Science of the Total Environment 724 (2020) 138102

Contents lists available at ScienceDirect

Science of the Total Environment

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

A multi-city air pollution population exposure study: Combined use of chemical-transport and random-Forest models with dynamic population data

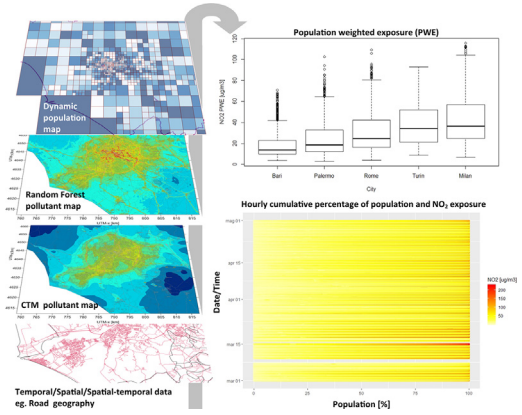
Claudio Gariazzo [a,*], Giuseppe Carlino [b], Camillo Silibello [c], Matteo Renzi [d], Sandro Finardi [c], Nicola Pepe [c], Paola Radice [c], Francesco Forastiere [e,f], Paola Michelozzi [d], Giovanni Viegi [e], Massimo Stafoggia [d], On behalf of the BEEP Collaborative Group [1]

[a] Occupational and Environmental Medicine, Epidemiology and Hygiene Department, Italian Workers' Compensation Authority (INAIL), Monte Porzio Catone (RM), Italy
[b] Simularia, Turin, Italy
[c] Arianet, Milan, Italy
[d] Department of Epidemiology, Lazio Regional Health Service, ASL Roma 1, Rome, Italy
[e] Institute of Biomedicine and Molecular Epidemiology "Alberto Monroy", National Research Council Palermo, Italy
[f] Environmental Research Group, King's College, London, UK

HIGHLIGHTS

- Machine learning methods were applied to obtain pollutant concentration in urban areas.
- Population weighted exposure was estimated using dynamic mobile phone location data.
- Long term NO_2, PM, and O_3 daily concentrations were provided for 6 urban areas.
- Differences among cities were found with spatial/geographical concentration gradients.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 29 November 2019
Received in revised form 13 March 2020
Accepted 20 March 2020
Available online 26 March 2020

ABSTRACT

Cities are severely affected by air pollution. Local emissions and urban structures can produce large spatial heterogeneities. We aim to improve the estimation of NO_2, O_3, $PM_{2.5}$ and PM_{10} concentrations in 6 Italian metropolitan areas, using chemical-transport and machine learning models, and to assess the effect on population exposure by using information on urban population mobility. Three years (2013–2015) of simulations were performed by the Chemical-Transport Model (CTM) FARM, at 1 km resolution, fed by boundary conditions provided

* Corresponding author.
E-mail address: c.gariazzo@inail.it (C. Gariazzo).

[1] Carla Ancona, Paola Angelini, Stefania Argentini, Sandra Baldacci, Lucia Bisceglia, Michela Bonafede, Sergio Bonomo, Laura Bonvicini, Serena Broccoli, Giuseppe Brusasca, Simone Bucci, Giuseppe Calori, Giuseppe Carlino, Achille Cernigliaro, Antonio Chieti, Annamaria Colacci, Francesca de' Donato, Moreno Demaria, Salvatore Fasola, Sandro Finardi, Francesco Forastiere, Claudia Galassi, Claudio Gariazzo, Paolo Giorgi Rossi, Stefania La Grutta, Gaetano Licitra, Sara Maio, Alessandro Marinaccio, Paola Michelozzi, Enrica Migliore, Antonino Moro, Alessandro Nanni, Marta Ottone, Federica Parmagnani, Nicola Pepe, Paola Radice, Andrea Ranzi, Matteo Renzi, Salvatore Scondotto, Matteo Scortichini, Camillo Silibello, Roberto Sozzi, Massimo Stafoggia, Gianni Tinarelli, Francesco Uboldi, Giovanni Viegi, Nicolas Zengarini.

