosting (XGBoost) (Chen and He, 2015) and Light Gradient Boosting Machine (LightGBM) (Ke et al., 2017), were used in this study to fit the annual, monthly, and daily land-use boosting models (with the "xgboost" (Chen and He, 2015; Chen et al., 2019b) and "lightgbm" (Ke et al., 2017, 2021) packages). Embedded variable selections (Guyon and Elisseeff, 2003) based on variable importance were applied prior to the fitting of the final XGBoost and LightGBM models (Appendix A.5 for details). The hyperparameters of the final XGBoost and LightGBM models, that minimize the model cross-validated RMSE loss, were selected based on the grid searches over a number of hyperparameter combinations in a parameter space for XGBoost and LightGBM (Table A2-A4).

2.3.6. Neural network (NN)

Neural networks are function approximators composed of interconnected neurons (Bishop, 1995; Goodfellow et al., 2016). The interconnected weights are the basic parameters adjusted during the training process with the backpropagation algorithm and stochastic-gradient-descent-based optimization, which tries to find the parameter that minimizes the error between the observed and predicted values (the loss function) (Bishop, 1995; Goodfellow et al., 2016; Mlakar and Boznar, 2011).

We developed neural network models for the regression of ground-level concentration with "Keras" (Allaire and Chollet, 2021). For the monthly and daily models, we applied a variable selection prior to the fitting of the final neural network model based on variable importance with Garson's algorithm (Garson, 1991; Goh, 1995) (Appendix A.5 for details). The architecture (hyperparameters) of the final model (number of hidden layers and neurons, regularization, dropout layer) and setting of the training process (batch size, training epochs) was selected based on grid searches over a number of hyperparameter combinations that minimize the model cross-validated MAE loss in a parameter space (Table A2-A4).

2.4. Model evaluation

The statistical performance of models was evaluated by comparing the observed and the predicted values of the models and by analyzing the prediction residuals. Besides the prediction based on the model developed with the complete training data (full model), we performed cross validations and compared the observed and the cross-validation prediction values to evaluate the robustness of the algorithms. Three different cross validation partitioning methods were applied: (1) The conventional 5-fold random-split cross validation randomly divided the 96 monitoring sites into 5 groups. In each iteration the data of one group

was held back in the training data when the model was fitted, and was used to evaluate the error of the fitted model by the predicted values. Furthermore, we applied a (2) 5-fold spatially-blocked cross validation. Instead of dividing the monitoring sites randomly, the study area was split into large chunks and the partitioning of the cross-validation folds was assigned to the chunks. All the monitoring sites within a same chunk were assigned with the same cross validation fold (Fig. 1 (a)). The spatially-blocked CV was a more conservative approach to evaluate the model's ability to generalize over space, since the model was applied to an area where no information was included in the training process. A previous review (Roberts et al., 2017) suggested that for the data with spatial, temporal or hierarchical structure, block cross validation was more appropriate to evaluate model performance than random cross validation when the goal was predicting to new data or predictor space. We also applied a (3) 12-fold temporally-blocked cross validation for the monthly and daily model to evaluate the models' robustness in time. The dataset was partitioned by 12 months where in each iteration the model was fitted with data from 11 months and predicted on the data of the held-out month.

We calculated the coefficient of determination (commonly referred to as R^2) and root mean square error (RMSE) to evaluate model performance. We also assessed the potential spatial autocorrelation of the model residuals with the global Moran's I statistic [61–64] of the residuals of the full training model.

Additionally, the fold-specific RMSEs of the spatially- and temporally-blocked (for monthly and daily models) cross validations were compared across folds to assess the space- and time-specific uncertainties of the models. The time series structure (trend and periodicity) of the daily-model residuals was also analyzed (Appendix A.6).

3. Results and discussion

3.1. Annual models

Overall, SLR and RF are the best performing models at annual average scale (Fig. 2). RF gave the highest random-split CV-R^2 (0.749) and SLR gave the highest spatially-blocked CV-R^2 (0.728) (Table A5).

SLR has the highest predictive performance at annual scale despite its simplicity and the potential limitation of the forward selection approach. The random-split (0.736) and spatially-blocked (0.728) CV-R^2 of SLR are close to the full-model R^2 (0.758), suggesting the robustness of the model in space. The statistical performance of GWR was little different from that of the SLR model on which the GWR was developed. Figure A2 shows the very low spatially-varying regression coefficients of GWR, probably due to the relatively small study area and the consequently limited heterogeneity. The random-split (0.749) and spatially-blocked (0.722) CV-R^2 values of RF are comparable to that of SLR. Whereas the full-model R^2 of the XGBoost model is almost 1, the R^2 measured by random-split (0.731) and spatial-blocked (0.710) CV models are lower than SLR and RF (Table A5). Likewise, the full-model R^2 of the LightGBM is around 90% but the spatially-blocked CV-R^2 (0.671) of the LightGBM models are the lowest among all of the annual models compared, suggesting that the model may be spatially not as robust as the other ones. The full-model R^2 of the neural network models is around 76%, and random-split CV-R^2 is the lowest (0.674) among all of the annual models compared.

In summary, SLR and RF are the best performing models predicting annual average concentrations. The two boosting models overfit on the training data, suggesting that the models are too complex for the data. Since monitoring data for only one year was used for training in this study, the performance of the boosting models might be better if the data size is bigger (e.g., multiple years) for better generalization.

Chen et al. (2019a) compared 16 algorithms predicting annual average $PM_{2.5}$ and NO_2 concentrations across Europe and showed that the NO_2 models performed similarly across different algorithms with 5-fold-CV R^2 ranging from 0.57 to 0.62 (Chen et al., 2019a). Despite the overall higher R^2 and lower RMSE values in this study, possibly linked to the smaller study area, we also found that the difference between the annual models was relatively small (5-fold random-split CV-R^2 ranging from 0.672 to 0.749). Chen et al. (2019a) suggested that the small differences between models may be related to the large number of training data (number of monitoring sites) and the lack of complex relationships between the predictor variables and the relatively stable annual average concentrations.

