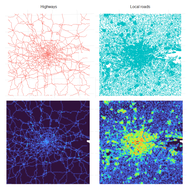
Ensemble machine learning and multi-stage data reconstruction methods for high-resolution air pollution mapping in Great Britain (2003-2021)

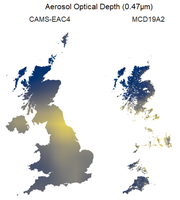
Arturo de la Cruz Libardi [a], Pierre Masselot [a], Rochelle Schneider [ef], Emily Nightingale [c], Ai Milojevic [bd], Jacopo Vanoli [ag], Malcolm Mistry [ah], Antonio Gasparrini [a]

Introduction

- Air pollution poses a significant risk to human health with exposure to particulate matter and nitrogen dioxide being independently linked to increased mortality and morbidity.
- Health research on air pollution effects requires accurate measurement and modelling of the pollutant concentrations and distributions.
- We aim to apply advanced machine learning-ensemble methods to reconstruct daily concentrations of NO_2, PM_{10}, and $PM_{2.5}$, over Great Britain within the period 2003-2021 in a grid with a 1x1km resolution.



Real-world features such as roads were transformed from non-gridded (vector) format to a gridded (raster) variant, representing their density.



Aerosol Optical depth is a measure of light extinction in the total atmospheric column. It is unitless and wavelength-dependent.

Data

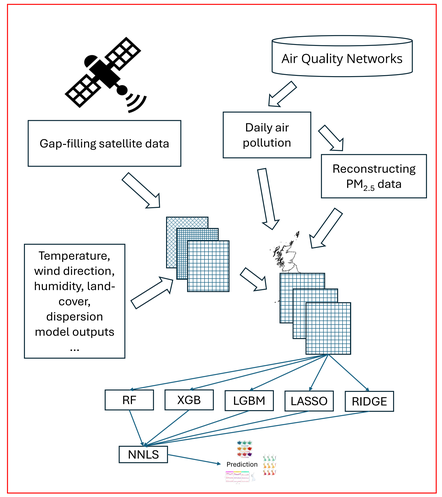
The developed model uses station measurements as the target, and an extensive set of land-use, atmospheric and demographic data as predictors.

Geographical and temporal scope: The 242,851 1km^2 grid cells covering GB yearly throughout 2003 to 2021.

Air pollution data: Daily-averaged ground station measurements from national and European air quality networks.

Spatio-temporal: Meteorological and atmospheric composition data from reanalysis, dispersion models, and satellite sources.

Spatial: Elevation, land cover, night-light radiance, population density, road and traffic data.



Methods

1. Sparse $PM_{2.5}$ data was reconstructed via co-located PM_{10} observations with a machine learning model.

2. Likewise, two satellite data products, Aerosol Optical Depth, and NO_2 column, were gap-filled using reanalysis products as predictors in machine-learning models

3. After all features were harmonized and linked to pollutant observations, the assembled dataset was used to evaluate and train five base-learners and a meta-learner creating an ensemble-model (also known as Super Learner).
   The base-learners used were random forest (RF), extreme gradient boosting (XGB), light gradient boosting machine (LGBM), ridge and lasso. Predictions from these models were aggregated by a non-negative least squares (NNLS) model which calculated the optimal positive and location- independent weighting for each base-learner from their cross-validated predictions on observed data.

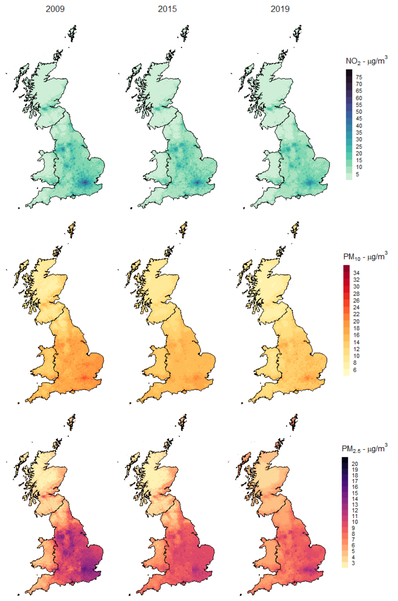
4. Finally, the trained ensemble was used to obtain predictions over the study area.

Results

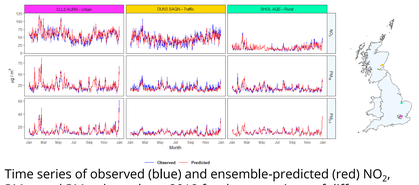
The ensemble models showed good performance across all pollutants with total mean R^2 values of 0.690 (min. year 0.611 - max. year 0.792) for NO_2, 0.704 (0.609-0.786) for PM_{10}, and 0.820 (0.746-0.888) for $PM_{2.5}$. Performance was highest in the most recent period (2015-2021).

