Spatial $PM_{2.5}$, NO_2, O_3 and BC models for Western Europe – evaluation of spatiotemporal stability

Kees de Hoogh[a,b]
a. Swiss Tropical and Public Health Institute, Socinstrasse 57, 4051 Basel, Switzerland
b. University of Basel, Petersplatz 1, Postfach 4001, Basel, Switzerland
c.dehoogh@swisstph.ch

Jie Chen[c]
c. Institute for Risk Assessment Sciences (IRAS), Utrecht University, Postbus 80125, 3508 TC, Utrecht, the Netherlands
j.chen1@uu.nl

John Gulliver[d]
d. School of Geography, Geology and the Environment, University of Leicester, University Road, Leicester, LE1 7RH, UK
jg435@leicester.ac.uk

Barbara Hoffmann[e]
e. Institute for Occupational, Social and Environmental Medicine, Centre for Health and Society, Medical Faculty, Heinrich Heine University Düsseldorf, Universitätsstraße 1, 40225 Düsseldorf, Germany
Barbara.Hoffmann@iuf-duesseldorf.de

Ole Hertel[f]
f. Department of Environmental Science, Aarhus University, 4000 Roskilde, Denmark
oh@envs.au.dk

Matthias Ketzel[f]
f. Department of Environmental Science, Aarhus University, 4000 Roskilde, Denmark
mke@envs.au.dk

Mariska Bauwelinck[g.h]
g. Interface Demography – Department of Sociology, Vrije Universiteit Brussel, Boulevard de la Plaine 2, 1050 Ixelles, Brussel, Belgium.
h. Unit Health & Environment — Sciensano, Rue Juliette Wytsmanstraat 14, 1050 Brussels, Belgium.
mariska.bauwelinck@vub.ac.be

Aaron van Donkelaar[i]
i. Department of Physics and Atmospheric Science, Dalhousie University, B3H 4R2 ,Halifax, Nova Scotia, Canada
kelaar@Dal.Ca

Ulla A. Hvidtfeldt[j]
j. Danish Cancer Society Research Center, Strandboulevarden 49, 2100 Copenhagen, Denmark
ullah@cancer.dk

Klea Katsouyanni[k,l]
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
kkatsouy@med.uoa.gr

Jochem Klompmaker[c,m]
c. Institute for Risk Assessment Sciences (IRAS), Utrecht University, Postbus 80125, 3508 TC, Utrecht, the Netherlands
m. . National Institute for Public Health and the Environment (RIVM), Antonie van Leeuwenhoeklaan 9, 3721 MA Bilthoven, Netherlands
jochem.klompmaker@rivm.nl

Randal V. Martin[i,n]
i. Department of Physics and Atmospheric Science, Dalhousie University, B3H 4R2 ,Halifax, Nova Scotia, Canada
n. Atomic and Molecular Physics Division, Harvard-Smithsonian Center for Astrophysics, Cambridge, 60 Garden St, Cambridge, MA 02138, US
Randall.Martin@Dal.Ca

Evangelia Samoli[k]
k. Department of Hygiene, Epidemiology and Medical Statistics, Medical School, National and Kapodistrian University of Athens, 75 Mikras Asias Str, 115 27 Athens, Greece
esamoli@med.uoa.gr

Per E. Schwartz[o]
o. Division of Environmental Medicine, Norwegian Institute of Public Health, PO Box 4404 Nydalen N-0403 Oslo, Norway
Per.Schwarze@fhi.no

Massimo Stafoggia[p,q]
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
m.stafoggia@deplazio.it

Tom Bellander[q]
q. Institute of Environmental Medicine, Karolinska Institutet, SE-171 77, Stockholm, Sweden
Tom.Bellander@ki.se

Maciek Strak[c]
c. Institute for Risk Assessment Sciences (IRAS), Utrecht University, Postbus 80125, 3508 TC, Utrecht, the Netherlands
M.M.Strak@uu.nl

Kathrin Wolf[r]
r. Helmholtz Zentrum München, German Research Center for Environmental Health (GmbH), Institute of Epidemiology II, Ingolstädter Landstr. 1, D-85764 Neuherberg, Germany
kathrin.wolf@helmholtz-muenchen.de

Danielle Vienneau[a,b]
a. Swiss Tropical and Public Health Institute, Socinstrasse 57, 4051 Basel, Switzerland
b. University of Basel, Petersplatz 1, Postfach 4001, Basel, Switzerland
danielle.vienneau@swisstph.ch

Bert Brunekreef[c,s]

c. Institute for Risk Assessment Sciences (IRAS), Utrecht University, Postbus 80125, 3508 TC, Utrecht, the Netherlands
s. Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX Utrecht, Netherlands
B.Brunekreef@uu.nl

Gerard Hoek[c]
c. Institute for Risk Assessment Sciences (IRAS), Utrecht University, Postbus 80125, 3508 TC, Utrecht, the Netherlands
G.Hoek@uu.nl

Corresponding author:

Kees de Hoogh, Department of Epidemiology and Public Health, Swiss Tropical and Public Health Institute, Socinstrasse 57, 4051 Basel, Switzerland, tel: +41 (0)61 284 8749
c.dehoogh@swisstph.ch

Abstract

Background

In order to investigate associations between air pollution and adverse health effects consistent fine spatial air pollution surfaces are needed across large areas to provide cohorts with comparable exposures. The aim of this paper is to develop and evaluate fine spatial scale land use regression models for four major health relevant air pollutants ($PM_{2.5}$, NO_2, BC, O_3) across Europe.

Methods

We developed West-European land use regression models (LUR) for 2010 estimating annual mean $PM_{2.5}$, NO_2, BC and O_3 concentrations (including cold and warm season estimates for O_3). The models were based on AirBase routine monitoring data ($PM_{2.5}$, NO_2 and O_3) and ESCAPE monitoring data (BC), and incorporated satellite observations, dispersion model estimates, land use and traffic data. Kriging was performed on the residual spatial variation from the LUR models and added to the exposure estimates. One model was developed using all sites (100%). Robustness of the models was evaluated by performing a five-fold hold-out validation and for $PM_{2.5}$ and NO_2 additionally with independent comparison at ESCAPE measurements. To evaluate the stability of each model's spatial structure over time, separate models were developed for different years (NO_2 and O_3: 2000 and 2005; $PM_{2.5}$: 2013).

Results

The $PM_{2.5}$, BC, NO_2, O_3 annual, O_3 warm season and O_3 cold season models explained respectively 72%, 54%, 59%, 65%, 69% and 83% of spatial variation in the measured concentrations. Kriging proved an efficient technique to explain a part of residual spatial variation for the pollutants with a strong regional component explaining respectively 10%, 24% and 16% of the R^2 in the $PM_{2.5}$, O_3 warm and O_3 cold models. Explained variance at fully independent sites vs the internal hold-out validation was slightly lower for $PM_{2.5}$ (65% vs 66%) and lower for NO_2 (49% vs 57%). Predictions from the 2010 model correlated highly with models developed in other years at the overall European scale.

Conclusions

We developed robust $PM_{2.5}$, NO_2, O_3 and BC hybrid LUR models. At the West-European scale models were robust in time, becoming less robust at smaller spatial scales. Models were applied to 100x100 m surfaces across Western Europe to allow for exposure

assignment for 35 million participants from 18 European cohorts participating in the ELAPSE study.

Keywords: LUR, spatiotemporal stability, $PM_{2.5}$, NO_2, ozone, black carbon

Abbreviations

CTM	Chemical Transport Models
SAT	Satellite-derived predictions
FULL	Models developed using 100% of the monitoring sites
HOV	Hold-Out-Validation models developed on 80% of the number of sites

Highlights

1. Robust $PM_{2.5}$, NO_2, BC and O_3 hybrid LUR models at a 100x100 m resolution for Western Europe were developed
2. Models included large scale satellite and chemical transport model estimates and fine scale traffic and land use and were further improved with kriging
3. Models were robust in time at European scale, becoming less robust at smaller spatial scales.

1. Introduction

Ambient air pollution remains one of the main causes of morbidity and mortality in the world (Cohen et al. 2017). WHO's global assessment of ambient air pollution exposure estimated that one in nine deaths annually are caused by ambient air pollution (WHO 2016). More recently, there is evidence showing that associations between mortality and morbidity and long-term exposure to outdoor air pollution might have no threshold, and extend to concentrations below current air quality limit values of the US EPA and EU (Beelen et al. 2015). Recent studies conducted in North-America have shown long-term exposure to $PM_{2.5}$ is associated with mortality also at low exposures (i.e. below the current WHO guideline of 10 µg/m^3) (Crouse et al. 2015; Di et al. 2017; Pinault et al. 2017). Particularly in North-America and Europe, tougher air quality policies have led to a reduction in emissions and a gradual decline in ambient air pollution concentrations (EEA 2017). Little, however, is known about the shape of the exposure-response curve at low concentrations, and thus the impact of low level concentrations on large populations remains uncertain.

The ELAPSE (Effects of Low-Level Air Pollution: A Study in Europe) study aims to fill this gap by investigating the relationship between long term air pollution and morbidity and mortality at low $PM_{2.5}$ (Particulate Matter <2.5 µg), nitrogen dioxide (NO_2), black carbon (BC) and ozone (O_3) exposures. Low levels are defined as air pollutant concentrations below EU and/or US air quality limit values and/or WHO guidelines. ELAPSE includes 11 cohorts with in-depth individual data on lifestyle and 7 large administrative/national cohorts across Europe (http://www.elapseproject.eu/). Cohorts were selected to represent a contrast in air pollution exposures between and within study areas. The 11 detailed individual-level cohorts will be analyzed as a pooled cohort, whereas the administrative cohorts will be analyzed separately. Taken together, the evidence should allow collective consideration and evaluation. This study therefore needs consistent models that can provide valid exposures at two different spatial extents in a Western Europe: combining all study regions of the detailed individual-level cohorts for the pooled analysis; and the national extents for the administrative/national cohorts. The previously developed ESCAPE LUR models (Beelen et al. 2013; Eeftens et al. 2012a) do not meet the requirements for the ELAPSE project because they do not cover the full national study areas. Secondly, methodological work by Basagana and Wang has shown that more stable models can be developed based on larger number of model training sites than the 20 sites that the ESCAPE PM models were based upon (Basagaña et al. 2012; M Wang et al. 2013). Finally, ESCAPE did not evaluate Ozone."

Cohorts in the ELAPSE study have different recruitment and follow-up periods going back as early as the 1990's. Epidemiological studies have used the back-extrapolation method to estimate exposures back in time (Beelen et al. 2014; Chen et al. 2017). The method uses a well validated air pollution surface as the base and assumes that the spatial structure of this surface remains stable over time. Monitoring data from routine monitoring sites are then used to re-scale the surface back or forward in time (Cesaroni et al. 2012; Chen et al. 2010). Few studies have been able to document the stability of spatial surfaces, mostly focusing on NO_2 and at the city level (Cesaroni et al. 2012; Eeftens et al. 2011; R Wang et al. 2013) or national scale (Gulliver et al. 2013). We thus evaluated the stability of these surfaces over time by comparing modelled estimates with historic monitoring data and by developing models for other years.

The aims of the paper are to:

1. develop and evaluate performance of fine spatial scale hybrid land use regression models for four major health relevant pollutants $PM_{2.5}$, NO_2, BC, O_3 across Western Europe;
2. investigate the temporal stability of the spatial contrast at the West-European and national scale.

This paper follows our recently published West-European fine scale air pollution exposure models for $PM_{2.5}$ and NO_2 (de Hoogh et al. 2016). Models were based on both 2010 ESCAPE and the European Environment Agency (EEA) AirBase routine monitoring data, and documented the contribution of satellite data and chemical transport models (CTM) to LUR models. An important finding was that models performed well when validated with data from the other measurement network (i.e. ESCAPE model validated with AirBase sites and vice versa). In the current paper we substantially extended this work, firstly by adding black carbon (BC) and ozone (O_3) which are both health relevant pollutants. We also improved the testing of the robustness of models by evaluating structure and predictions using five-fold hold-out-validation (HOV), following a study on land use regression models for ultrafine particles (van Nunen et al. 2017). We further assessed improving the LUR models using kriging and added new predictor variables with improved granularity, including 1x1 km satellite $PM_{2.5}$ to the previously used 10x10 km satellite data. Finally we added an assessment of the temporal stability of the models.