https://doi.org/10.1016/j.scitotenv.2020.138102

Editor: Lidia Morawska

Keywords:
Machine learning
Dispersion model
Urban area
Population mobility
Particulate matter
Gaseous pollutants

by national-scale simulations, local emission inventories and meteorological fields. A downscaling of daily air pollutants at higher resolution (200 m) was then carried out by means of a machine learning Random-Forest (RF) model, considering CTM and spatial-temporal predictors, such as population, land-use, surface greenness and vehicular traffic, as input. RF achieved mean cross-validation (CV) R^2 of 0.59, 0.72, 0.76 and 0.75 for NO_2, PM_{10}, $PM_{2.5}$ and O_3, respectively, improving results from CTM alone. Mean concentration fields exhibited clear geographical gradients caused by climate conditions, local emission sources and photochemical processes. Time series of population weighted exposure (PWE) were estimated for two months of the year 2015 and for five cities, by combining population mobility data (derived from mobile phone traffic volumes data), and concentration levels from the RF model. PWE_RF metric better approximated the observed concentrations compared with the predictions from either CTM alone or CTM and RF combined, especially for pollutants exhibiting strong spatial gradients, such as NO_2. 50% of the population was estimated to be exposed to NO_2 concentrations between 12 and 38 μg/m^3 and PM_{10} between 20 and 35 μg/m^3. This work supports the potential of machine learning methods in predicting air pollutant levels in urban areas at high spatial and temporal resolutions.

1. Introduction

Air pollution is known to cause health effects on the general population. According to different Organizations and Commissions, about 4.2 million of deaths were attributable to $PM_{2.5}$ ambient concentrations during 2015 (Cohen et al., 2017; Ostro et al., 2018). In addition to mortality, air pollution is associated with incidence of various debilitating diseases such as chronic obstructive pulmonary disease, ischemic heart diseases and cerebrovascular events (Brook et al., 2017; Cesaroni et al., 2014; Scheers et al., 2015; Schikowski et al., 2014; Stafoggia et al., 2014). Recent studies also identified a positive association between nitrogen oxides exposures and prevalence of diabetes (Renzi et al., 2018) and incidence of dementia related to air pollution exposure (Chen et al., 2017).

About 55% of the world's population lives in urban areas. Some of them, and particularly those living in metropolitan areas, are exposed to high pollution levels caused by urban sources (i.e. traffic, domestic heating, industry), low dispersion conditions determined by the presence of buildings and emissions from sources located outside the city (regional background) such as agriculture, natural sources and distant industries (Harrison, 2018). In these areas, strong spatial inhomogeneities in air pollution levels occur, particularly for NO_2, that need to be accurately estimated to properly quantify the burden of mortality and morbidity attributable to ambient air pollutants. To obtain information on the spatial distribution of pollutant levels, two main approaches have been used by the epidemiologists: air quality models and statistical methods.

Gaussian plume or puff models are often used to describe the distribution of traffic related air pollutants at local (up to 1 km) and urban (up to 10 km) scales (Forehead and Huynh, 2018). Since these models generally do not include the treatment of gas-phase reactions and aerosol processes, their application is limited to primary pollutants. Chemical Transport Models (CTM) represent an alternative to Gaussian models thanks to their ability to estimate secondary pollutants such as ozone, NO_2 and particulate matter secondary components (Zhang et al., 2012; Kukkonen et al., 2012, 2016; Gariazzo et al., 2007). In most cases, their application has been limited to a spatial resolution of 1 km, therefore resulting in a misclassification of local scale phenomena such as those occurring in street canyons.

Statistical linear regression models have been often preferred to air quality models, as they are easier to apply and reiterate on a yearly basis. Land Use Regression (LUR) models are examples of such an approach (Hoek, 2017; Hoek et al., 2008; Cesaroni et al., 2012). LUR models use the measurements as the dependent variable and land-use related data, such as population, distance from main roads, land cover, etc. as the independent variables (predictors) in a multivariate regression model. Air pollution levels are then predicted at any location, such as individual addresses, using the parameter estimates derived from the regression model. Changes over time of land use characteristics and/or emissions may however introduce spatial-temporal errors in population exposure estimates.