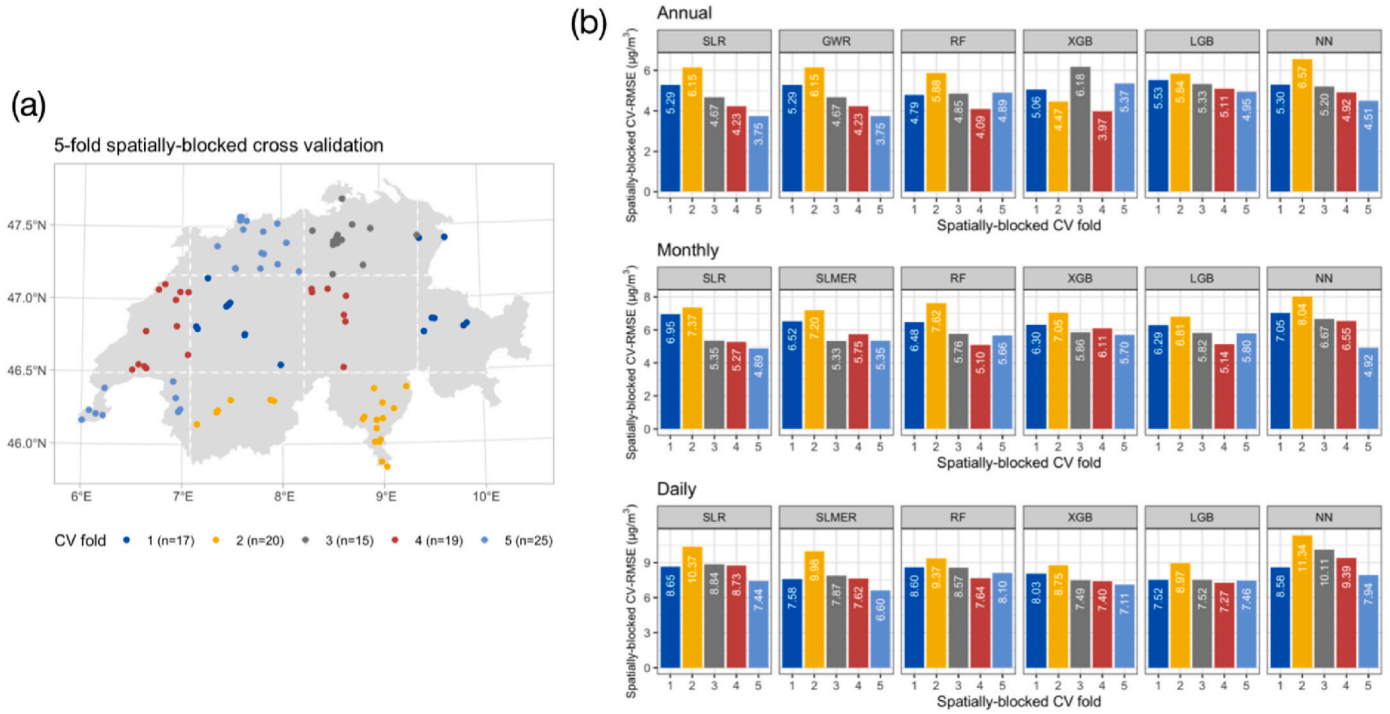

Fig. 1. (a) The spatial partitioning of the spatially-blocked cross validation. (b) The fold-specific RMSEs (μg/m^3) of the spatially-blocked cross validated models.

In most of the annual models, TROPOMI-observation NO_2 product was not selected by the models (Table A10; Figure A5), whereas the traffic-related variables (traffic intensity, major road density, traffic-source NOx emission) at various radii consistently were the variables with the highest ranking. The improvement of the performance of the SLR models by including satellite-derived column NO_2 (OMI) was larger in the West-European models (de Hoogh et al., 2016) than in this study. This is likely because of the little extra information added by the satellite observation on the spatial variability of NO2 concentrations already covered by the land-use terms. This as opposed to large areas, where the satellite-derived column NO_2 is able to explain large scale variability, which, at the European scale, improved the model performance (de Hoogh et al., 2016).

3.2. Monthly models

The two gradient-boosting models, XGBoost and LightGBM, have the highest random-split CV-R^2 (XGB: 0.784; LGB: 0.787) and temporally-blocked CV-R^2 (both 0.901) (Fig. 2, Table A6). The models with the highest spatially-blocked CV-R^2 are LightGBM (0.728), SLR (0.727), and SLMER (0.722). The overall performance of the neural network models is the lowest among the algorithms compared. The difference between the full-model R^2 and CV-R^2 is largest for the two gradient-boosting models (full-model $R^2 > 0.99$), suggesting an overfitting.

A pattern of underestimation at lower fitted values (in the range $\hat{y} < 0$) can be observed in SLR and SLMER, showing the limitation of linear predictions of the linear regression models. An overestimation at higher fitter values (in the range $\hat{y} > 30$) can be observed with the NN model (Figure A3 (b)).

Unlike in the annual models, the TROPOMI-observation NO_2 was selected as a predictor variable in all models (Table A10; Figure A5). The traffic-related variables, including traffic density, major road density and traffic-source NOx emission, remain highly relevant predictor variables across the different algorithms. Month (and its cosine-transformed numeric variable) is also an important variable. Meteorological variables, whose variable importance ranking was relatively low in the annual models, are also present among the most important variables, including air temperature (in SLMER, XGBoost, RF, LightGBM, NN), boundary layer height (in RF, XGBoost, LightGBM, NN), and total cloud cover (in SLR, SLMER, RF, XGBoost, LightGBM, NN).