2. Materials and methods

2.1 Air pollution monitoring data

$PM_{2.5}$, NO_2 and O_3 daily concentration data for 2010 were derived from the AirBase v8 dataset (EEA). Only sites with ≥ 75% completeness of the total hours (NO_2 and O_3) or days ($PM_{2.5}$) were accepted, and an annual average was calculated for $PM_{2.5}$ and NO_2. For O_3, we calculated the maximum running 8-hour mean for each day and then averaged to obtain an annual, warm season (April through September) and cold season (January through March and October through December) average maximum running 8-hour mean. For BC, which is not available through AirBase, we used the ESCAPE annual mean BC concentrations (measured as $PM_{2.5}$ absorbance based on reflectance measurement of the filters) reflecting the time period 2009-2010. A detailed description of the ESCAPE measurement campaign can be found elsewhere (Eeftens et al. 2012b). Table S1 describes the number of sites and summary statistics of the air pollution measurement data. The locations of the monitoring sites used for the 2010 models are shown in Figure S1. For temporal stability analysis we additionally included NO_2 and O_3 daily concentration data for 2000 and 2005 from AirBase v8 and daily $PM_{2.5}$ concentration data for 2013 from Air Quality e-Reporting (www.eea.europa.eu/data-and-maps/data/aqereporting-8). There were insufficient $PM_{2.5}$ sites across Western Europe before 2010.

2.2 Predictor variables

2.2.1 Satellite derived air pollution data

In addition to the satellite (SAT) $PM_{2.5}$ product (v3.01) used in the previous paper (de Hoogh et al. 2016), we tested two additional different SAT $PM_{2.5}$ products, which have become available only recently, as potential predictors. These were obtained from the global dataset reported in Van Donkelaar et al. (2015). Aerosol Optical Depth (AOD) retrievals from the NASA MODIS (Moderate Resolution Imaging Spectroradiometer), MISR (Multi-angle Imaging Spectroradiometer) and SeaWiFS instruments were related to near-surface concentrations using aerosol vertical profiles and scattering properties simulated by the GEOS-Chem CTM, to produce an annual average $PM_{2.5}$ dataset at a 0.1° x 0.1° (~10km) resolution for 2010. In the previous paper we used a dataset inferred from 2009-2011 (optimized for 2010), here we additionally tested the inferred data from 2010 data only. We further included the current, purely geophysical, global $PM_{2.5}$ dataset (V4.GL.02.NoGWR), which includes some information at the finer resolution of 0.01° x 0.01° (~1km) published by van Donkelaar et al. (2016). The pre-Geographically Weighted Regression dataset used here includes AOD from multiple satellite products (MISR, MODIS Dark Target, MODIS and SeaWiFS Deep Blue, and MODIS MAIAC) together with simulation-based sources, with information content below ~10km provided by the MAIAC AOD retrieval. $PM_{2.5}$ satellite data was offered as a predictor to the $PM_{2.5}$ models. No BC satellite data were available and because BC is a major component of $PM_{2.5}$, $PM_{2.5}$ satellite data were also offered to the BC models.

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. Like $PM_{2.5}$, the satellite column-integrated retrievals were related to ground-level concentrations using the global GEOS-Chem model, producing an annual gridded NO_2 surface at a 10km resolution (Bechle et al. 2013, 2015; Novotny et al. 2011). NO_2 satellite predictors were offered to the NO_2 models. No O_3 satellite data were available but, because NO_2 is related to O_3 formation and scavenging, NO_2 satellite data was also offered to the O_3 models.

2.2.2 Chemical transport model (CTM) data

Pollutant estimates for 2010 from two long range CTM's were obtained as potential predictor variables for the models. Annual $PM_{2.5}$, NO_2 and O_3 estimates were derived from the MACC-II ENSEMBLE model at a 0.1º x 0.1º (~10km) resolution (Inness et al. 2013). The ENSEMBLE model provides a value at each pixel which is defined as the median value of seven individual CTMs: CHIMERE, EMEP, EURAD, LOTOS-EUROS, MATCH, MOCAGE and SILAM. Annual MACC-II ENSEMBLE averages for $PM_{2.5}$, NO_2 and O_3 were offered to the respective LUR models. We additionally acquired a second CTM dataset from the Danish Eulerian Hemispheric Model (DEHM_v31102016) for $PM_{2.5}$, NO_2, O_3 and BC at a monthly 50x50 km resolution (Brandt et al. 2012). Annual DEHM averages were calculated for all pollutants and offered to the respective LUR models, while warm and cold averages of O_3 were offered to the warm and cold season models.

2.2.3 Other predictor variables

The GIS predictor variables used in this study are described in more detail elsewhere (de Hoogh et al. 2016; Vienneau et al. 2013). In brief, road data, classified as 'all' and 'major' roads, were extracted from the 1:10,000 EuroStreets digital road network (version 3.1 based on TeleAtlas MultiNet TM, year 2008). Land cover data were extracted from European Corine Land Cover 2006 data (ETC-LC) except for Greece for which Corine Land Cover 2000 was used (ETC-LC). The 100 m resolution Corine datasets, with an initial 44 land classes, were grouped into six main land cover groups. Elevation was extracted from the SRTM Digital Elevation Database version 4.1 which has a resolution of one arc second (approximately 90 m) and a vertical error <16 m (CGIAR-CSI). We additionally obtained 1x1 km population data for 2011 from Eurostat (European Commission (Eurostat).

Both road and land cover databases were intersected with a 100x100 m base polygon and the sum of road length (for 'all' and 'major' roads) and sum of land cover area (for the six grouped land classes) were calculated. The 100x100 m polygons were converted to grids and a focalsum procedure was applied to calculate these predictor variables for different distances, i.e. "buffers". All potential predictor variables are listed in Table S2, and GIS analysis was conducted in ESRI ArcGIS 10.5.

2.3 Model development and evaluation

A two-stage statistical procedure was applied to explain the spatial variation in the measurement data. Firstly, separate standard LUR models were developed based on all measurements for each pollutant. LUR models were developed according to the ESCAPE protocol; i.e. supervised stepwise linear regression as used in our previous paper (de Hoogh et al. 2016). Predictor variables were only allowed to enter the model if they adhered to the predefined direction of effect (see Table S2). We allowed significant predictor variables to enter the model when they added to the adjusted R^2 of the previous model step. Secondly, using the urban and rural background sites only, we explored the remaining broad scale variation in the residuals. Ordinary kriging was applied to the residuals using the GSTAT R package (LUR + kriging). If kriging was not successful (i.e. we could not fit a kriging function through the residuals) we offered longitude and/or latitude to the LUR model as additional predictors.

For each pollutant, six LUR models for 2010 were developed. The main model was developed using all sites (FULL). To test the robustness and stability of this model we additionally developed five hold out validation (HOV) models (HOV1, HOV2,…, HOV5), each built on 80% of the monitoring sites with the remaining 20% used for validation. Sites were selected into five groups (20% of sites) at random, stratified by site type and country.

HOV was performed after the LUR modelling and after the kriging (when applicable) using the criteria R^2 and root mean square error (RMSE). The main model (FULL, developed on all available sites) was evaluated against the 5 HOV samples.

For $PM_{2.5}$ and NO_2 we were able to perform an additional independent comparison with the ESCAPE monitoring datasets. Comparisons were performed at different scales: 1) overall (all ESCAPE sites); 2) overall ELAPSE (ESCAPE sites falling in ELAPSE study areas); and 3) matched to individual ELAPSE study areas (both detailed individual-level and administrative cohorts). Since the BC model was developed using the ESCAPE measurements, no independent comparison was possible.

2.4 Stability of spatial structure

In back extrapolation we assume that the spatial structure remains the same going back in time. To investigate the stability of the spatial structure of the models, and to test this assumption, we developed models for NO_2 and O_3 (2000 and 2005) using the same methods described in section 2.3. For $PM_{2.5}$ it was not possible to develop models for 2000 and 2005 due to the lack of monitoring data (12 and 165 in 2000 and 2005 respectively), instead we developed a model for 2013 (number of included monitoring sites = 732). The FULL models were mapped at a 100x100 m resolution across the study area and for the different years we visually inspected the spatial patterns.

As we did not have access to cohort geocodes, we created a random point file of 150,000 points across the full rectangular extent of the study area. After intersecting with the study area boundary, approximately 44,000 points remained which was considered a sufficient number to evaluate the stability. These points were intersected with all the raster surfaces: 2010 for $PM_{2.5}$, NO_2 and O_3 (annual, cold season and warm season); 2013 for $PM_{2.5}$; and 2005 and 2000 for NO_2 and O_3. Comparisons of model predictions were made for the West-European countries combined and at the national scale reporting R^2, RMSE and fractional bias (FB). In addition we calculated population weighted annual means for $PM_{2.5}$, NO_2 and O_3, using the 1x1 km GEOSTAT population database (European Commission (Eurostat).

We additionally evaluated the correlation of annual average measurements (plus summer and winter average for O_3) for those AirBase stations with measurements going sufficiently back in time.

2.5 Population exposure

For 2010, we calculated the total population of West-European countries (based on the GEOSTAT 2011 population grid dataset (European Commission (Eurostat)) residing in $PM_{2.5}$ and NO_2 concentration classes.

3. Results

3.1 Air pollution models 2010

The performance statistics (squared Pearson correlation (R^2) and RSME) and model structure of the FULL hybrid models for all pollutants are presented in Table 1 including the LUR component and, where applicable, the combined LUR + kriging component. The variograms of the kriging models for $PM_{2.5}$, O_3 in the warm and cold season are shown in Figure S2. A detailed model description, including constants, coefficients, incremental R^2 and RMSE can be found in Table 2 for $PM_{2.5}$ and the Supplementary material for the other pollutants (Table S3) and years (Table S4). Figure 1 shows the mapped surfaces at a 100x100 m resolution of the FULL models for all pollutants.

<INSERT Table 1 around here>

<INSERT Table 2 around here>

3.1.1 $PM_{2.5}$ models

The $PM_{2.5}$ LUR model developed on all available monitoring sites (FULL) explained 62% of spatial variation of the measured $PM_{2.5}$ concentrations (Table 1). Apart from satellite and CTM estimates, the LUR model included altitude, all roads, natural areas, ports and residential area. The satellite variable was the strongest predictor in all models explaining approximately 48% of the spatial variation in measured $PM_{2.5}$ concentrations. Comparing the predicted increase in $PM_{2.5}$ across a change from the 1st to the 99th percentile of each predictor, satellite and CTM $PM_{2.5}$ were associated with the largest contrast in $PM_{2.5}$. The model included large scale predictors (CTM, SAT at 10x10 km) and small-scale road, natural and residential land (50-200m) predictors. Kriging increased the explained variation to 72%.

The difference between the calibration and HOV R^2 of the FULL $PM_{2.5}$ model was small (72% vs 66%) confirming that overfitting was unlikely to be a big problem in the model development (Table 2). Similar predictor variables as in the FULL model were retained in the validation models, with only ports and urban green not always present in each model. Consistently, predictions of the six models (FULL and 5 HOV) at the 44,000 randomly selected sites were very highly correlated documenting the robustness of the model (Figure S3).

The mapped FULL $PM_{2.5}$ model (see Figure 1) showed predicted levels of $PM_{2.5}$ > 20 µg/m^3 in major cities and the Po area (the Po river basin running from the Western Alps to the Adriatic Sea) in Italy. Large parts of Northern Europe had low (<10 µg/m^3) predicted $PM_{2.5}$ concentrations.

<INSERT Figure 1 around here>

We tested the three different $PM_{2.5}$ satellite products in preliminary $PM_{2.5}$ model development and found that the 0.1° x 0.1°inferred 2009-2011 product v3.01 produced the best results (see the Supplementary material section 1 and Table S5 for a more detailed description).