Recent studies used machine-learning (ML) methods to predict high-resolution pollution maps (Chen et al., 2018a, 2018b; Stafoggia et al., 2019; de Hoogh et al., 2019; Araki et al., 2018; Di et al., 2019). Random-Forest (RF) algorithms represent a family of ML methods consisting in an ensemble of decision trees (forest) that are able to capture complex and non-linear relationships between predictor variables. Data about observations and spatial-temporal predictors are used to construct a set of trees from which an ensemble prediction of the target variable (e.g. pollutant concentration) is obtained. This method allows deriving accurate long time-series of daily pollutant concentrations at a very fine spatial scale, typically 1 square km for nation-wide studies, eligible for epidemiology studies. RFs models have been used to predict nation-wide particulate matter, starting from satellite aerosol optical depth data and spatial-temporal predictors (Chen et al., 2018a, 2018b; Stafoggia et al., 2019). Another recent study predicted NO_2 at high resolution across Switzerland using both mixed and RF models driven by satellite NO_2 data collected by Ozone Monitoring Instrument (OMI) (de Hoogh et al., 2019). Araki et al. (2018) applied a RF approach for estimating metropolitan monthly NO_2 exposure in Japan using satellite derived OMI NO_2 data. Di et al. (2019) used an ensemble-base ML model to predict $PM_{2.5}$ over United States at 1 km × 1 km grid cells, with a focus over the Great Boston Area at 100 m × 100 m using a downscaling model.

Although ML methods are promising techniques to estimate exposure, studies at urban scale are rare. The estimation of long time-series of multi-pollutants concentrations at high spatial and temporal resolution would be needed to carry out high quality epidemiology studies on urban resident population.

Despite the aforementioned literature on computational methods developed for population exposure, there is a need to increase the spatial and/or temporal resolution of concentrations estimates in urban areas, aimed at reducing misclassification of exposure (Özkaynak et al., 2013; Health Effects Institute, 2009). A possible way to improve the accuracy of concentration estimates is to combine modelling techniques in which the results from a model are used as an input to the following one. Former studies applied different methods for models combination such as: neural network downstream to a dispersion model (Pelliccioni et al., 2003; Pelliccioni and Tirabassi, 2006), hybrid modelling frameworks combining regional and local scale models (Parvez and Wagstrom, 2019), Bayesian ensemble approaches combining satellite and CTM $PM_{2.5}$ data (Murray et al., 2019) and ensemble-based machine-learning models (Di et al., 2019; Shtein et al., 2020).

The aim of this study was to improve the air quality assessment in urban areas, by combining CTM and RF results, to support epidemiological studies at high spatial resolution. This study was carried out in the frame of BEEP (Big data in Environmental and occupational EPidemiology) project whose main goals are to improve exposure assessment and to support environmental epidemiological studies in Italy by collecting,

linking and analysing large amount of data from different sources. Within the BEEP project, a multi-city exposure assessment study has been carried out in six urban areas across Italy from north to south. For five of them a time-limited dynamic population exposure study has been conducted using mobile phone traffic data. These data have been coupled in time and space with pollutants concentrations, provided by the above RF modelling techniques, allowing for a dynamic population exposure analysis. In the following section (Section 2), the urban areas and their monitoring networks, the urban CTM, the RF application to downscale at finer spatial resolution and the dynamic population data used to derive population-weighted exposure are described. Section 3 presents the validation of model's results with observations and population exposure results by city. A discussion follows in Section 4.

2. Materials and methods

2.1. Urban areas description

The BEEP project analysed the health effects of air pollution in six regions located in different parts of the country in order to consider different environmental and climate situations, urbanization levels and economic conditions related to country's heterogeneities. The regions chosen by the BEEP study were Piedmont, Lombardy, Emilia Romagna, Lazio, Apulia and Sicily and their capital cities Turin, Milan, Bologna, Rome, Bari and Palermo were selected as target urban areas. Fig. 1 shows a map of the location of the six urban areas. These cities, spanning from North to South, are characterised by different urban structures and environmental/climate conditions (see Table 1). Health effects caused by air pollution in these areas have been previously reported (Cesaroni et al., 2012, 2014; Tramuto et al., 2011; Baccini et al., 2017; Chiusolo et al., 2011).