3.3. Daily models

At daily scale the variability of NO_2 concentration is higher and the relationship between the variables is more complex, and in this case, it is observed that the boosting models are the best performing models. The predictions of the two gradient-boosting models, XGBoost and LightGBM, have the highest random-split CV-R^2 (XGB: 0.721; LGB: 0.724) and spatially-blocked CV-R^2 (XGB: 0.676; LGB: 0.673) (Fig. 2, Table A7). Following gradient boosting, mixed effect model (SLMER) has the third highest random-split CV-R^2 (0.686) and spatially-blocked CV-R^2 (0.660), higher than that of random forest. Random forest and LightGBM are the models with the highest temporally-blocked CV-R^2 (RF: 0.741; LGB: 0.742). SLR performs poorly at daily scale (CV-R^2s around 58%), suggesting that the daily variation is too complex to be predicted with a simple model like SLR. The performance of the mixed effect model improves compared to SLR when daily variation was taken into consideration as the random effect. Note that the temporal-CV R^2 is lower (0.400) because only the fixed-effect predictions (instead of random-effect BLUPs) of the mixed effect model can be estimated for temporally-blocked cross validation (see section 2.3.3). It can also be observed from the temporally-blocked CV that higher variability exists between months than within months (SLMER: pooled temporal-CV $R^2 = 0.400$, minimum temporal-CV $R^2 = 0.415$, maximum temporal-CV $R^2 = 0.638$). With lower CV-R^2 and higher CV-RMSE, neural network performed the worst among the four machine-learning algorithms. Note that the difference between the full-model (training) R^2 and CV-R^2 is largest for the two gradient-boosting models (XGBoost: full-model $R^2 = 0.927$, LightGBM: full-model $R^2 = 0.991$), suggesting an overfitting despite the cross-validation hyperparameter grid search. Including observation data of more years for training the models may improve the overfitting issue for a better generalization.

We aggregated the daily concentration predictions of the daily models to an annual average and compared the predicted values (full-model, random-split CV, spatially-blocked CV) to the annual average measured concentrations (Table A8). Compared to the models that were developed directly at the annual-average scale (Table A5), the CV- and spatial-CV- R^2 of the aggregated models are higher except for random forest and neural network. Particularly the spatially-blocked CV-R^2 of the two gradient boosting models (XGBoost, LightGBM) increased the most in the aggregated models (+5–13%). The aggregated XGBoost and LightGBM models have the highest CV- and spatial-CV- R^2 values, followed by SLR. This is somewhat different from the models developed directly at the annual-average scale, where random forest and SLR were the outperforming models. Besides the difference in data complexity as well as the corresponding suitable model complexity from annual to daily scale, a reason could be the difference in training data size for the data-hungry boosting learners. This suggests that with the accessibility and availability of data, modeling daily concentrations may still be useful even if the objective is modeling long-term ambient concentration. However, the two gradient boosting models highly overfit on the training data (full-model-$R^2 = 1.000$, RMSE<1 μg/m^3) whereas SLR did not overfit as strongly. Also, although SLR performed relatively poorly in predicting daily concentrations (CV- $R^2 < 0.60$), the aggregated predicted concentrations at annual-average scale highly correlate to the observed annual-average concentration (CV-$R^2 = 0.76$). This suggests that the linear-regression-based model is more capable of explaining spatial variations than temporal. On the other hand, the performances of random forest and neural network were better when the models were developed based on annual average than the annual average aggregated from daily estimates, therefore the choice of daily model for aggregated annual average concentration estimates may be model-specific.

Table A9 summarizes and compares several daily land-use statistical and machine-learning models for various pollutants from previous studies. The performance of random forest, gradient boosting and neural network models did not differ greatly in Di et al. (2019), whereas the difference is significant in this study. This difference may be related to the data-hungry nature of the machine-learning algorithms, the difference of study area sizes and the resulting difference in sample data size, with the contiguous United States being more than 200 times larger than Switzerland. The random-split CV-R^2 of the linear mixed effect models (SLMER) in this study are slightly higher than that of the mixed effect model in a previous study (de Hoogh et al., 2019) modeling daily NO_2 concentration across Switzerland (at 1×1 km; the "second-stage" model in that study). Anand and Monks (2017) developed similar mixed effect models based on supervised stepwise variables selection (Eeftens et al., 2012) for daily NO_2 concentrations in Hong Kong. The CV-R^2 (0.775) was about 9% higher than that of the mixed effect models in the current study, however the scale of the study area was much smaller (city-scale; with 11 monitoring stations) and the spatial resolution was coarser. Compared to the outperforming gradient boosting models, the mixed effect models in this study performed similarly in spatially-blocked cross validation and slightly lower in random-split cross validation. However, the mixed effect models still performed better than random forest and neural network. SLR performed poorly at daily scale, explaining only <60% of the variability of the spatiotemporal distribution of the daily concentrations. Few studies used linear regression models for modeling daily concentrations compared to annual. Rahman et al. (2017) applied a stepwise linear regression model to estimate daily NO_2 concentration in Brisbane metropolitan area and used a periodic function of "the day of year" and "the day of week" fitted with penalized splines to incorporate

seasonality. Similarly, the study showed a relatively low CV-R^2 (0.23; LOOCV) (note the limited number of monitoring sites in the study). Our results suggested that the high temporal variation of concentrations at a daily scale may be too challenging to be captured by SLR, a multiple linear regression with limited complexity and predictor variables, while the mixed effect models developed upon supervised stepwise variables selection can better model the variation by the adjustment with random-effect structure.