3.1.2 NO_2 models

The FULL NO_2 model explained 59% of the spatial variation (Table 1 and Table S3). In all models the CTM variable was the strongest predictor explaining approximately 29% of variation in NO_2 concentrations, followed by the small (100-300m) and larger scale (2000m) road variables. All roads, major roads, natural and residential predictor variables consistently appeared in every model. Predictions of the six models (FULL and 5 HOV) models at the 44,000 randomly selected sites were very highly correlated (Figure S3). None of the variogram models adequately fit the residuals at the NO_2 background monitoring sites, nor did including longitude and/or latitude help explain the residuals (p-value of coefficient not significant). The mapped NO_2 estimates (Figure 1) showed more variation compared to

$PM_{2.5}$. Major roads and cities clearly stood out with predicted concentrations generally > 30 µg/m^3. Away from sources in rural areas, NO_2 levels dropped below 15 µg/m^3.

3.1.3 O_3 models

Around half of the spatial variation in the annual O_3 measurements was explained by the CTM (MACC-O_3) variable. Other variables consistently entering all 6 annual models were roads, residential land cover and altitude (Table S3). Ports entered the FULL model and 4 of the 5 HOV models. The CTM was associated with much larger contrast in O_3 than the other predictors. Predictions of the 6 models (FULL and 5 HOV) models at the 44,000 randomly selected sites were very highly correlated (Figure S3). No reliable kriging function could be fit through the residuals of O_3 background monitoring sites. However, latitude and longitude variables were fit to the models. The FULL model had a R^2 of 65% (HOV models ranging from 63 to 68%).

Like the annual O_3 model, the cold season O_3 model was dominated by the MACC predictor variable, explaining nearly 60% of the spatial variation in measured O_3 concentrations. Roads, residential land and altitude variable entered in all 6 cold season models. Kriging explained, on average, an additional 16% of the spatial variation, bringing the final performance of the FULL O_3 cold model to 83% (80% to 85% for the 5 validation models).

The O_3 warm season models also contained a CTM variable, but unlike the annual and cold season O_3 models where the annual MACC CTM variable entered, here the warm season DEHM CTM variable was the stronger predictor. Other variables entering in all models were roads, ports, residential land and altitude. The performance of LUR models was moderate (R^2 ranging from 44 to 48%) but with additionally fitted kriging functions, we increased the explained variation to 70% for the FULL model (67% to 73% for the 5 validation models).

Maps of the FULL O_3 models (Figure 1 and S4) showed similar general patterns for annual and cold season, with the highest predicted O_3 concentrations in Southern Europe and lower concentrations in more central areas (England, the Netherlands, Germany and northern Italy). Areas of high altitude also tended to have higher predicted O_3 levels compared areas of lower altitudes. Predicted O_3 concentrations for the warm season showed a somewhat different spatial pattern with a much clearer negative North-South gradient than the cold season model.

3.1.4 BC models

For the FULL BC LUR model we achieved an explained variation of 54% (FULL model) and between 52 and 57% for the 5 HOV models (Table 1, Table S3). For all 6 models, the CTM MACC-$PM_{2.5}$ contributed 24 to 30% of the explained spatial variation. Roads, $PM_{2.5}$ SAT estimates, urban green land, residential land and natural land were also included consistently in FULL and HOV models. Predictions of the 6 (FULL and 5 HOV) models at the 44,000 randomly selected sites were very highly correlated (Figure S3). The BC model included large contributions from large-scale predictors (CTM $PM_{2.5}$, Y-coordinate and residential density) and small-scale predictors (roads and residential density).

Due to the clustered nature of the BC monitoring data it was not possible to perform kriging. Latitude was best able to explain the residuals.

When mapped across Western Europe (Figure 1), BC predicted concentrations showed a distinct North – South division, with low (<=0.8 $10^{-5}m^{-1}$) BC concentrations in Scandinavia and the north of the UK, and higher >0.8 $10^{-5}m^{-1}$ in the rest of Western Europe. Mediterranean Europe had the highest concentration > 1.2 $10^{-5}m^{-1}$. Traffic sources were also clearly identifiable in the inset with major roads visible around Paris.

3.2 Comparison at ESCAPE sites

We performed an independent external comparison for $PM_{2.5}$ and NO_2 FULL models using measured concentration data from the ESCAPE study. Table 3 shows the correlations at different scales including the mean and standard deviation of measured concentrations at the ESCAPE measurement sites.

<INSERT Table 3 around here>

The $PM_{2.5}$ FULL model explained 65% of variance overall (n=416) with a small fractional bias (FB = -2%). The explained variance is almost identical to the HOV R^2 of 66% (Table 1). Restricting the analysis to the overall area with ELAPSE cohorts (n = 255) led to a slight decrease in the explained variance (59%) and a small overestimation (FB = -10%). The comparison at each ELAPSE study areas separately (detailed individual-level and administrative cohorts) revealed a large range in the explained variation, 8% for EPIC Oxford and English administrative cohort to 66% for HNR, also with the FB varying from -2 to -30%. We note that the number of sites is relatively small for the individual area comparisons.

NO_2 FULL models also showed reasonable associations for overall (49%) and overall ELAPSE (46%). The explained variance was modestly lower than the HOV R^2 of 57% (Table 1). FB indicated a small overestimation of 13% for the ELAPSE overall area. At the ELAPSE detailed individual-level cohorts the correlations for NO_2 were generally better than for $PM_{2.5}$: all were >47% except for HUBRO (7%) and EPIC VARESE (34%). FB showed overestimation for all areas, except for ELAPSE areas in Italy.

3.3 Air pollution models for different time periods and stability analysis

3.3.1 Models for 2000, 2005 (NO_2 and O_3) and 2013 ($PM_{2.5}$)

The performance statistics of the $PM_{2.5}$, NO_2 and O_3 models for different years are presented in Table S4. The 2013 $PM_{2.5}$ LUR models explained 64% of spatial variation in the $PM_{2.5}$ measurements. The LUR models had some similarities with the 2010 models, with MACC, SAT, roads and natural land entering all models. Neither reliable kriging models nor longitude/latitude variables improved the models.

No NO_2 MACC CTM estimates were available for the years 2000 and 2005, so only DEHM NO_2 for 2000 and 2005 estimates were offered to the NO_2 model development. Otherwise the NO_2 models showed a similar structure with the 2010 NO_2 LUR models (CTM, roads, natural land, residential land and ports in all models), but performed slightly less well (R^2 NO_2 2000 = 56%; R^2 NO_2 2005 = 52%).

O_3 models for 2000 and 2005 were able to respectively explain 60% and 49% (annual), 82 and 42% (warm season), 52 and 70% (cold season) of the variation in measured concentrations. The 2000 and 2005 annual and warm O_3 models contained DEHM CTM variables whereas no DEHM variable entered the cold season models. Kriging models explained an additional ~ 25% of spatial variation in the 2000 warm season and the 2005 cold season models. Latitude and longitude variables were entered to the other models.

Figure 1 shows the maps of $PM_{2.5}$ (2013, 2010), NO_2 and O_3 warm season (2010, 2005, 2000). Similar patterns over multiple years were observed with, for example, high predicted $PM_{2.5}$ concentrations for both 2010 and 2013 in the Po valley in North Italy and low $PM_{2.5}$ concentrations in Scandinavia. Spatial patterns in the NO_2 and O_3 concentrations maps for the 3 years also appeared broadly similar.

<INSERT Table 4 around here>

3.3.2 Comparison of model predictions for Western Europe across years

Table 4 (and Figure S5) shows the results of the stability tests at country level. Agreement in spatial variation was generally high at the overall EU country and combined ELAPSE country level (>76%) for all comparisons, except for the O_3 cold season surface (44% when 2000 model compared to 2010). At the national level, focusing on ELAPSE countries only, we observed some heterogeneity in the associations. Both 2000 and 2005 NO_2 surfaces showed a high agreement with the 2010 NO_2 surface (all ELAPSE countries >80%). The agreement between $PM_{2.5}$ surfaces developed for 2010 and 2013 showed more variability, with four ELAPSE countries >80% (UK, Sweden, Belgium and Italy), the Netherlands 70% and the rest between 48 and 60%. There was a high variability between the associations of the different O_3 surfaces. The agreement between O_3 annual surfaces of 2000 and 2005 with

2010 was reasonable, all ELAPSE countries had >60% explained spatial variability, with the exception of Sweden (2000) with 45%. Except for the 2005 O_3 cold (all ELAPSE countries > 60%), the O_3 cold and warm season surfaces were less stable over time with large ranges of explained spatial variability. Italy performed poorly with 1.6%, 11.9% and 16.6% for respectively 2000 warm season, 2005 warm season and 2000 cold season (combined with the largest RMSE's).

NUTS areas are standard administrative divisions of EU countries for statistical purposes. We performed the stability analysis using the same 44,000 random points at the NUTS1 area level (see Figure S6) to gain a better understanding of the stability at the sub-national level. Similar to the national level, there was a good agreement for all areas for NO_2 2000 and 2005 when compared to the 2010 surface (R^2 >0.60). For more details see the Supplementary material section 2.

3.3.3 Comparison of measurements

We additionally evaluated the relationship between measured average concentrations for those AirBase stations with measurements going sufficiently back in time between 2010 to 2005 and 2000 (Table 5). In Western Europe the measured concentrations between the different years yielded high correlations. When focusing on ELAPSE participating countries, high correlations were also observed for the majority of the countries and years.

<*INSERT Table 5 around here*>

3.4 Population exposure

Based on our modelled concentrations (FULL models), a respective 8 million (2%) and 371 million (89%) people live in areas with estimated $PM_{2.5}$ concentrations greater than the EU annual $PM_{2.5}$ limit value of 25 µg/m^3 and the WHO annual guideline of 10 µg/m^3. 32 million (8%) of people live in areas with modelled NO_2 concentration greater than the EU and WHO annual NO_2 guideline of 40 µg/m^3 (see Table S6). Table S7 shows that population weighted concentrations levels across the whole of our study area do not drastically fluctuate over time and are generally low ($PM_{2.5}$ ~ 11 µg/m^3 and NO_2 < 20 µg/m^3).

4. Discussion

We developed West-European LUR models at a 100x100 m spatial scale for four priority pollutants. The models including large scale satellite data and CTM and small-scale traffic and land use predictors explained between 54% (BC) and 83% (O_3 cold season) of the measured variability in concentrations. The explained variance at fully independent sites was only slightly less than the internal hold-out validation: 65% vs 66% for $PM_{2.5}$ and 49% vs 57% for NO_2. Predictions from the 2010 model correlated highly with models developed for 2000 and 2005 (2013 for $PM_{2.5}$) at the overall European scale, with squared correlations larger than 76%, except for the O_3 cold season of 2000 (44%). The temporal correlation was more variable when evaluated at the country and especially at the NUTS1 level. Correlations between measured concentrations at the EU level between 2010 - 2005 and 2010 - 2000 for NO_2 and O_3 (R^2 between 68% to 87%) and for $PM_{2.5}$ 2010 - 2013 (R^2 79%) were even higher than modeled concentrations. Based on our modelled surfaces, 371 million and 32 million people in Western Europe live in areas with air pollution levels exceeding the WHO annual guidelines for $PM_{2.5}$ and NO_2 respectively.

4.1 Interpretation of 2010 models

$PM_{2.5}$ SAT and CTM available at a 10x10 km scale were the strongest predictors in the $PM_{2.5}$ models, consistent with $PM_{2.5}$ being a largely regionally varying pollutant. Eeftens et al. (2012a) reported that 81% of the variability in the ESCAPE annual average $PM_{2.5}$ concentrations was due to between study area contrast. The modest contrast related to the small-scale road variable is consistent with the overall mean ratio of 1.14 comparing traffic and background sites within ESCAPE (Eeftens et al. 2012a). Roads, ports and residential areas represent the contribution of local sources, with altitude, and nature/urban green representing pollution sinks. Applying kriging to the residuals of the LUR model explained an extra 10% of the variation, suggesting that the SAT and CTM predictors did not fully capture the large scale variation of $PM_{2.5}$ across Europe. Alternatively, the number of sites was insufficient to train the model. Kriging was not feasible for the 2013 model, possibly due to the larger number of sites.