Turin, Milan and Bologna are in the Po Valley, a well-known highly entropized area located in Northern Italy, characterised by the presence of many urban settings, industrial facilities, agricultural activities and breeding farms. During wintertime, this area is affected by long-lasting and intense atmospheric stability periods that cause rapid degradation of air quality (Pernigotti et al., 2012; Bigi and Ghermandi, 2014; Perrino et al., 2014). The mean annual NO_2 and PM_{10} concentrations and the corresponding number of days with daily PM_{10} concentration exceeding 50 μg/m^3 (maximum allowed 35 days per year) and number of hours with NO_2 concentrations higher than 200 μg/m^3 are provided

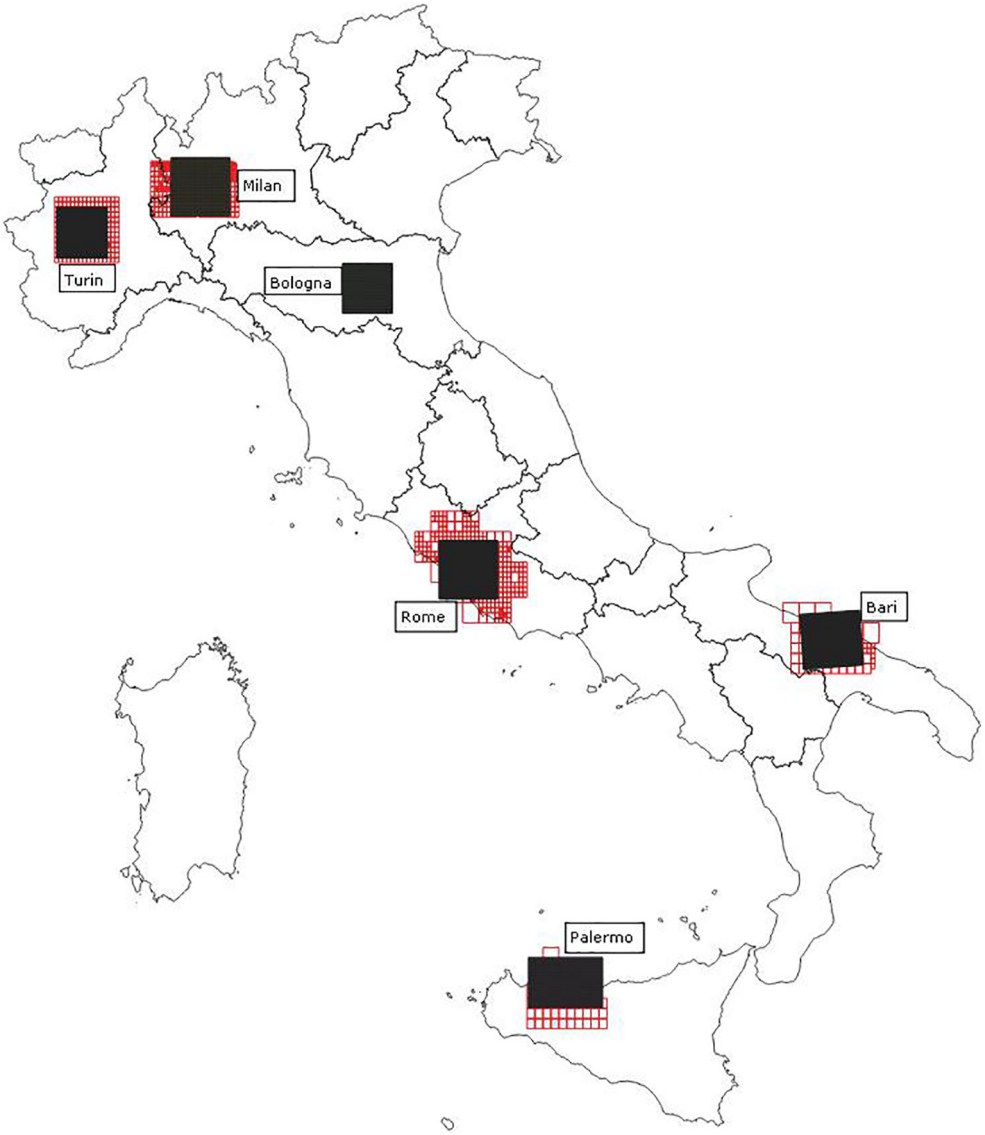


Fig. 1. Map of cities location (upper figure) and grids of chemical-transport model (grey) and of dynamic population data (red) over the city of Milan in its Province (blue) (bottom figure). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