No clear pattern between the residuals and the model-fitted values were observed for the random forest, XGBoost and LightGBM models. An overestimation at high model-fitted values was observed in SLR and SLMER (Figure A3 (c)). A similar pattern of overestimation at higher model-fitted values was also observed in the temporally-blocked cross validation of the neural network models.

The relative variable importance of the satellite-observation NO_2 products is higher at daily scale compared to the monthly and annual models for each algorithm (Table A10; Figure A5). Date (or day of year and its cosine-transformed numeric variable) is another important variable for all models. Traffic intensity (at various radii of moving window size) is also a highly important predictor variable for all models. Boundary layer height and wind speed are the meteorological variables that are most important in the different models.

3.4. Uncertainties

Annual, monthly or daily models that were trained with the observations in spatially-blocked CV-folds 1, 3, 4, 5 and predicted on observations in fold 2 have the highest bias (fold-specific CV-RMSEs; yellow bars in Fig. 1(b)). In comparison, the differences between the CV-RMSEs of the other folds are relatively small. Figure A4 summarizes the fold-specific RMSE of the temporally-blocked cross validation of the monthly and daily models. It can be observed in all algorithms that the bias in winter months is higher than the bias in summer months. The highest bias exists in the hold-out predictions of February.

3.5. Model-estimated concentration surfaces

Fig. 3(a) shows the model-estimated annual average NO_2 concentrations at 100 × 100 m for 2019 zoomed in Zürich and the surrounding area with the Lake of Zürich in the middle. All models predicted high annual mean concentrations (~40 μg/m^3) at Zürich downtown and the surrounding heavy traffic corridors. More variations, mainly surrounding road networks, were depicted in the concentration surfaces estimated by SLR and NN, whereas the pattern was less obvious in the surfaces estimated by RF, XGBoost and LightGBM. RF and XGBoost models predict a smoother surface and wider decay of concentration from the major roads. One difference of the surfaces estimated by XGBoost and LightGBM to the other models is the island-like fragmented distribution of higher concentration areas, mainly along road networks. This pattern may be caused by the overfitting nature of gradient boosting models reflected in the relatively lower spatially-blocked CV. Furthermore, the Schwyz area, located at the bottom right corner of the maps, was estimated with lower concentration by SLR and NN compared to RF, XGBoost and particularly LightGBM. This points to a characteristic while using decision-tree-based models for regression tasks: the predicted values are never outside the training set values for the response variable. On the one hand, this means that the models would not predict negative or extremely high values as linear-regression models do. On the other hand, the predicted values are restricted to the range of the training data.

Similarly, in the model-predicted monthly average concentrations (Fig. 3 (c)), high concentrations were predicted at Zürich downtown and the surrounding heavy traffic corridors by all models. A seasonality of higher concentrations in winter months and lower concentrations in summer months can be observed. The seasonality was most obvious in the estimated surfaces of SLR and SLMER and was less in that of LightGBM. SLR and SLMER predicted more areas with negative concentration values, reflecting the observation of underestimation pattern at lower fitted values in the residual distributions (Figure A3). In comparison, the linear-regression-based models predict similar spatial patterns at a monthly level but with varying concentration levels, while for the machine-learning models, not only the levels vary between months but also the spatial patterns.

3.6. Limitations and potential improvements

Although the models presented here were able to explain a large portion of the NO_2 variation across Switzerland, some regions seemed to be more challenging to predict. The highest uncertainties for all models occurred in Ticino (Southwestern Switzerland) with a constant underestimation and higher prediction error in spatially-blocked CV (Fig. 2). We suspect that the pollution characteristics in Southern Switzerland may be different from that in the rest of the country, related to the area's geographical proximity to the Po Valley in northern Italy which is known for its high air pollutant levels and is considered an area with the worst air quality in Europe (Bigi et al., 2012). The Ticino basin located on the south of the Alps is climatically different from the Swiss Central Plateau with generally warmer temperatures and geographically more similarly to the Po valley. The overall higher pollution level in the area and difference in climate could result in the higher estimation bias in the area despite the inclusion of the satellite observation which may reflect the difference in pollution level.

In terms of temporal variation, February was most challenging to predict for all models (Figure A5). The estimated concentration levels of February 2019 were already the highest compared to the other months (Fig. 3), yet a constant underestimation was still observed. Temporally-blocked CV suggested that the behavior of NO_2 concentration in February was most different from the other months as the models trained with the observations of the other months generalized and predicted poorly on the observations in February (Figure A4).

This study has limitations, which may contribute to uncertainties of the modeling results and predictive performance. The spatial-temporal predictor variable, the meteorology dataset, is a product of climate reanalysis dataset combining global data from models and observations, which comes with uncertainties (Hersbach et al., 2018). Uncertainties also exist in the spatial predictor variables like land use, traffic, and emission inventories. For example, the NOx emission inventory data was conducted in 2015, and the road network and traffic volume database were from 2010. Any changes in the emissions and land use between the time the data was produced and 2019 are not reflected in the data, causing miss-alignment of real-world information in the models. There are also uncertainties in the satellite-derived column NO_2 data, including the NO_2 retrieval algorithm, the imputation models (full-model OOB-R^2: 0.898), and the CAMS modeled NO_2 data that was used in the imputation models. These uncertainties in the input data may introduce uncertainties of the modeling results.