In the BC models, satellite and CTM $PM_{2.5}$ also contributed strongly, raising potential concerns when applying the $PM_{2.5}$ and BC models in the epidemiological analysis as it might be difficult to tease apart their respective contribution to health effects. Compared to the $PM_{2.5}$ models, small-scale road predictors contributed more to the BC prediction. The FULL model contained three road variables with a similar magnitude to the CTM and SAT predictors. This is consistent with the observation in ESCAPE that 52% of the variability was due to within-study area variability (Eeftens et al. 2012a). The overall ratio of BC concentrations measured at traffic /urban background sites was 1.38 (Eeftens et al. 2012a). The residuals of our initial model showed a clear north-south gradient, which was captured by a Y-coordinate in the model, documenting that the models did not predict the large scale contrast of BC across Europe sufficiently. MACC and satellites do not represent BC, whereas DEHM modelled BC at a larger scale (50x50 km scale). It is likely that limitations in emission data for BC may have impacted the performance of the models.

After the CTM predictor variable, small-scale road variables were the strongest predictors in the NO_2 models. Motorized traffic is a dominant source of local NO_2 concentrations, as illustrated by the overall ratio of 1.63 for concentrations measured at traffic vs. urban background ESCAPE monitoring sites (Cyrys et al. 2012). In ESCAPE, 60% of the variability of NO_2 was due to within-study area variability (Cyrys et al. 2012). The NO_2 models could not be further improved by kriging or geographical coordinates, suggesting that the CTM adequately captured the large scale variation across Europe. We previously suggested that CTM's were better developed for NO_2 than for $PM_{2.5}$ when discussing the contribution of CTM and SAT to $PM_{2.5}$ and NO_2 LUR models (de Hoogh et al. 2016).

In O_3 models, CTM (the ensemble MACC for the annual and cold period and DEHM for the warm season) were the dominant predictor variables, consistent with O_3 being a regional pollutant. The model further predicted higher concentrations at higher altitude, in accordance with a previous European LUR model (Beelen et al. 2009). Predicted lower concentrations near roads was consistent with scavenging of O_3 by NO_2. In both the warm and cold season, kriging substantially improved the models, likely illustrating limitations in the CTM. Kriging did not contribute to the annual model, possibly because the annual average combined the two different spatial patterns of the cold and warm seasons.

Few studies have combined LUR and kriging in air pollution models. Young et al. (2016) evaluated the additional value of satellite data and/or kriging on NO_2 LUR models across the USA for 1990 – 2012. Models with both satellite data and kriging performed best, increasing the average cross-validation R^2 from 0.72 (just applying LUR) to 0.85. Satellite or kriging alone yielded respective average R^2's of 0.81 and 0.84. Although we found improvement of model performance with kriging for the $PM_{2.5}$ and O_3 models, we did not see the same result in our NO_2 models. This might be due to the difference in scale of the two studies. Young et al. (2016) estimated NO_2 concentrations at a 25 x 25 km resolution, thereby not explaining intra-urban variation but rather focusing on more regional background. This study operates at a much smaller resolution (100x100 m) and, at least for NO_2, the residual concentrations after LUR were too variable, even at background sites, for reliable kriging functions. In a previous study distinguishing global, regional and urban scales, universal kriging improved PM10, O_3 and NO_2 European models compared to regression models (Beelen et al. 2009). In that study, the analysis was based on 1 * 1 km estimates.

Relatively few studies have tested the robustness by developing HOV models and assessing the structure of the models. Johnson et al. (2010) evaluated $PM_{2.5}$, NO_x and benzene LUR models in New Haven, CT, USA by including hold-out validation using varying sizes of training/testing groups. van Nunen et al. (2017) performed a 10-fold cross validation when developing UFP LUR models in six study European areas. We observed that the model predictions from our FULL model correlated very highly with the 5 HOV models at the 44,000 independent sites, suggesting that the developed models were robust. The correlations in our study were higher than that observed for the UFP models based on short-term monitoring at 160 sites in some of the cities (van Nunen et al. 2017).

4.2 Comparison with other European models

Previously we published the development of hybrid $PM_{2.5}$ and NO_2 LUR models for the same study area, showing that satellite-derived (SAT) estimates and CTM estimates contribute considerably to the explained variance in $PM_{2.5}$ and NO_2 measurements (de Hoogh et al. 2016). The models presented in this paper confirm our previous findings. Moreover, by additionally including kriging to explain residuals at background monitoring sites, we improved the $PM_{2.5}$ hybrid models from 62 to 72% (R^2). This improvement was also observed when tested using the independent ESCAPE monitoring dataset, showing an improvement from 53 to 65% (R^2). For NO_2 models, where the inclusion of longitude explained some of the residuals, the R^2 remained the same (both 58%); but the improved NO_2 model described here yielded a higher independent validation (R^2) of 49% compared to 43% in de de Hoogh et al. (2016). Additionally we evaluated the performance of SAT and CTM derived estimates by comparing monitored AIRBASE data and satellite derived $PM_{2.5}$ (R^2 = 0.48) and NO_2 (R^2 = 0.13) and MACC $PM_{2.5}$ (R^2 = 0.41) and NO_2 (R^2 = 0.29). SAT and CTM (MACC) surfaces explain less of the measured spatial variation than when these datasets are used within a hybrid LUR framework as presented as in this paper.

Vienneau et al. (2013) also developed European NO_2 and PM_{10} LUR models, for 2005-2007, showing that the inclusion of satellite data substantially improved model performance. The NO_2 model explained a comparable fraction of the variation (46-56%) to our models. The CTM predictor outperformed the satellite data in our NO_2 model, a predictor variable not available in the study by Vienneau et al. (2013).

To date few studies have attempted to model pollutants other than NO_2 and PM. European O_3 LUR models have been previously developed by Beelen et al. (2009) for the year 2001 at the global (R^2 = 0.53), rural (R^2 = 0.63) and urban (R^2 = 0.06) scale. Our annual O_3 model performance for 2000 yielded a higher R^2 (0.63) possibly due to the inclusion of DEHM estimates in our model. In addition we further developed seasonal O_3 models.

4.3 Application of 2010 models in epidemiological studies

The models developed and described here will be used for the exposure assessment in ELAPSE for 7 administrative cohorts and a pooled cohort comprising of 11 local cohorts across 11 countries in Europe (Norway, Sweden, Denmark, United Kingdom, the

Netherlands, Belgium, Germany, France, Switzerland, Austria and Italy). For the pooled cohort, the (moderately) high explained variance in hold-out validation and external validation over the full area suggests that exposure assessment is robust. For individual cohorts, comparison with ESCAPE data in the respective study areas showed more variable results, especially for $PM_{2.5}$. This implies that our West European model should be applied with caution in a small area (part of a country) unless local validation is possible. The difference between NO_2 and $PM_{2.5}$ could be due to the relatively small number of sites for $PM_{2.5}$ and the smaller contrast in $PM_{2.5}$ within cohorts compared to NO_2.

For the administrative cohorts, direct comparisons of the Dutch, Rome and to some extent national English and Swiss (NO_2 only) study areas with the ESCAPE data are possible due to overlaps between the ESCAPE and ELAPSE study areas/regions. The West European ELAPSE models explained variation well, except for $PM_{2.5}$ in the Netherlands (possibly due to small variation) and NO_2 in Switzerland. The findings for Switzerland do not directly apply to the Swiss cohort, as the evaluation was limited to three cities whereas the Swiss cohort includes the entire population including those in rural and Alpine areas. We have no ready explanation for these findings, and can only speculate that a more locally generated model may better capture area-specific small-scale concentration differences than a pan-European model, which tends to smooth intra-urban differences over several very different study areas.

4.4 Spatial stability of models and measurements over time

This is one of the few studies which has tested the stability of spatial structure of air pollution exposure models at a continental scale, by developing models for different time points and comparing the respective estimates. Most studies evaluated LUR models at a national or sub-national scale by linear regression using historical monitoring data, allowing the constant and coefficient to change (Cesaroni et al. 2012; Chen et al. 2010; Eeftens et al. 2011; Gulliver et al. 2013; Gulliver and de Hoogh 2015; Levy et al. 2015). Gulliver et al. (2016), however, produced separate NO_2 LUR models for 1991 and 2009 for the UK and found that the year-specific 1991 model yielded similar exposures as the back-extrapolated 2009 model. R Wang et al. (2013) developed NO_2 LUR models for 2003 and 2010 for Vancouver, Canada, and when applied to measurements of the other year were able to explain 52 to 61% (2003 model to 2010 measurements) and 44 to 49% (2010 model to 2003 measurements) of the spatial variation. These studies suggest that the spatial structure of the different models were similar, at least at a national or city level. It is difficult to compare the findings of the analyses carried out in this study with the studies conducted at the sub-continental scale. In this study we specifically assessed the stability of the spatial structure by comparing the concentration surfaces of the different models based on a set of ~44,000 random points spread across the study area. At the EU scale (all countries combined and

ELAPSE countries combined) there was a high squared correlation (>76%) between the other year models ($PM_{2.5}$ 2013, NO_2 and O_3 2000, 2005) and the corresponding 2010 models, with the only exception the O_3 2000 cold season model (~45%). Other countries that performed poorly for O_3 2000 cold were Germany and the Netherlands. The poorer temporal correlation for O_3 may be due to the smaller spatial contrast when evaluating at a smaller spatial scale. Another explanation may be that there are different CTM predictions used in the LUR models for 2010 (MACC-O_3 for annual and cold O_3) compared to 2000 and 2005 for which only the DEHM model was available.

Correlations between annual average measured concentrations at sites that were in operation for an extended time period were even higher. The higher correlation for measurements was probably due to the only moderately high explained variance of the models and difference in availability of predictor variables across years. A difficulty in the interpretation of monitoring data is the limited number of sites with continuous data, especially for $PM_{2.5}$.

The temporal stability of the estimated spatial surface for most of the pollutants has positive consequences for further application in long-term epidemiological studies especially those including cohorts which started one or two decades ago and which will have had several follow-ups since then. The 2010 surfaces produced here can be used with some confidence as the base for back-extrapolation.

For several areas we now have study-area specific ESCAPE models and Europe wide ELAPSE models. The ESCAPE models are based upon a smaller number of training sites but may be more specific for the area. The spatial extent of ESCAPE PM models has limited the analysis of some ESCAPE cohorts (e.g. only Paris in the national French E3N cohort and Copenhagen in the Danish DCH cohort). The ELAPSE model can be applied to larger areas e.g. entire France, Denmark. In general, Europe wide models may be better when large areas are studied. In international multi-center studies, the use of a single harmonized model is important to standardize exposure assessment. We do not recommend the use of our ELAPSE models in single cohort analyses e.g. in a cohort exclusively based in Stockholm, unless local validation data documents that the European model can explain small-scale variation in the specific city

5. Conclusions

We were able to develop robust $PM_{2.5}$, NO_2, BC and O_3 LUR models. At the West-European scale models were robust in time, becoming less robust at smaller spatial extents. In terms of model performance we improved on previously published European NO_2 and $PM_{2.5}$ models and developed new models for BC and O_3 explaining large fractions of the variance.

We showed, by five-fold hold-out validation plus an independent comparison, that the models were spatially robust at the West-European and, to a lesser degree, at the national scale. At the West-European scale, $PM_{2.5}$, NO_2 and O_3 models were robust in time. For BC models we were not able to perform a stability analysis. At smaller spatial scales, models were less robust in time, especially for O_3. The models presented here will be used to assign exposures in the ELAPSE study and will be made available for other studies in Europe.

Acknowledgements / Disclaimer

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. We would like to acknowledge Jesper H. Christensen and Jørgen Brandt for providing the DEHM calculations.

References

Basagaña X, Rivera M, Aguilera I, Agis D, Bouso L, Elosua R, et al. 2012. Effect of the number of measurement sites on land use regression models in estimating local air pollution. Atmos Environ 54.