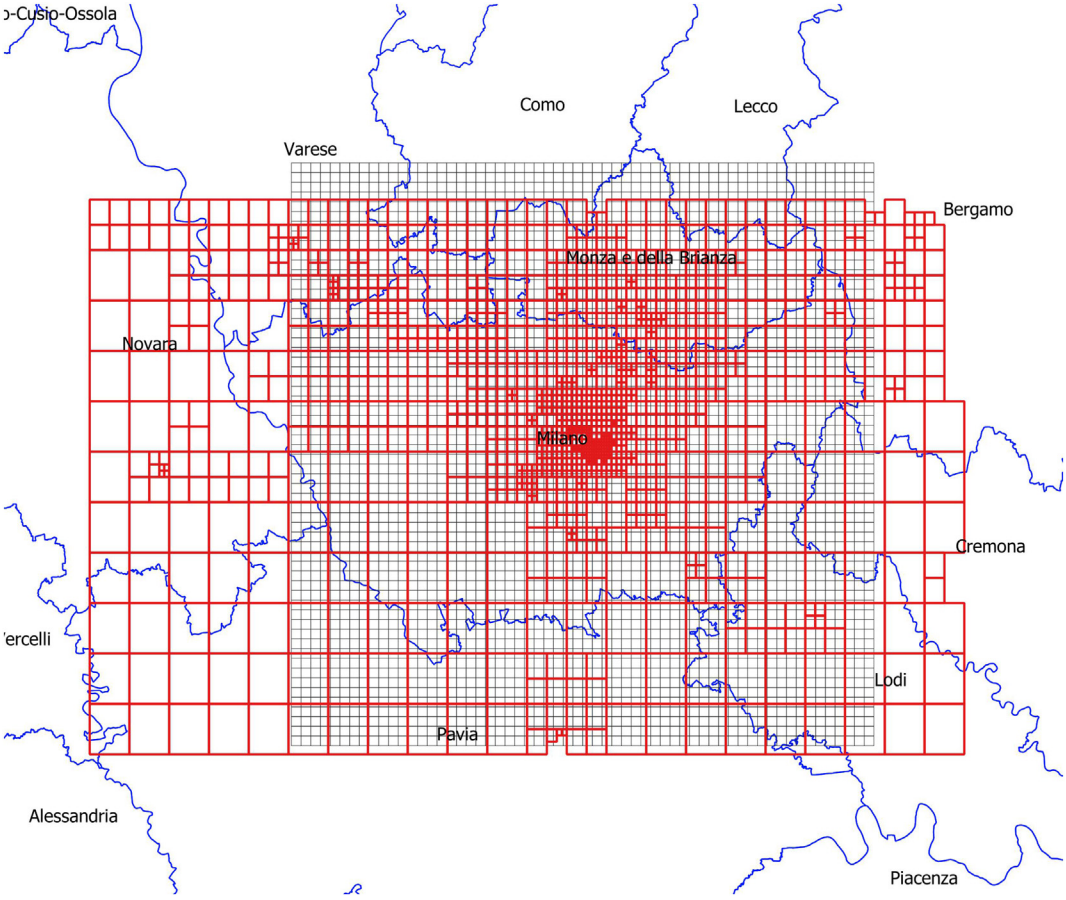


Fig. 1 (*continued*).

in Table 1, confirming the impact of these stagnant conditions on air quality. The metropolitan area of Rome is also affected by air pollution episodes (Gariazzo et al., 2007), although the proximity of the shoreline limits their number, due to better dispersion conditions caused by frequent sea-land breezes. As for Bari and Palermo, both located in Southern Italy, the temperate climate limits the use of domestic heating and urban air pollution is mainly caused by traffic-related sources, although the proximity of commercial harbours may contribute as well.

2.2. Urban areas air pollution monitoring networks

Air pollutants observed concentrations were used as training data for the development of the RF models that will be described in Section 2.4. Hourly values of measured NO_2 and O_3 concentrations, as well as values of PM_{10} and $PM_{2.5}$ daily concentrations were collected from the local institutional monitoring networks for the years 2013–2015. Gaseous pollutants were then averaged on a daily basis to be used in the training of the RF model. According to the EC legislation, the air pollution monitoring stations are classified based on the "Type of area" and "Type of station". The former refers to the environment on a scale of several kilometres ("urban", "suburban", "rural") while the latter refers to the impact (or absence) of near-by emissions ("traffic", "industrial", "background"). Table 2 shows the number of stations used in the study divided by type, pollutant and year. Most of the stations represented urban traffic and urban/suburban background conditions. On average, 84, 42, 73 and 35 stations were used for NO_2, O_3, PM_{10} and $PM_{2.5}$ respectively.