Another limitation is related to the scale of this study. Switzerland is relatively small, and so is the heterogeneity across the study area. In this study, geographically weighted regression (GWR) performed not too different from basic multiple linear regression (Fig. 1(b), Table A5) and the spatial variability of the regression coefficients of GWR was low (Figure A2). For larger areas the difference could be more significant. The influence of landscape heterogeneity on model performance was observed in this study as the higher prediction errors occurred in Southern Switzerland. The higher uncertainties in Southern Switzerland with differing pollution characteristics showed a potential limitation of the transferability of the models. As traffic-related LUR models often perform poorly in areas with significant non-traffic sources (Beelen et al., 2010; Novotny et al., 2011), our models which were developed with the majority of the monitoring sites on the north of the Alps predicted poorly in Southern Alps areas where the contribution from emissions in the Po Valley were not captured. Marcon et al. (2015) also

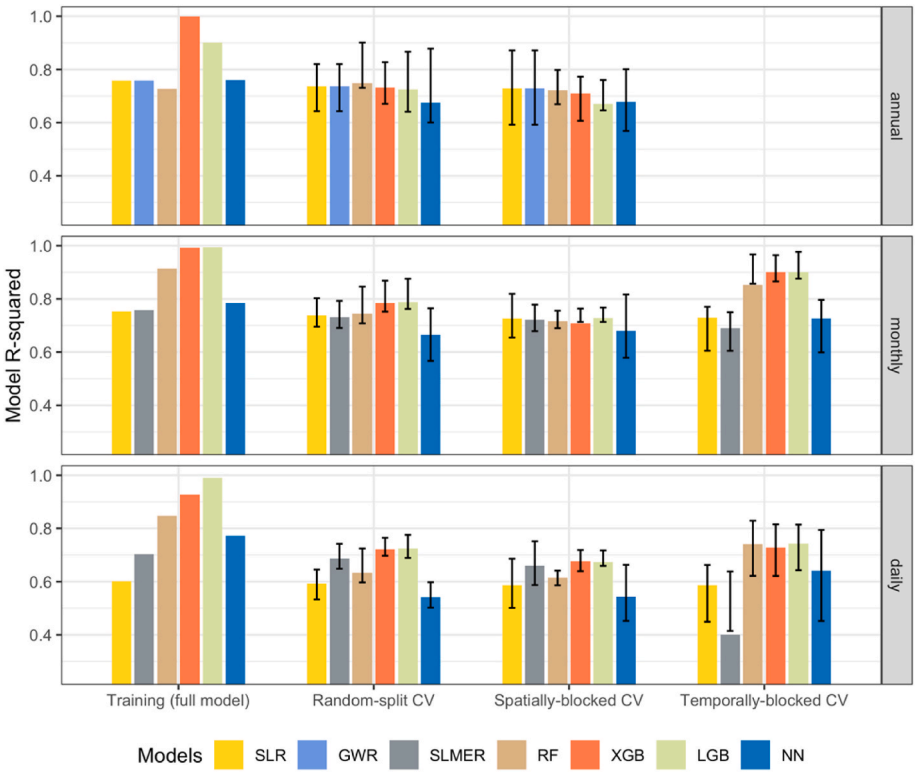

Fig. 2. A graphical summary of the model R-squared values at annual, monthly, and daily scale. The black lines indicate the maximum and minimum fold-specific R-squared values.

pointed out the limitation of transferring LUR models to nested areas with different characteristics. Therefore, the observations in this study may not be consistent if similar approaches were adopted for larger study areas or areas with different pollution characteristics.

The relative performance of the models depends on many factors, and the sample size used for training the models is highly relevant. The study developed models for only a year limited by the availability of the TROPOMI data. Including data of more years or more monitoring sites and the consequent larger sample size for model training may improve the performance of some models. For example, the performance of boosting models might be better with milder overfitting if the data size is bigger for better generalization. Besides sample size, we believe that temporal resolution (the long-/short-term variability) is still the most relevant factor to the performance of the models based on our observations in the study. The increasing complexity of the data from annual to daily temporal resolution can directly relate to the increasing complexity of the best-performing models.

There are some possible ways to improve the predictions, for example the development of an ensemble model aggregating the estimation of different models. Recent studies (Di et al., 2019, 2020) integrating random forest, gradient boosting and neural network with a geographically-weighted GAM ensemble model estimating NO_2 and $PM_{2.5}$ over the contiguous United States showed that despite the small incremental R-squared compared to the individual base learner models, the ensemble models resulted in a more linear relationship between measured and predicted NO_2 and $PM_{2.5}$ and were more stable across years, seasons, locations or pollution concentration levels (Di et al., 2019, 2020). The model projection maps (Fig. 3) also showed inconsistency of pollution patterns even between the best-performing models measured by cross-validation R^2 and RMSE, and overfitting gradient-boosting models may introduce noise in the estimated concentration surfaces. An ensemble model may be able to smooth the estimations of different models and achieve a more stabilized overall estimation. Furthermore, analysis of the time-series structure of the daily mean full-model residuals revealed the presence of periodicity and autoregressive structure in the residuals (Figure A5), indicating that there were still temporal variations remained unexplained by the models. The inclusion of temporal cosine waves as model predictor variables may be helpful to take into account the periodicity pattern and adjust the model estimations (Shumway and Stoffer, 2017). Including lagged predicted values in the models may also adjust the model estimates and correct the autoregressive structure in the residuals. For linear-regression-based models, generalized least square models with the assignment of AR(p) or ARMA(p,q) dependency structure could be considered as an alternative to adjust the models. For neural network, recurrent neural network (RNN), which uses sequential data or time-series data, is commonly used for time-series models in machine learning. RNNs are distinguished by their "memory" as they take information from prior inputs to influence the current input and output (Goodfellow et al., 2016). While recent studies (Liu et al., 2020; Tsai et al., 2018) have used RNNs for air pollution forecasting at fixed points, developing land-use air pollution models with RNNs for spatiotemporal distribution across surfaces are seldom investigated. Using RNNs may improve the model performance in time series modeling compared to the basic neural networks. Last but not least, static traffic volume data (version 2010) was used as a proxy of traffic-related emission sources. The common underestimation at the site "Camignolo", a traffic site in rural area, suggested a potential underestimation of traffic volume. The daily model residual time-series showed a weekly periodicity, which also exists in the pattern of traffic flow from weekdays to weekends. NO_2 is a highly traffic-relevant pollutant and the traffic-related variables showed high importance in the models. The inclusion of an updated or a dynamic instead of static traffic volume data may be helpful to reflect the variation and periodic pattern of emitted NO_2 in the models.