Bechle MJ, Millet DB, Marshall JD. 2013. Remote sensing of exposure to no 2: Satellite versus ground-based measurement in a large urban area. Atmospheric Environment 69:345-353.

Bechle MJ, Millet DB, Marshall JD. 2015. National spatiotemporal exposure surface for no2: Monthly scaling of a satellite-derived land-use regression, 2000–2010. Environmental science & technology 49:12297-12305.

Beelen R, Hoek G, Pebesma E, Vienneau D, de Hoogh K, Briggs DJ. 2009. Mapping of background air pollution at a fine spatial scale across the european union. SCIENCE OF THE TOTAL ENVIRONMENT 407:1852-1867.

Beelen R, Hoek G, Vienneau D, Eeftens M, Dimakopoulou K, Pedeli X, et al. 2013. Development of no2 and nox land use regression models for estimating air pollution exposure in 36 study areas in europe - the escape project. ATMOSPHERIC ENVIRONMENT 72:10-23.

Beelen R, Raaschou-Nielsen O, Stafoggia M, Andersen ZJ, Weinmayr G, Hoffmann B, 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. The Lancet 383:785-795.

Beelen R, Hoek G, Raaschou-Nielsen O, Stafoggia M, Andersen ZJ, Weinmayr G, et al. 2015. Natural-cause mortality and long-term exposure to particle components: An analysis of 19 european cohorts within the multi-center escape project. ENVIRONMENTAL HEALTH PERSPECTIVES 123:525-533.

Brandt J, Silver JD, Frohn LM, Geels C, Gross A, Hansen AB, et al. 2012. An integrated model study for europe and north america using the danish eulerian hemispheric model with focus on intercontinental transport of air pollution. Atmospheric Environment 53:156-176.

Cesaroni G, Porta D, Badaloni C, Stafoggia M, Eeftens M, Meliefste K, et al. 2012. Nitrogen dioxide levels estimated from land use regression models several years apart and association with mortality in a large cohort study. Environmental Health 11:48.

CGIAR-CSI. Srtm 90m digital elevation data.

Chen H, Goldberg MS, Crouse DL, Burnett RT, Jerrett M, Villeneuve PJ, et al. 2010. Back-extrapolation of estimates of exposure from current land-use regression models. Atmospheric Environment 44:4346-4354.

Chen H, Kwong JC, Copes R, Hystad P, van Donkelaar A, Tu K, et al. 2017. Exposure to ambient air pollution and the incidence of dementia: A population-based cohort study. Environ Int 108:271-277.

Cohen AJ, Brauer M, Burnett R, Anderson HR, Frostad J, Estep K, et al. 2017. Estimates and 25-year trends of the global burden of disease attributable to ambient air pollution: An analysis of data from the global burden of diseases study 2015. The Lancet 389:1907-1918.

Crouse DL, Peters PA, Hystad P, Brook JR, van Donkelaar A, Martin RV, et al. 2015. Ambient pm2.5, o(3), and no(2) exposures and associations with mortality over 16 years of follow-up in the canadian census health and environment cohort (canchec). Environ Health Perspect 123:1180-1186.

Cyrys J, Eeftens M, Heinrich J, Ampe C, Armengaud A, Beelen R, et al. 2012. Variation of no2 and nox concentrations between and within 36 european study areas: Results from the escape study. Atmos Environ 62.

de Hoogh K, Gulliver J, Donkelaar Av, Martin RV, Marshall JD, Bechle MJ, et al. 2016. Development of west-european pm2.5 and no2 land use regression models incorporating satellite-derived and chemical transport modelling data. Environmental Research 151:1-10.

Di Q, Wang Y, Zanobetti A, Wang Y, Koutrakis P, Choirat C, et al. 2017. Air pollution and mortality in the medicare population. New England Journal of Medicine 376:2513-2522.

EEA. 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 13 January 2015].

EEA. 2017. Air quality in europe — 2017 report.

Eeftens M, Beelen R, Fischer P, Brunekreef B, Meliefste K, Hoek G. 2011. Stability of measured and modelled spatial contrasts in no₂ over time. Occupational and Environmental Medicine.

Eeftens M, Beelen R, de Hoogh K, Bellander T, Cesaroni G, Cirach M, et al. 2012a. 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. ENVIRONMENTAL SCIENCE & TECHNOLOGY 46:11195-11205.

Eeftens M, Tsai M-Y, Ampe C, Anwander B, Beelen R, Bellander T, et al. 2012b. 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.

ETC-LC. Corine land cover (clc2006), raster database (version 12/2013). .

ETC-LC. Corine land cover (clc2000), raster database (version 12/2009). .

European Commission (Eurostat JRCaDRP-R-G. Geostat 2011 grid dataset.

Gulliver J, de Hoogh K, Hansell A, Vienneau D. 2013. Development and back-extrapolation of no2 land use regression models for historic exposure assessment in great britain. ENVIRONMENTAL SCIENCE & TECHNOLOGY 47:7804-7811.

Gulliver J, de Hoogh K. 2015. Environmental exposure assessment: Modelling air pollution concentrations. In: Oxford textbook of global public health Part 6 (Detels R, Gulliford M, Abdool Karim Q, Chuan C, eds). Oxford:Oxford University Press.

Gulliver J, de Hoogh K, Hoek G, Vienneau D, Fecht D, Hansell A. 2016. Back-extrapolated and year-specific no2 land use regression models for great britain - do they yield different exposure assessment? Environment International 92–93:202-209.

Inness A, Baier F, Benedetti A, Bouarar I, Chabrillat S, Clark H, et al. 2013. The macc reanalysis: An 8 yr data set of atmospheric composition. Atmos Chem Phys 13:4073-4109.

Johnson M, Isakov V, Touma JS, Mukerjee S, Özkaynak H. 2010. Evaluation of land-use regression models used to predict air quality concentrations in an urban area. Atmos Environ 44.

Levy I, Levin N, Yuval, Schwartz JD, Kark JD. 2015. Back-extrapolating a land use regression model for estimating past exposures to traffic-related air pollution. Environmental Science & Technology 49:3603-3610.

Novotny EV, Bechle MJ, Millet DB, Marshall JD. 2011. National satellite-based land-use regression: No2 in the united states. Environmental science & technology 45:4407-4414.

Pinault L, van Donkelaar A, Martin RV. 2017. Exposure to fine particulate matter air pollution in canada. Health Reports 28:9-16.

Van Donkelaar A, Martin RV, Brauer M, Boys BL. 2015. Use of satellite observations for long-term exposure assessment of global concentrations of fine particulate matter. Environmental health perspectives 123:135.

van Donkelaar A, Martin RV, Brauer M, Hsu NC, Kahn RA, Levy RC, et al. 2016. Global estimates of fine particulate matter using a combined geophysical-statistical method with information from satellites, models, and monitors. Environmental Science & Technology 50:3762-3772.

van Nunen E, Vermeulen R, Tsai M-Y, Probst-Hensch N, Ineichen A, Davey M, et al. 2017. Land use regression models for ultrafine particles in six european areas. Environmental Science & Technology 51:3336-3345.

Vienneau D, de Hoogh K, Bechle MJ, Beelen R, van Donkelaar A, Martin RV, et al. 2013. Western european land use regression incorporating satellite- and ground-based measurements of no2 and pm10. ENVIRONMENTAL SCIENCE & TECHNOLOGY 47:13555-13564.

Wang M, Beelen R, Basagana X, Becker T, Cesaroni G, de Hoogh K, et al. 2013. Evaluation of land use regression models for no2 and particulate matter in 20 european study areas: The escape project. ENVIRONMENTAL SCIENCE & TECHNOLOGY 47:4357-4364.

Wang R, Henderson SB, Sbihi H, Allen RW, Brauer M. 2013. Temporal stability of land use regression models for traffic-related air pollution. Atmospheric Environment 64:312-319.

WHO. 2016. Ambient air pollution: A global assessment of exposure and burden of disease.

Young MT, Bechle MJ, Sampson PD, Szpiro AA, Marshall JD, Sheppard L, et al. 2016. Satellite-based no2 and model validation in a national prediction model based on universal kriging and land-use regression. Environmental Science & Technology 50:3686-3694.

Table 1. Model structure[a] and performance of 2010 LUR models

Pollutant	Stage	Method	N sites	R^2	RMSE[b]	Full LUR model[c]
$PM_{2.5}$	Training	LUR	543	62.2	3.17	3.19 +13.24*SAT-PM25 +7.08*MACC-PM25 – 3.82* ALT + 2.17*$ALRD_{100}$ -2.07*NAT_{50} +2.39*POR_{800} +1.41*RES_{200}
		LUR +Kriging		72.2	2.71	
	HOV	LUR		58.7	3.30	
		LUR +Kriging		66.4	2.97	
BC[4]	Training	LUR	436	54.4	0.56	0.99 + 0.85* MACC-PM25 + 0.30* SAT-PM25 + 0.68*$MJRD_{100}$ +0.40* $ALRD_{50}$ + 0.45* $ALRD_{700}$ +0.90*RES_{3000} -0.12*UGR_{1000} – 1.16*Y
	HOV	LUR		51.4	0.58	
NO_2[d]	Training	LUR	2399	58.8	9.38	3.30 + 22.73*MACC-NO_2 + 7.04* $ALRD_{50}$ + 3.92* $ALRD_{300}$ + 12.32* $MJRD_{100}$ + 15.73*$ALRD_{2000}$ -3.38*NAT_{400} + 4.1*POR_{700} + 5.8*RES_{300}
	HOV	LUR		57.5	9.51	
O_3 annual[d]	Training	LUR	1747	65.1	6.73	40.54 +25.51*MACC-O_3 -2.49*$ALRD_{50}$ – 4.75* $ALRD_{200}$ – 3.24* $MJRD_{200}$ –1.57*POR_{4000} -1.94*RES_{500} -4.13*RES_{2000} +8.82*ALT + 2.48*X – 10.05*Y
	HOV	LUR		63.4	6.87	
O_3 warm	Training	LUR	1730	45.5	10.07	30.00 +32.57*DEHM-O_3 -6.87* $ALRD_{200}$ –6.03* $MJRD_{100}$ -5.95*$PORT_{5000}$ -4.79*RES_{2000} +5.70*ALT
		LUR +Kriging		69.6	7.51	
	HOV	LUR		44.5	10.15	
		LUR +Kriging		59.9	8.63	
O_3 cold	Training	LUR	1716	67.7	7.43	1.00 + 37.62*MACC-O_3 -3.35* $ALRD_{200}$ -3.48* $MJRD_{50}$ -1.61* $MJRD_{700}$ +5.81*NAT_{700} -4.18*RES_{1200} -1.10*TBU_{100}+2.21*UGR_{1000} +6.84*ALT
		LUR +Kriging		83.3	5.33	
	HOV	LUR		66.5	7.55	
		LUR +Kriging		75.3	6.99	

a. Regression slope in µg/m^3, except BC ($10^{-5}m^{-1}$), multiplied by the difference between the 1st and 99th percentile of each predictor to allow comparison across predictors
b. RMSE in µg/m^3, except BC ($10^{-5}m^{-1}$)
c. ALT = altitude, ALRD = all roads, MJRD = major roads, IND = industry, POR = ports, UGR = urban green, TBU = total build up, NAT = natural land, RES = residential, POP = sum of population, X = North-South trend, Y = East–West trend, SAT = satellite, MACC = MACC dispersion model, DEHM = DEHM CTM. Number in subscript depicts the buffer size (e.g. $ALRD_{100}$ = sum of all road length within 100 m)
d. No valid variograms were possible on the residuals of these models

Table 2. Structure and performance of LUR models[a] for $PM_{2.5}$ for full dataset and five hold-out validation datasets for 2010