Maps of locations of monitoring stations by pollutant and city are shown in Supplementary Materials (SM) (Fig. S1).

2.3. The Air Quality Modelling System (AQMS)

The numerical simulations of meteorological parameters and airborne pollutants have been performed by an AQMS based on the chemical transport model (CTM) FARM (Flexible Air quality Regional Model) (Gariazzo et al., 2007, Silibello et al., 2008). Physical and chemical processes influencing the concentration fields within the modelling domains are described in FARM by a system of partial differential equations expressing the time variation of the average concentrations. A detailed description of AQMS application is illustrated in SM. The AQMS includes subsystems used to reconstruct the atmospheric flow and related turbulence parameters and to apportion data from the emission inventories to grid cells. The meteorological fields were produced by the prognostic non-hydrostatic model WRF (Skamarock et al., 2008). To better capture the influence of urban areas on meteorological fields, the so-called Building Environment Parameterization (BEP), a multilayer urban canopy model (Martilli et al., 2002), has been used. Urban parameters were derived from building data available at "Geoportale Nazionale" (http://www.pcn.minambiente.it/mattm/). By carrying out WRF analysis by city, different climate conditions can be taken into account in estimations of air pollution concentrations. The emissions on the modelling domains (see Table 1) have been derived from regional inventories. The INEMAR (http://www.inemar.eu/xwiki/bin/view/Inemar/WebHome) inventories developed for Piedmont, Lombardy, Emilia Romagna and Apulia regions, have been used to reconstruct emissions for the cities of Turin, Milan, Bologna and Bari domains respectively. The emissions for Palermo and Rome domains have been provided by the respective

Table 1
Main characteristics of the six selected cities and data used in this study.

	Rome	Milan	Turin	Bologna	Bari	Palermo
Region	Lazio	Lombardy	Piedmont	Emilia Romagna	Apulia	Sicily
Altitude	21 m	122	239 m	54 m	5 m	14 m
Climate	Mediterranean	Continental	Continental	Continental	Mediterranean	Mediterranean
Metropolitan area population	4352 k	3127 k	1437 k	1015 k	1252 k	1253 k
Total domain emission						
NO_x [t/y]	28,276.8	46,474.3	20,263.3	16,786.9	13,745.8	11,520.7
PM_{10} [t/y]	8416.1	10,701.3	5461.8	3031.7	3958.8	5151.8
PM_{10} [$\mu g/m^3$][b]	31	40	46	29	27	34
	(26)	(97)	(118)	(40)	(14)	(26)
NO_2 [$\mu g/m^3$][c]	62	64	80	46	43	60
	(14)	(11)	(25)	(0)	(0)	(0)
FARM/RF[a] models						
Domain extension [km^2]	59 × 59	59 × 59	51 × 51	50 × 50	60 × 55	75 × 50
Dynamic population data						
Number of grid cells	923	1419	571	n.a	144	165
Domain extension [km^2]	114 × 114	90 × 57	65 × 67	n.a	97 × 72	83 × 82
Cell size (min/max) [km^2]	0.26 × 0.34/16 × 20	0.5 × 0.65/4 × 5	0.26 × 0.34/4 × 5	n.a	1.0 × 1.3/16 × 20	0.5 × 0.65/16 × 20

[a] Flexible Air quality Regional Model (FARM) and Random Forest (RF) models have been applied with a horizontal resolution of 1 km and 200 m respectively. Emission data provided by regional inventories.
[b] Mean observed annual PM_{10}, between parenthesis days with $PM_{10} > 50\ \mu g\ m^{-3}$ (based on 2017 data provided by National Agency for Environment (ISPRA)).
[c] Mean observed annual NO_2, between parenthesis hours with $NO_2 > 200\ \mu g\ m^{-3}$ (based on 2017 data provided by National Agency for Environment (ISPRA)).