4. Conclusions

This study compared various land-use statistical and machine-

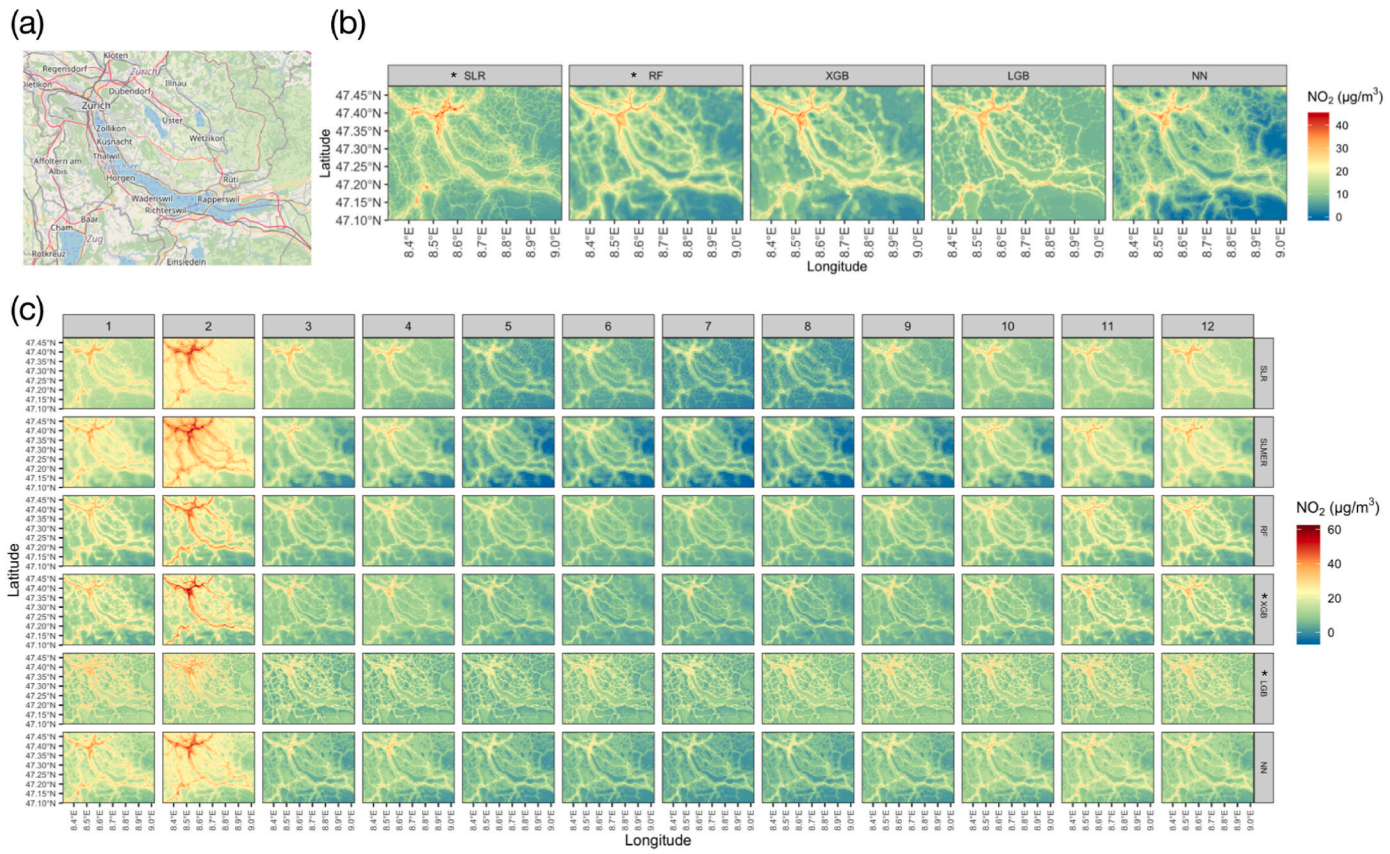

Fig. 3. The model-estimated NO_2 concentrations at 100 × 100 m for 2019 around Zürich. (a) An OpenStreetMap indicating the area (© OpenStreetMap contributors). (b) Annual average. The estimated concentration surface of GWR was not presented because the GWR models were almost identical to SLR given the low variability of regression coefficients in space. (c) Monthly average. Columns represent the months (1–12) and rows represent the different models. (*: best performing models).

learning models estimating ground-level NO_2 concentrations of 2019 across Switzerland at high spatial resolution (100 × 100 m) at annual, monthly and daily temporal resolution. Overall, an increasing complexity of the best performing models can be observed together with the increasing complexity of data from annual to daily temporal resolution. Linear-regression-based models are powerful in explaining long-term spatial distribution of NO_2, and when the sample size is limited to one single year like in this study. At annual- and monthly-average scale, spatially-blocked cross validation showed the robustness of the supervised stepwise linear models in space. Linear-regression-based models are, however, limited in explaining short-term temporal variation. Underestimation at low fitted values and overestimation at high fitted values exist in the prediction of linear-regression-based models at monthly and daily scale because the effects were modeled and extrapolated assuming linearity. For small study areas, like Switzerland, with limited heterogeneity, GWR may not predict better compared to basic multiple linear regression. In contrast, machine-learning models showed high capability in explaining short-term temporal variation and are particular helpful for modeling NO_2 at a high temporal resolution (daily). Some algorithm-specific characteristics were also observed. Gradient boosting machines (XGBoost and LightGBM) performed the best at daily scale but persistently overfit on the training data despite the control of overfitting with cross validation hyperparameter grid search. The overfitting nature of gradient boosting may result in unusual patterns in the model-estimated concentration surface. Random forest is relatively stable compared to gradient boosting despite lower statistical performance. Neural network performed not as high as the ensemble-tree models because of the smaller study area and sample size. For predicting daily concentrations, SLMER or XGBoost may be considered whilst acknowledging the potential overestimation at high levels for SLMER and the overfitting for XGBoost. In general, the temporal and spatial scale of the study is a particularly important factor on which the choice of the suitable model type should be based. Validation is required whatever approach is used and enlarging the sample size (e.g., data of more years) for model development can be favored to improve model predictive performance.