Theme	Variable[b]	FULL[c]	HOV1	HOV2	HOV3	HOV4	HOV5
	(Constant)	3.19	3.46	3.53	3.14	3.49	3.32
Satellite	SAT-PM25	13.24	12.98	12.39	13.19	12.68	13.55
CTM	MACC-PM25	7.08	7.32	7.45	7.17	7.09	6.93
Altitude	ALT	-3.82	-3.82	-4.10	-3.93	-3.54	-3.73
Roads	$ALRD_{100}$	2.17	2.89		2.23	2.00	
	$MJRD_{50}$						1.98
	$MJRD_{100}$			2.26			
Urban green	UGR_{700}		-1.08				
	UGR_{800}					-0.98	
Nature	NAT_{50}	-2.07	-2.24			-2.72	-2.26
	NAT_{100}			-2.31	-2.12		
	NAT_{300}						
	NAT_{400}						
Ports	POR_{800}	2.39	3.19		2.95	2.46	2.35
Residential	RES_{50}		0.89				
	RES_{200}	1.41		1.72	1.44	1.48	
	RES_{300}						1.39
Training (LUR)	R^2	62.2	62.0	63.1	61.1	60.8	66.0
	RMSE	3.17	3.26	3.10	3.30	3.22	2.95
HOV (LUR)	R^2	58.7	62.2	53.9	67.4	68.1	50.3
	RMSE	3.30	2.93	3.67	2.68	3.01	3.94
Training (LUR + Kriging)	R^2	72.2	71.4	70.5	76.8	76.0	63.3
	RMSE	2.71	2.55	2.94	2.26	2.61	3.38
HOV (LUR + Kriging)	R^2	66.4	67.7	66.0	72.3	74.0	57.9
	RMSE	2.97	2.71	3.15	2.47	2.72	3.61

a. Regression slope $\mu g/m^3$ were multiplied by the difference between the 1st and 99th percentile of each predictor to allow comparison across predictors

b. ALT = altitude, ALRD = all roads, MJRD = major roads, IND = industry, POR = ports, UGR = urban green, TBU = total build up, NAT = natural land, RES = residential, POP = sum of population, X = North-South trend, Y = East-West trend, SAT = satellite, MACC = MACC dispersion model, DEHM = DEHM CTM. Number in subscript depicts the buffer size (e.g. $ALRD_{100}$ = sum of all road length within 100 m)

c. FULL refers to all sites; HOV1 is first holdout validation dataset (80% stratified random sample)

Table 3. Comparison of $PM_{2.5}$ and NO_2 ELAPSE models at ESCAPE monitoring sites

	Pollutant	$PM_{2.5}$					
					Measurements		
	Name ESCAPE area	R^2	RMSE	FB[a]	Mean	SD	N[b]
Overall		64.8	3.41	-0.02	15.86	5.73	416
Overall ELAPSE		58.7	2.85	-0.10	14.16	4.43	255
ELAPSE cohorts							
HUBRO	Oslo, NO	18.4	2.04	-0.30	8.59	2.20	19
CEANS	Stockholm County, SE	39.0	1.32	-0.04	8.29	1.64	19
DCH	Copenhagen, DK	40.1	1.26	-0.18	11.12	1.58	20
EPIC-NL	NL	12.6	1.71	-0.02	17.35	1.80	34
EPIC OXFORD	London- Oxford, Manchester, UK	7.6	2.23	-0.26	10.55	2.29	39
HNR	Ruhr Area, GER	65.5	0.97	-0.06	18.52	1.61	20
KORA	Munich-Augsburg, GER	31.5	1.44	-0.16	14.34	1.70	20
VHM&PP	Vorarlberg, AU	22.4	1.74	-0.19	13.34	1.92	20
E3N	Paris, FR	38.7	3.30	-0.24	16.02	4.10	20
EPIC VARESE	n.a.	-	-	-	-	-	-
DNC	n.a.	-	-	-	-	-	-
Administrative ELAPSE cohorts							
Dutch	NL	12.6	1.71	-0.02	17.35	1.80	34
English	London- Oxford, Manchester, UK	7.6	2.23	-0.26	10.55	2.29	39
Rome	Rome, IT	43.0	2.51	0.16	19.77	3.24	20
Danish	n.a.	-	-	-	-	-	-
Norwegian							
Swiss[3]	Lugano. CH	-	-	-	-	-	
Belgian[3]	Antwerp, BE	-	-	-	-	-	
	Pollutant	NO_2					
					Measurements		
	ESCAPE area	R^2	RSME	FB	Mean	SD	N
Overall		49.4	11.47	-0.08	29.32	16.12	1396
Overall ELAPSE		45.8	10.28	-0.13	29.74	13.95	780
ELAPSE cohorts							
HUBRO	Oslo, NO	7.0	12.74	-0.19	24.29	13.05	39
CEANS	Stockholm County, SE	55.0	5.03	-0.50	15.49	7.44	39
DCH	Copenhagen, DK	59.0	5.99	-0.54	17.82	9.21	41
EPIC-NL	NL	75.9	5.10	-0.26	28.76	10.32	68
EPIC OXFORD	London -Oxford, Manchester, Bradford, UK	53.9	8.64	-0.17	29.82	12.67	119
HNR	Ruhr Area, GER	54.0	6.74	-0.20	33.16	9.76	40
KORA	Munich-Augsburg, GER	64.0	5.79	-0.13	26.82	9.58	40
VHM&PP	Vorarlberg, AU	47.0	5.29	-0.10	22.59	7.17	40
E3N	Paris, Grenoble, Lyon, Marseille, FR	52.6	12.37	-0.01	34.42	17.90	160
EPIC VARESE	Varese, IT	34.0	13.78	0.10	36.53	16.54	20
DNC	n.a.	-	-	-	-	-	-
Administrative ELAPSE cohorts							
Dutch	NL	75.9	5.10	-0.26	28.76	10.32	68
English	London-Oxford, Manchester, Bradford, UK	53.9	8.64	-0.17	29.82	12.67	119
Rome	Rome, IT	51.0	9.72	0.23	42.64	13.71	40
Danish	n.a.	-	-	-	-	-	-
Norwegian	n.a.	-	-	-	-	-	-
Swiss	Basel, Geneva, Lugano. CH	13.7	7.55	-0.16	30.03	8.09	121
Belgian[c]	Antwerp, BE	-	-	-	-	-	

a. FB = Fractional Bias calculated as 2 * (mean observations - mean predictions)/(mean observations + mean predictions)
b. N = number of ESCAPE monitoring sites (the same for black carbon and $PM_{2.5}$)
c. Covers only a small part of the area, with insufficient number of sites

Table 4. Stability analysis at country level: predictions of the 2010 LUR model versus models from other years at randomly selected points (in squared correlation, R^2 in percentages, RMSE in μg/m^3)

Region	$PM_{2.5}$ 2013		NO_2 2005		NO_2 2000		O_3 2005a[a]		O_3 2000a[a]		O_3 2005c[a]		O_3 2000c[a]		O_3 2005w[a]		O_3 2000w[a]		N
	R^2 (%)	RMSE	R^2 (%)	RMSE	R^2 (%)	RMSE	R^2 (%)	RMSE	R^2 (%)	RMSE	R^2 (%)	RMSE	R^2 (%)	RMSE	R^2 (%)	RMSE	R^2 (%)	RMSE	
All West-European countries	88.2	1.9	91.9	1.9	90.9	2.0	85.8	3.5	78.8	4.3	80.4	4.3	44.3	7.3	84.3	4.6	76.4	5.6	44.000
ELAPSE countries																			
Combined	89.3	1.9	92.6	2.0	91.4	2.1	82.7	3.2	82.0	3.3	87.0	3.3	45.1	6.9	81.6	4.6	78.3	5.0	34762
Austria	60.1	2.0	86.7	1.3	87.4	1.9	81.9	3.7	82.7	4.1	80.9	3.4	67.4	6.7	82.5	3.4	64.5	3.9	1050
Belgium	84.1	1.0	90.9	1.4	84.6	2.3	81.5	1.9	87.4	1.9	89.6	1.9	81.7	2.4	86.5	2.0	70.6	2.5	352
Switzerland	52.5	1.9	91.5	1.2	92.6	1.8	94.6	2.4	95.2	2.7	88.2	3.3	85.5	5.1	87.9	3.5	88.7	4.6	503
Germany	57.6	1.2	85.0	1.3	80.5	2.2	64.0	2.7	69.2	3.0	75.5	3.1	29.4	4.7	47.3	3.8	63.7	4.4	4232
Denmark	48.8	1.1	88.8	0.8	84.8	1.6	73.0	1.2	71.1	1.3	71.0	1.6	59.6	1.8	63.6	1.5	73.2	1.6	527
France	57.4	1.5	89.0	1.1	82.9	1.9	83.2	2.7	80.4	3.5	87.6	3.0	55.0	5.2	76.3	3.4	86.8	4.1	6475
Italy	82.6	1.7	81.9	1.6	82.6	2.3	59.9	4.4	64.8	4.9	90.0	4.3	16.6	9.8	11.9	5.2	1.6	12.3	3548
Netherlands	70.1	0.9	87.9	1.6	81.9	2.7	60.4	2.2	71.8	2.1	73.0	2.3	35.6	3.0	79.3	2.2	53.1	2.6	454
Norway	59.3	0.9	83.3	0.5	83.4	0.8	88.6	1.7	79.4	2.4	79.0	2.2	71.7	3.1	61.1	3.0	79.4	2.4	3449
Sweden	86.2	0.9	93.1	0.5	91.3	0.8	65.5	1.6	45.1	2.2	78.9	1.7	63.3	2.9	76.6	1.6	87.4	1.7	5353
United Kingdom	89.8	1.2	95.3	1.1	93.0	2.0	71.8	2.0	78.1	2.1	81.9	3.3	74.3	3.4	52.2	2.5	53.0	3.3	2845
Non ELAPSE countries																			
Greece	64.4	1.2	86.5	0.9	83.3	1.6	40.9	3.7	49.5	3.8	14.2	6.6	6.0	7.6	34.7	3.9	19.4	5.1	1549
Finland	44.2	1.0	92.7	0.4	89.7	0.8	52.4	1.0	46.3	1.2	25.2	2.4	67.9	1.6	70.2	1.3	69.7	1.6	4008
Hungary	53.9	0.9	84.3	0.9	84.8	1.2	50.8	1.3	38.4	1.6	21.6	3.9	59.4	2.5	54.1	1.2	38.6	2.3	1118
Ireland	73.9	0.8	92.7	0.6	90.2	1.0	52.0	1.3	49.1	1.4	79.1	2.4	68.8	2.2	61.1	1.2	61.6	2.3	841
Lithuania	56.3	0.9	89.7	0.6	85.1	1.0	52.9	1.1	40.8	1.2	65.3	1.9	74.7	1.4	54.8	0.9	24.4	1.7	780
Luxembourg	68.3	0.9	89.0	1.3	77.9	2.2	73.9	1.3	75.4	1.4	74.1	2.6	78.3	2.2	47.2	1.8	57.7	1.4	31
Portugal	63.8	1.1	85.4	1.0	87.0	1.6	71.3	1.9	67.4	2.2	62.1	3.3	51.5	3.5	33.0	2.4	37.4	3.9	1015
Spain	69.4	1.1	77.8	1.2	79.7	1.7	65.6	2.8	58.5	3.6	62.8	4.4	41.4	5.6	42.9	3.4	38.9	7.0	5974

a. O3 a for annual, c for cold season and w for warm season.

Table 5. Correlations between concurrent AirBase measurements (background sites only) in 2010 with 2000 and 2005 (NO_2, O_3 annual, warm and cold season) and 2013 ($PM_{2.5}$) in R^2 (number of sites) for EU and separately for ELAPSE countries.