The ensemble had generally low bias as indicated by the intercept and slope values obtained by comparing observed and cross-validated predictions and shown in the table and scattered-density plot.

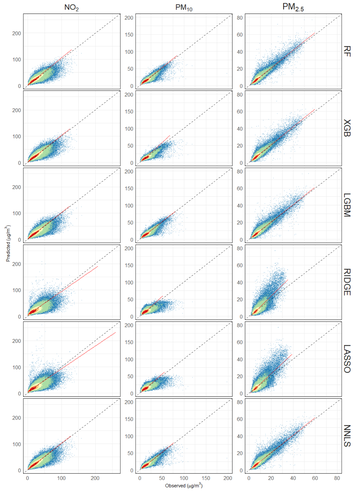
Daily variations are visible in the Greater London panels and follow closely the observations as shown in the yearly time-series figure. Spatial patterns are also captured at a national and regional scale.



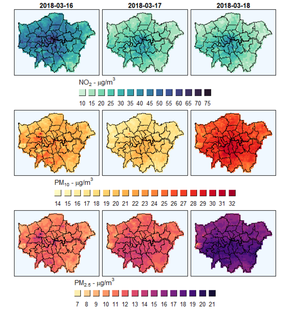
Annual average maps.



Time series of observed (blue) and ensemble-predicted (red) NO_2, PM_{10}, and $PM_{2.5}$ throughout 2018 for three monitors of different types. Levels are in micrograms per cubic metre and are shown at the corresponding grid-cell locations in Great Britain.



Hexagon-binned density and regression line of cross-validated predictions against observed values in 2019. All base-learners and the ensemble (NNLS) are shown for each pollutant (NO_2, PM_{10} and $PM_{2.5}$). Bisector line in dashed black. Axis limits are symmetrical and fixed by pollutant for ease of comparison.



Daily predicted concentrations of NO_2, PM_{10}, and $PM_{2.5}$ from Friday 16th to Sunday 18th of March 2018.

10-fold monitor-blocked cross-validated ensemble (NNLS) performance by period means.

	Period	R^2	RMSE	Inter.	Slope
NO_2	2003-2008	0.705	12.983	1.764	1.004
	2009-2014	0.653	14.103	2.302	1.000
	2015-2021	0.712	10.080	0.920	1.023
	Mean	0.690	12.389	1.662	1.009
PM_{10}	2003-2008	0.679	7.702	0.301	1.024
	2009-2014	0.700	6.372	0.153	1.033
	2015-2021	0.735	5.041	0.164	1.029
	Mean	0.704	6.371	0.206	1.029
$PM_{2.5}$	2003-2008	0.789	3.220	0.090	1.020
	2009-2014	0.816	3.570	0.119	1.024
	2015-2021	0.856	2.635	0.107	1.022
	Mean	0.820	3.142	0.105	1.022

Conclusion

Over 5 billion data points were generated with high cross-validated accuracy and resolution. These data can be linked to health datasets and contribute to environmental health research. Limitations of this study include the unbalanced distribution of ground monitoring stations and the low resolution of predictor features.
Further studies could explore novel model and map evaluation strategies as well as data and methods to increase the spatial resolution.

Contact: Arturo.de-la-Cruz-Libardi@lshtm.ac.uk

Download the references (pdf)

https://adlcruz.github.io/assets/poster_turingcliminf_2024/poster_turingcliminf_refs_adlcl.pdf

Affiliations:
a) Environment & Health Modelling (EHM) Lab, Department of Public Health Environments and Society, London School of Hygiene & Tropical Medicine, 15-17 Tavistock Place, WC1H 9SH London, United Kingdom
b) Department of Public Health, Environments and Society, London School of Hygiene & Tropical Medicine, 15-17 Tavistock Place, WC1H 9SH London, United Kingdom
c) Department of Global Health and Development, London School of Hygiene & Tropical Medicine, Keppel Street, WC1E 7HT, London, United Kingdom
d) Centre on Climate Change & Planetary Health, London School of Hygiene & Tropical Medicine, Keppel Street, WC1E 7HT, London, United Kingdom
e) Φ-lab (Phi-lab), European Space Agency (ESA), Frascati, Italy
f) Forecast Department, European Centre for Medium-Range Weather Forecast (ECMWF), Reading, United Kingdom
g) School of Tropical Medicine and Global Health, Nagasaki University, Nagasaki, Japan
h) Department of Economics, Ca' Foscari University of Venice, Italy