Credit author statement

Tze-Li Liu: Conceptualization, Methodology, Validation, Visualization, Formal analysis, Writing - Original Draft **Benjamin Flückiger**: Resources, Methodology, Writing - Review & Editing **Kees de Hoogh**: Conceptualization, Methodology, Supervision, Writing - Review & Editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.apr.2022.101611.

References

Alimissis, A., Philippopoulos, K., Tzanis, C., Deligiorgi, D., 2018. Spatial estimation of urban air pollution with the use of artificial neural network models. Atmos. Environ. 191, 205–213.

Allaire, J., Chollet, F., 2021. Keras. R Interface to 'Keras.

Anand, J.S., Monks, P.S., 2017. Estimating daily surface NO 2 concentrations from satellite data–a case study over Hong Kong using land use regression models. Atmos. Chem. Phys. 17, 8211–8230.

Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Software 67, 1–48.

Beelen, R., 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 383, 785–795.

Beelen, R., Voogt, M., Duyzer, J., Zandveld, P., Hoek, G., 2010. Comparison of the performances of land use regression modelling and dispersion modelling in estimating small-scale variations in long-term air pollution concentrations in a Dutch urban area. Atmos. Environ. 44, 4614–4621.

Bigi, A., Ghermandi, G., Harrison, R.M., 2012. Analysis of the air pollution climate at a background site in the Po valley. J. Environ. Monit. 14, 552–563.

Bishop, C.M., 1995. Neural Networks for Pattern Recognition. Oxford university press.

Bovensmann, H., et al., 1999. SCIAMACHY: mission objectives and measurement modes. J. Atmos. Sci. 56, 127–150.

Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.

Briggs, D.J., et al., 1997. Mapping urban air pollution using GIS: a regression-based approach. Int. J. Geogr. Inf. Sci. 11, 699–718.

Cabaneros, S.M., Calautit, J.K., Hughes, B., 2020. Spatial estimation of outdoor NO2 levels in Central London using deep neural networks and a wavelet decomposition technique. Ecol. Model. 424, 109017.

Cabaneros, S.M., Calautit, J.K., Hughes, B.R., 2019. A review of artificial neural network models for ambient air pollution prediction. Environ. Model. Software 119, 285–304.

Chen, J., et al., 2019a. A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide. Environ. Int. 130, 104934.

Chen, J., et al., 2018. OP III–4 Exposure Assessment Models for No2 and Pm2. 5 in the Elapse Study: a Comparison of Supervised Linear Regression and Machine Learning Approaches. BMJ Publishing Group Ltd.

Chen, T., He, T., 2015. Xgboost: Extreme Gradient Boosting, pp. 1–4. R package version 0.4-2 1.

Chen, T., He, T., Benesty, M., Khotilovich, V., 2019b. Package 'xgboost'. R version 90.

Chen, T.-M., Kuschner, W.G., Gokhale, J., Shofer, S., 2007. Outdoor air pollution: nitrogen dioxide, sulfur dioxide, and carbon monoxide health effects. Am. J. Med. Sci. 333, 249–256.

Copernicus sentinel-5P (processed by ESA), 2021. TROPOMI level 2 nitrogen dioxide total column products. In: European Space Agency.

de Hoogh, K., et al., 2016. Development of West-European PM2. 5 and NO2 land use regression models incorporating satellite-derived and chemical transport modelling data. Environ. Res. 151, 1–10.

de Hoogh, K., Héritier, H., Stafoggia, M., Künzli, N., Kloog, I., 2018. Modelling daily PM2. 5 concentrations at high spatio-temporal resolution across Switzerland. Environ. Pollut. 233, 1147–1154.

de Hoogh, K., et al., 2019. Predicting fine-scale daily NO2 for 2005–2016 incorporating OMI satellite data across Switzerland. Environ. Sci. Technol. 53, 10279–10287.

Di, Q., et al., 2019. An ensemble-based model of PM2. 5 concentration across the contiguous United States with high spatiotemporal resolution. Environ. Int. 130, 104909.

Di, Q., et al., 2020. Assessing NO2 concentration and model uncertainty with high spatiotemporal resolution across the contiguous United States using ensemble model averaging. Environ. Sci. Technol. 54, 1372–1384.

Di, Q., et al., 2016. Assessing PM2. 5 exposures with high spatiotemporal resolution across the continental United States. Environ. Sci. Technol. 50, 4712–4721.

Didan, K., 2015. In: DAAC, N.E.L.P. (Ed.), MYD13Q1 MODIS/Aqua Vegetation Indices 16-Day L3 Global 250m SIN Grid V006.

Eeftens, M., et al., 2012. Development of land use regression models for PM2. 5, PM2. 5 absorbance, PM10 and PMcoarse in 20 European study areas; results of the ESCAPE project. Environ. Sci. Technol. 46, 11195–11205.

Eskes, H., et al., 2020. Sentinel-5 Precursor/TROPOMI Level 2 Product User Manual Nitrogendioxide. Royal Netherlands Meteorological Institute, Ministry of Infrastructure and Water Management.