	NO_2		O_3 annual		O_3 warm		O_3 cold		$PM_{2.5}$
	2000	2005	2000	2005	2000	2005	2000	2005	2013
EU	85.8 (546)	86.7 (794)	71.6 (572)	72.3 (836)	68.3 (576)	67.7 (843)	77.9 (555)	79.5 (817)	79.3 (247)
Austria	86.1 (66)	94.6 (77)	87.8 (77)	89.9 (86)	72.1 (79)	79.5 (88)	91.3 (75)	92.4 (84)	96.7 (8)
Belgium	95.4 (16)	93.2 (26)	88.2 (22)	88.1 (28)	76.7 (22)	75.9 (29)	91.5 (22)	94.6 (25)	85.5 (19)
Switzerland	97.7 (21)	94.7 (21)	90.9 (21)	89.2 (23)	75.0 (21)	86.0 (23)	97.5 (21)	92.1 (23)	n.a. (0)
Germany	90.9 (185)	93.5 (213)	73.3 (181)	77.6 (206)	58.4 (182)	59.9 (206)	80.5 (175)	88.3 (201)	46.4 (63)
Denmark	n.a. (2)	93.4 (6)	n.a. (0)	41.0 (6)*	n.a. (0)	18.6 (6)*	n.a. (0)	72.7 (6)	95.5 (3)*
France	86.0 (169)	90.1 (261)	70.9 (179)*	82.5 (301)	66.3 (184)	82.0 (307)	80.1 (173)	85.7 (294)	52.5 (57)
Great Britain	88.2 (27)	90.0 (44)	72.9 (35)	71.7 (55)	67.5 (31)	66.1 (51)	77.7 (35)	76.8 (54)	59.4 (28)
Italy	65.9 (30)	73.7 (109)	38.0 (26)	20.5 (87)	20.3 (26)	1.2 (90)*	74.9 (23)	68.4 (88)	84.5 (44)
Netherlands	89.2 (23)	92.5 (26)	30.0 (19)	30.0 (25)	1.1 (19)*	2.6 (25)*	59.5 (20)	69.6 (23)	68.3 (15)
Norway	n.a. (2)	100 (3)	2.8 (6)	49.7 (7)	46.3 (6)*	72.4 (7)	73.2 (6)	91.1 (7)	15.5 (5)*
Sweden	96.6 (5)	96.8 (8)	67.5 (6)	0.8 (12)*	40.9 (6)*	15.4 (11)*	93.2 (5)	30.1 (12)	84.5 (5)

*not significant ($p>0.05$)

Supplementary material

1. Analysis of different $PM_{2.5}$ satellite products

We offered three different $PM_{2.5}$ satellite products to the $PM_{2.5}$ model development; (1) 10km product inferred 2009-2011; (2) 10km product for 2010; (3) 1km product for 2010. In preliminary models, the first data set led to better $PM_{2.5}$ models compared to the other 2 datasets. We further investigated the raw squared correlation coefficients (R^2) of the 3 data products (annual mean) with the annual mean $PM_{2.5}$ measurements from AirBase for the year 2010 (see Table S5 for more details). The difference in explained variance seems to be in the time period of the 3 products. The products 2 and 3 focusing on the year 2010 yielded similar correlations, irrespective of the 10 or 1km spatial resolution, explaining around 40% of variation. Product 1, which was inferred for 2009 to 2011 and optimized for 2010, explained 46% of variation. For the final $PM_{2.5}$ model we therefore decided to only offer the first $PM_{2.5}$ product.

2. Stability analysis at regional (NUTS1) level

NUTS areas are standard administrative divisions of EU countries for statistical purposes. The NUTS1 level is the first level. We also performed the stability analysis using the same 44,000 random points at the NUTS1 area level (see Figure S6). Like at the country level, there is a good agreement for all areas for NO_2 2000 and 2005 when compared to the 2010 surface (R^2 >0.60). For the other pollutants there is more heterogeneity in the correlation coefficients across areas. When comparing the $PM_{2.5}$ surfaces (2010 vs. 2013), the majority of the NUTS1 areas have a correlation coefficient > 0.40, with only a handful of areas dropping between 0.20 and 0.40. The comparison of the O_3 surfaces (2000, 2005 vs. 2013) shows a clear difference between annual and cold season versus the warm season. Both the 2000 and 2005 comparisons for warm season show a number of areas in the south of Europe with correlations of less than 0.20. This pattern is not observed in the annual and cold season comparisons.

Table S1. Descriptive statistics of $PM_{2.5}$, NO_2, BC, O_3 concentrations for 2010 used in the modelling procedure.

Air pollutant	Type	N[a]	Mean (µg/m³)	Median (µg/m³)	Std. Deviation (µg/m³)	Percentiles (µg/m³)			
						5	25	75	95
$PM_{2.5}$	Traffic	149	16.28	16.63	4.93	8.33	12.75	19.67	23.76
	Background	341	15.75	15.77	5.16	7.25	12.65	18.78	23.92
	Industrial	53	15.12	15.27	5.45	7.51	10.36	19.18	25.73
	All	543	15.84	15.88	5.13	7.65	12.42	19.25	23.94
NO_2	Traffic	740	40.23	38.69	14.62	19.90	30.39	47.75	66.06
	Background	1287	21.47	21.01	9.82	5.67	14.81	27.94	37.27
	Industrial	372	19.43	18.22	10.13	4.47	11.52	26.72	38.06
	All	2399	26.94	25.12	14.59	6.73	16.34	34.70	53.69
BC[b]	Traffic	207	2.28	2.16	0.90	0.98	1.65	2.83	3.99
	Background	229	1.51	1.47	0.54	0.74	1.08	1.85	2.44
	Industrial	0	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.
	All	436	1.88	1.78	0.83	0.84	1.26	2.33	3.50
O_3 annual[c]	Traffic	228	65.01	63.86	12.26	45.80	57.10	73.44	88.72
	Background	1323	72.10	70.92	10.66	56.25	65.71	77.76	91.86
	Industrial	194	75.77	76.59	11.76	57.42	67.12	84.31	94.49
	All[c]	1747	71.59	70.61	11.35	53.75	64.54	78.31	91.85
O_3 warm[c]	Traffic	225	81.62	81.34	15.03	57.65	70.69	92.51	106.04
	Background	1311	89.99	89.30	12.99	69.09	82.22	98.45	111.84
	Industrial	192	90.70	92.20	13.45	67.13	81.61	100.18	110.61
	All[d]	1730	88.98	88.95	13.62	66.39	80.97	97.96	111.08
O_3 cold[c]	Traffic	223	48.31	47.13	12.29	29.96	39.44	57.52	72.03
	Background	1304	54.05	53.18	12.69	35.34	44.92	62.17	77.63
	Industrial	188	60.16	62.26	13.55	37.90	49.77	69.77	82.50
	All[e]	1716	53.98	53.04	13.05	34.11	44.34	62.72	77.42

a. Number of sites
b. BC monitoring data from ESCAPE, no industrial type monitoring sites were used. Measured as absorbance of $PM_{2.5}$ ($10^{-5}m^{-1}$)
c. O_3 concentrations were calculated as the average of the daily maximum running 8-hour mean; warm season is from 1 April to 30 September; cold season is from 1 October to 31 March falling in the same year.
d. These include 2 sites with site type ‘Unknown’
e. These include 1 site with site type ‘Unknown’

Table S2. GIS predictor variables

Data set	Predictor variable	Name variable	Year	Buffer Size (radius in m) or point estimate	Pre-specified direction of effect for $PM_{2.5}$, NO_2, BC (O_3)
$PM_{2.5}$ ($\mu g/m^3$) derived from MODIS on board the Terra satellite: ~10km	Surface $PM_{2.5}$ concentration derived from satellite	SAT-PM25	2010, 2013	Point	+
$PM_{2.5}$ ($\mu g/m^3$) derived from MODIS on board the Terra satellite: ~1km	Surface $PM_{2.5}$ concentration derived from satellite	SAT-PM25-1k	2010, 2013	Point	+
NO_2 ($\mu g/m^3$) derived from OMI on board the Aura satellite: ~10km	Surface NO_2 concentration derived from satellite	SAT-NO_2	2010	Point	+(-)
$PM_{2.5}$, NO_2, O_3 ($\mu g/m^3$) estimated by MACC-II Ensemble model: ~10km	Surface $PM_{2.5}$,NO_2 and O_3 concentration from dispersion model	MACC-PM25 MACC-NO_2 MACC-O_3	2010, 2013	Point	+(-)
$PM_{2.5}$, NO_2, O_3 and BC ($\mu g/m^3$) estimated by DEHM: ~50km	Surface $PM_{2.5}$, NO_2, O_3 and BC concentration from dispersion model	DEHM-$PM_{2.5}$ DEHM-NO_2 DEHM-O_3 DEHM-BC	2000, 2005, 2010	Point	+(-)
EuroStreets roads (length in m)	Major roads All roads	MJRD ALRD	2010	50; 100; 200; 300; 400; 500; 700; 1000; 2000; 5000; 10000	+(-)
Corine land cover: 100m	Industry/commercial Ports Urban green Total built up[a] Natural land Residential[b]	IND POR UGR TBU NAT RES	2006	50; 100; 200; 300; 400; 500; 600; 700; 800; 1000; 1200; 1500; 1800; 2000; 2500; 3000; 3500; 4000; 5000; 6000; 7000; 8000; 10000	+(-) +(-) -(+) +(-) -(+) +(-)
GEOSTAT 2011 population grid dataset: ~1km	Sum of population	POP	2011	Point	+(-)
Altitude SRTM DTM: ~90 m	Altitude – transformed[c]	ALT		Point	-(+)
Trend	North-South and East-West trend	X, Y[d]		Point	n.a.

[a]Residential + Ind/comm + Port + transport infrastructure, airports, mines, dumps and construction sites
[b]continuous urban fabric (high density housing) + discontinuous urban fabric (low density housing)
[c]Transformed altitude is calculated as $\sqrt{(nalt/max(nalt))}$, where nalt = altitude − min(altitude).
[d]Coordinates were truncated : X =x –xmin / (xmax – xmin); Y =y –ymin / (ymax – ymin)

Table S3. Structure and performance of LUR models for BC, NO_2 and O_3 for full dataset and five hold-out validation datasets

BC models[a]

Theme	Variables[b]	FULL	HOV1	HOV2	HOV3	HOV4	HOV5
	Constant	0.99	0.85	0.99	0.94	0.99	1.10
CTM	MACC-PM25	0.85	0.95	0.80	0.92	0.95	0.65
Satellite	SAT-PM25	0.30	0.34	0.39	0.32	0.22	0.34
Major roads	$MJRD_{50}$		0.64				
	$MJRD_{100}$	0.68		0.68	0.67	0.65	0.44
All roads	$ALRD_{50}$	0.40			0.39	0.40	0.52
	$ALRD_{1000}$				0.34		
	$ALRD_{700}$	0.45	0.67	0.40		0.43	0.57
Residential	RES_{2500}			0.93	0.86		
	RES_{3000}	0.90	0.86			0.86	0.94
Urban green	UGR_{1000}	-0.12	-0.18	-0.14	-0.11	-0.19	-0.08
Ycoord	Y	-1.16	-1.09	-1.16	-1.16	-1.16	-1.30
Training	R^2	54.4	52.2	52.3	53.9	57.6	55.3
(LUR)	RMSE	0.56	0.59	0.56	0.58	0.54	0.572
HOV	R^2	51.4	56.7	56.9	53.8	43.2	49.3
(LUR)	RMSE	0.58	0.52	0.62	0.52	0.67	0.54

a. Regression slope in $10\text{-}5m^{-1}$ were multiplied by the difference between the 1st and 99th percentile of each predictor to allow comparison across predictors

b. ALT = altitude, ALRD = all roads, MJRD = major roads, IND = industry, POR = ports, UGR = urban green, TBU = total build up, NAT = natural land, RES = residential, POP = sum of population, X = North-South trend, Y = East-West trend, SAT = satellite, MACC = MACC dispersion model, DEHM = DEHM CTM. Number in subscript depicts the buffer size (e.g. $ALRD_{100}$ = sum of all road length within 100m)

NO_2 models[a]

Theme	Variable[b]	FULL	HOV1	HOV2	HOV3	HOV4	HOV5
	Constant	3.30	3.30	3.70	4.70	3.30	3.20
CTM	MACC-NO_2	22.73	22.73	23.30	22.73	22.16	22.73
All Roads	$ALRD_{50}$	7.04		8.60	6.65	9.38	
	$ALRD_{100}$		9.68				10.29
	$ALRD_{300}$	3.92			7.50		
	$ALRD_{1000}$					12.27	
	$ALRD_{2000}$	15.73	16.71	16.71			15.73
Major roads	$MJRD_{50}$		5.70				
	$MJRD_{100}$	12.32		12.32	10.67	12.32	12.32
	$MJRD_{200}$		6.95				
Natural	NAT_{400}	-3.38			-3.97	-3.09	-3.14
	NAT_{500}		-3.24	-3.89			
Ports	POR_{200}			1.82			1.80
	POR_{700}	4.10	4.51			2.87	
	POR_{1800}				4.14		
Residential	RES_{200}				2.73		
	RES_{300}	5.80		6.38			
	RES_{400}		6.86				6.37
	RES_{2500}				14.10		
	RES_{8000}					7.09	
Total build up	TBU_{300}					6.09	
Urban green	UGR_{100}				-3.20		
	UGR_{200}			-2.40			
Training	R^2	58.8	59.1	58.3	58.4	59.0	59.6
(LUR)	RMSE	9.38	9.38	9.39	9.38	9.36	9.36
HOV	R^2	57.5	57.8	59.9	60.2	54.8	54.7
(LUR)	RMSE	9.51	9.36	9.44	9.44	9.81	9.51

a. Regression slope in $\mu g/m^3$ were multiplied by the difference between the 1st and 99th percentile of each predictor to allow comparison across predictors

b. ALT = altitude, ALRD = all roads, MJRD = major roads, IND = industry, POR = ports, UGR = urban green, TBU = total build up, NAT = natural land, RES = residential, POP = sum of population, X = North-South trend, Y = East-West trend, SAT = satellite, MACC = MACC dispersion model, DEHM = DEHM CTM. Number in subscript depicts the buffer size (e.g. $ALRD_{100}$ = sum of all road length within 100m)

O_3 models[a]

		Annual						Warm						Cold					
Theme	Variable[b]	FULL	HOV1	HOV2	HOV3	HOV4	HOV5	FULL	HOV1	HOV2	HOV3	HOV4	HOV5	FULL	HOV1	HOV2	HOV3	HOV4	HOV5
	Constant	40.54	42.72	40.91	40.02	39.30	40.16	30.00	30.00	30.00	32.00	29.00	29.00	1.00	0.23	1.10	1.60	0.19	0.14
MACC	MACC-O_3	25.51	25.04	24.93	25.38	26.35	26.45							37.62	37.62	37.62	36.87	37.62	37.62
DEHM	DEHM_O_3[c]							32.57	32.57	32.57	31.71	33.43	33.43						
Roads	ALRD50	-2.49	-3.70		-3.14				-3.04	-3.35	-4.87				-3.35				
	ALRD200	-4.75		-7.04	-4.25	-7.13	-6.29	-6.87				-6.09	-4.52	-3.53		-3.14	-4.52	-3.53	-3.34
	ALRD500								-5.81	-6.18	-5.54								
	ALRD700						-5.17						-3.40						
	ALRD1000		-5.83													-3.04			
	ALRD2000											-4.99							
	MJRD50			-3.69			-3.03							-3.48		-3.48	-4.09		-2.66
	MJRD100					-4.67		-6.03		-5.85		-6.63	-7.24		-3.32				
	MJRD200	-3.24	-4.42		-3.48				-5.85		-5.96							-4.01	
	MJRD700													-1.61	-1.90				-2.49
Nature	NAT700													5.81	6.11	5.07		5.81	6.56
	NAT800																6.70		
Ports	POR4000	-1.57		-1.67		-1.63						-5.68							
	POR5000				-1.63		-2.20	-5.95	-4.95	-5.45	-5.95		-7.43						
Residential	RES100		-1.84									-1.75							
	RES200																		-3.38
	RES400															-3.58			
	RES500	-1.94			-1.91	-4.29													
	RES700														-4.77				
	RES800																	-5.32	
	RES1000						-5.02												
	RES1200												-3.49	-4.18					
	RES1500																-5.16		
	RES2000	-4.13	-2.38	-6.02	-3.81			-4.79	-3.99										
	RES3000									-4.01	-3.54								
Industrial/ commercial	IND50		-2.42												-2.20				
	IND200															-2.21			
Total build up	TBU100													-1.10					
Urban green	UGR1000													2.21	2.21		2.78	2.59	
Altitude	ALT	8.82	7.82	10.02	9.33	8.67	8.38	5.70	4.99	6.56	5.99	5.13	5.27	6.84	6.13	8.12	7.13	6.98	6.27
Coordinate	X	2.48	1.65	3.57	2.55	2.62													
	Y	-10.05	-11.01	-10.93	-9.34	-10.07	-8.82												
Training (LUR)	R^2	65.1	65.4	68.5	63.4	64.6	63.9	45.5	44.9	48.9	45.0	45.6	44.7	67.7	68.1	69.8	66.2	67.6	67.0
	RMSE	6.73	6.66	6.39	6.93	6.81	6.81	10.07	10.09	9.69	10.10	10.17	10.18	7.43	7.35	7.25	7.60	7.53	7.43
HOV (LUR)	R^2	63.4	64.1	51.1	73.5	64.3	67.7	44.5	48.1	34.9	48.8	45.1	49.5	66.5	65.8	58.7	73.8	67.4	68.5
	RMSE	6.87	7.00	7.94	5.71	6.65	6.57	10.15	9.98	11.33	9.83	9.69	9.58	7.55	7.84	8.16	6.74	7.11	7.67
Training (LUR + Kriging)	R^2							69.6	69.5	73.3	70.0	68.5	67.3	83.3	84.4	85.2	80.3	83.8	82.4
	RMSE							7.51	7.49	6.98	7.44	7.73	7.81	5.33	5.13	5.05	5.78	5.32	5.42
HOV (LUR + Kriging)	R^2							59.9	61.5	48.0	61.9	6.50	65.8	75.3	71.6	71.8	83.3	76.8	75.1
	RMSE							8.63	8.60	10.13	8.48	7.74	7.88	6.49	7.15	6.75	5.39	5.99	6.82

a. Regression slope $\mu g/m^3$ were multiplied by the difference between the 1st and 99th percentile of each predictor to allow comparison across predictors

b. ALT = altitude, ALRD = all roads, MJRD = major roads, IND = industry, POR = ports, UGR = urban green, TBU = total build up, NAT = natural land, RES = residential, POP = sum of population, X = North-South trend, Y = East-West trend, SAT = satellite, MACC = MACC dispersion model, DEHM = DEHM CTM. Number in subscript depicts the buffer size (e.g. $ALRD_{100}$ = sum of all road length within 100m)

c. DEHM estimates were calculated for each season (annual, warm and cold)

Table S4. Details of 2000, 2005 (NO_2 and O_3) and 2013 ($PM_{2.5}$) FULL models[a]

Theme	Variable[b]	$PM_{2.5}$ 2013	NO_2 2000	NO_2 2005	O_3 2000a	O_3 2000w	O_3 2000c	O_3 2005a	O_3 2005w	O_3 2005c
	Constant	0.45	6.48	8.39	56.06	35.88	59.90	72.00	61.92	56.60
CTM	MACC-PM25	13.21								
	DEHM-NO_2		11.55	12.49						
	DEHM-O_3[c]				11.19	32.33		7.24	20.34	
Satellite	SAT-PM25	6.54								
Roads	$ALRD_{50}$	1.09	14.07	9.96				-6.39	-10.29	-8.16
	$ALRD_{100}$					-7.84	-6.07			
	$ALRD_{200}$				-7.15			-6.76		
	$ALRD_{1000}$			9.20						
	$ALRD_{2000}$					-21.31			-17.76	
	$ALRD_{5000}$		32.43		-16.57			-21.84		
	$MJRD_{50}$				-6.78		-4.12			
	$MJRD_{100}$			11.67		-6.92		-2.21	-5.96	
	$MJRD_{500}$									-2.93
	$MJRD_{10000}$						-15.42			-16.64
Nature	NAT_{400}			-3.23			3.87			
	NAT_{500}		-4.05							
	NAT_{600}									5.20
	NAT_{1000}	-3.17								
	NAT_{10000}						5.16			4.85
Urban green	UGR_{100}					6.00				
	UGR_{1000}	-1.09								
Ports	POR_{300}								-2.18	
	POR_{1000}		4.42							
	POR_{6000}			3.73					-4.38	
Industrial/ commercial	IND_{200}									-5.07
Residential	RES_{300}			4.64						
	RES_{700}		8.20							-7.30
Total build up	TBU_{300}	1.51								
	TBU_{400}							-3.53		
	TBU_{500}				-5.67					
	TBU_{600}						-8.25			
Altitude	ALT				15.30	14.38	18.36	12.75	9.85	12.60
X-coord	X				9.94		-12.42			
Y-coord	Y				-11.83			-14.40	-11.52	
Training (LUR)	R^2	65.6	55.9	52.2	59.6	53.5	51.8	48.8	42.0	45.0
	RMSE	2.87	10.59	11.34	8.56	11.39	9.17	9.45	11.93	10.13
HOV (LUR)	R^2	0.64	53.8	50.1	58.0	52.2	50.1	46.6	40.6	43.3
	RMSE	2.93	10.57	11.57	8.69	11.53	9.31	9.62	12.04	10.25
Training (LUR + Kriging)	R^2					81.5				69.9
	RMSE					7.17				7.47
HOV (LUR + Kriging)	R^2					63.8				63.4
	RMSE					10.02				8.24

a. Regression slope μg/m^3 were multiplied by the difference between the 1st and 99th percentile of each predictor to allow comparison across predictors

b. ALT = altitude, ALRD = all roads, MJRD = major roads, IND = industry, POR = ports, UGR = urban green, TBU = total build up, NAT = natural land, RES = residential, POP = sum of population, X = North-South trend, Y = East-West trend, SAT = satellite, MACC = MACC dispersion model, DEHM = DEHM CTM. Number in subscript depicts the buffer size (e.g. $ALRD_{100}$ = sum of all road length within 100m)

c. DEHM O_3 estimates were calculated for each season (annual, warm and cold)

Table S5. Comparison of predictions of different satellite derived $PM_{2.5}$ (SAT) products with routine $PM_{2.5}$ concentrations.

AirBase Site type	Inferred 2009-2011, 10x10km, v3.01		Inferred 2010, 10x10km, v3.01		Inferred 2010, 1x1km, v4.GL.02.NoGWR		
	R^2	RMSE	R^2	RMSE	R^2	RMSE	N
All sites	0.46	3.77	0.39	4.01	0.40	3.98	545
No industrial sites	0.47	3.71	0.38	4.02	0.38	4.00	492
Traffic only	0.46	3.62	0.42	3.75	0.42	3.76	149
Background only	0.49	3.68	0.37	4.11	0.38	4.07	342

Table S6. Population (2010) by classes of modelled air pollution estimates ($PM_{2.5}$ and NO_2 FULL models)

Pollutant	Concentration class (µg/m^3)	Area (1,000,000 m2)	Percentage of total (%)	Population[a] (millions)	Percentage of total (%)
$PM_{2.5}$	<5	1769.41	11.32	4.51	1.08
	5-<10	3592.20	22.98	42.31	10.12
	10-<15	7016.49	44.88	173.87	41.57
	15-<20	2762.81	17.67	155.38	37.15
	20-<25	404.38	2.59	33.68	8.05
	25-<30	87.30	0.56	7.96	1.90
	>30	2.17	0.01	0.53	0.13
Total		15634.76		418.24	
NO_2	<10	7335.38	46.93	32.72	7.82
	10-<15	4156.22	26.59	60.36	14.43
	15-<20	2351.02	15.04	77.17	18.45
	20-<30	1500.96	9.60	139.49	33.35
	30-<40	240.35	1.54	76.54	18.30
	40-<60	45.95	0.29	30.92	7.39
	>60	1.54	0.01	1.07	0.26
Total		15631.42		418.26	

a. Population was derived from 2011 census data

Table S7. Population weighted annual concentration (µg/m^3) averaged over Western Europe

	Pollutant (µg/m^3)				
Year	$PM_{2.5}$	NO_2	O_3 annual	O_3 cold	O_3 warm
2000		20.09	55.37	39.94	70.02
2005		15.41	57.97	43.48	71.86
2010	11.17	18.85	57.11	42.70	70.51
2013	10.58				