Garson, D.G., 1991. Interpreting neural network connection weights. Artif. Intell. Expert 6, 46–51.

Goh, A.T., 1995. Back-propagation neural networks for modeling complex systems. Artif. Intell. Eng. 9, 143–151.

Gollini, I., Lu, B., Charlton, M., Brunsdon, C., Harris, P., 2015. GWmodel: an R package for exploring spatial heterogeneity using geographically weighted models. J. Stat. Software 63, 1–50.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT press.

Gulliver, J., de Hoogh, K., 2015. Environmental exposure assessment: modelling air pollution concentrations. In: Oxford Textbook of Global Public Health, 6 ed. Oxford University Press. Tan, R.D.M.G.Q.A.K.C.C.

Guyon, I., Elisseeff, A., 2003. An introduction to variable and feature selection. J. Mach. Learn. Res. 3, 1157–1182.

Hersbach, H., et al., 2018. ERA5 hourly data on single levels from 1979 to present. Copernicus Clim. Change Serv. (C3S) Clim. Data Store (CDS).

Hoek, G., et al., 2008. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmos. Environ. 42, 7561–7578.

Hu, X., et al., 2017. Estimating PM2. 5 concentrations in the conterminous United States using the random forest approach. Environ. Sci. Technol. 51, 6936–6944.

Hu, Z., 2009. Spatial analysis of MODIS aerosol optical depth, PM 2.5, and chronic coronary heart disease. Int. J. Health Geogr. 8, 1–10.

Ke, G., et al., 2017. Lightgbm: a highly efficient gradient boosting decision tree. Adv. Neural Inf. Process. Syst. 30, 3146–3154.

Ke, G., et al., 2021. Lightgbm: light gradient boosting machine. R package version 3.2.1.

Lamsal, L.N., et al., 2020. OMI/Aura NO2 tropospheric, stratospheric & total columns MINDS daily L3 global gridded 0.25 degree x 0.25 degree. In: NASA Goddard Space Flight Center. Goddard Earth Sciences Data and Information Services Center (GES DISC).

Lee, J.-H., et al., 2014. Land use regression models for estimating individual NOx and NO2 exposures in a metropolis with a high density of traffic roads and population. Sci. Total Environ. 472, 1163–1171.

Liaw, A., Wiener, M., 2002. Classification and regression by randomForest. R. News 2, 18–22.

Liu, D.R., Lee, S.J., Huang, Y., Chiu, C.J., 2020. Air pollution forecasting based on attention-based LSTM neural network and ensemble learning. Expet Syst. 37, e12511.

Lu, M., Schmitz, O., de Hoogh, K., Kai, Q., Karssenberg, D., 2020. Evaluation of different methods and data sources to optimise modelling of NO2 at a global scale. Environ. Int. 142, 105856.

Marcon, A., de Hoogh, K., Gulliver, J., Beelen, R., Hansell, A.L., 2015. Development and transferability of a nitrogen dioxide land use regression model within the Veneto region of Italy. Atmos. Environ. 122, 696–704.

Mlakar, P., Boznar, M.Z., 2011. Artificial neural networks-a useful tool in air pollution and meteorological modelling. In: Advanced Air Pollution. IntechOpen.

Montagne, D., et al., 2013. Agreement of land use regression models with personal exposure measurements of particulate matter and nitrogen oxides air pollution. Environ. Sci. Technol. 47, 8523–8531.

Novotny, E.V., Bechle, M.J., Millet, D.B., Marshall, J.D., 2011. National satellite-based land-use regression: NO2 in the United States. Environ. Sci. Technol. 45, 4407–4414.

Rahman, M.M., Yeganeh, B., Clifford, S., Knibbs, L.D., Morawska, L., 2017. Development of a land use regression model for daily NO2 and NOx concentrations in the Brisbane metropolitan area, Australia. Environ. Model. Software 95, 168–179.

Roberts, D.R., et al., 2017. Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure. Ecography 40, 913–929.

Röösli, M., Vienneau, D., 2014. Epidemiological exposure assessment. Epidemiol. Electromagnet. Fields 35–55.

Samet, J.M., Utell, M.J., 1990. The risk of nitrogen dioxide: what have we learned from epidemiological and clinical studies? Toxicol. Ind. Health 6, 247–262.

Shumway, R.H., Stoffer, D.S., 2017. Time Series Analysis and its Applications with R Examples. Springer, New York.

Stafoggia, M., et al., 2019. Estimation of daily PM10 and PM2. 5 concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model. Environ. Int. 124, 170–179.

Thapa, R.B., Estoque, R.C., 2012. Geographically weighted regression in geospatial analysis. In: Progress in Geospatial Analysis. Springer, pp. 85–96.

Tsai, Y.-T., Zeng, Y.-R., Chang, Y.-S., 2018. Air pollution forecasting using RNN with LSTM. In: 2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/DataCom/CyberSciTech). IEEE, pp. 1074–1079.

Van Roode, S., Ruiz-Aguilar, J., González-Enrique, J., Turias, I., 2019. An artificial neural network ensemble approach to generate air pollution maps. Environ. Monit. Assess. 191, 1–15.

Veefkind, J., et al., 2012. TROPOMI on the ESA Sentinel-5 Precursor: a GMES mission for global observations of the atmospheric composition for climate, air quality and ozone layer applications. Remote Sens. Environ. 120, 70–83.

World Health Organization, 2006. Air Quality Guidelines: Global Update 2005: Particulate Matter, Ozone, Nitrogen Dioxide, and Sulfur Dioxide. World Health Organization.

Wright, M.N., Ziegler, A., ranger, 2015. A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *arXiv preprint arXiv:1508.04409*.

Yassi, A., Kjellström, T., De Kok, T., Guidotti, T.L., 2001. Basic Environmental Health. Oxford University Press, USA.