regional authorities. A description of emission inventory for Sicily can be found in a related publication (ARPA Sicilia, 2012). As for the domain of Rome, the regional inventory has been integrated with detailed vehicular traffic information provided, for the city of Rome, by the "Roma Mobilità" municipal agency and, for the rest of the region, calculated by means of a traffic model. Details about emission inventories can be found in SM. All the above inventories were developed according to the EMEP/EEA air pollutant emission inventory guidebook that provides guidance on estimating emissions from both anthropogenic and natural emission sources. Pollutant emissions estimates are divided into sectors, representing groups of homogeneous processes, and the sources are classified according to their impact on air quality and the possibility of identification. Diffuse and point emission are considered, the former related to smaller scattered sources for which it is impractical to collect reports from each individual source (e.g. domestic heating, road traffic, agricultural activities, etc.), the latter related to relevant plants such as thermal power plants, cement plants, refineries, etc. The spatial disaggregation of diffuse emissions on the grid cells was based on a statistical procedure, which involves the use of spatial variables (proxy variables) that are assumed to be related both to the type of emissions and to the different territorial units (e.g. municipality). Point emissions were allocated on the grid cells according to their geographic coordinates, geometrical (stack height and diameter) and emissive data. Total NO_x and PM yearly emissions over the six metropolitan modelling domains are shown in Table 1.

Finally, boundary conditions to WRF and FARM models were provided by previous national scale simulations, performed by same models and described elsewhere (Silibello et al., 2019).

The simulations with the above modelling system have been performed for the years 2013, 2014 and 2015 over the 6 urban areas (see Fig. 1 and model domain parameters in Table 1), on an hourly basis, with an horizontal resolution of 1 km. Daily data were then produced to feed the RF model.

2.4. The Random-Forest model

The AQMS modelled daily NO_2, PM_{10}, $PM_{2.5}$ and ozone urban concentrations fields were processed by a Random Forest (RF) model (Breiman, 2001) to derive concentration fields at the target spatial resolution of 200 m. RF models consist of an ensemble of decision trees (forest), suitable both for classification and for regression problems, which has been developed to solve the high variance errors typical of a single decision tree. Each tree is built with a bootstrap of the input data and each node is split by choosing the best subset of randomly chosen predictors (Liaw and Wiener, 2002). The final output of the ensemble is computed by averaging the outputs of each single tree. The application of the RF model was divided into two phases: the first aimed at its formation, i.e. its ability to reproduce concentrations

Table 2
Number of monitoring stations by type of area, type of station, pollutant and year used in the study.

		Rural	Suburban	Urban	Suburban	Urban	Suburban	Urban	Total
		Background			Industrial		Traffic		
NO_2	2013	4	17	23	3	4	3	29	83
	2014	4	17	24	3	4	3	29	84
	2015	4	16	25	3	4	2	31	85
O_3	2013	4	16	17	1	1	2	1	42
	2014	4	16	17	1	1	2	1	42
	2015	4	16	17	1	1	2	0	41
PM_{10}	2013	3	14	22	3	3	2	26	73
	2014	3	14	24	3	3	1	25	73
	2015	3	14	26	3	3	1	23	73
$PM_{2.5}$	2013	2	4	13	1	2	1	10	33
	2014	2	4	16	1	2	1	10	36
	2015	2	4	16	1	2	1	10	36

observed in monitoring sites (training phase) based on a set of predictors. The second (generalization phase) aimed to estimate the concentrations elsewhere and in particular to the cells of the grid where no observations are available. The model also provides an estimate of the "importance" of each predictor by quantifying how much prediction error increases when data for that variable is permuted while all others are left unchanged (Liaw and Wiener, 2002).

The list of temporal, spatial and spatial-temporal predictors, chosen to capture the peculiar time and space fluctuations of the concentration fields, is summarised in Table 3, with each data source and resolution. The target resolution was chosen as a compromise between the spatial size of the census blocks (from 300 m^2 to 5.6 km^2), to which concentrations had to be provided for the epidemiologic studies carried out in the BEEP project, and the spatial resolution of the covariates used as predictors (between 30 and 300 m, see Table 3). The main spatial-temporal predictor is represented by the concentration fields computed by the FARM model at 1 km resolution (see Section 2.3). The Normalized Difference Vegetation Index (NDVI), Julian day, day of week and month were included in the input dataset to predict the temporal variability of the pollutants. Spatial information related to the land-use were provided by CORINE land use dataset and the so-called "Imperviousness Layer" (IL).

NDVI is a spatial-temporal indicator of the greenness of the land surface that is derived from satellite reflectance measurements and was included with a decade long temporal resolution.

The CORINE Land Cover (CLC) inventory was initiated in 1985 (reference year 1990) and updates have been produced in following years (http://land.copernicus.eu/). It consists of an inventory of land cover in 44 classes that we have reduced to a subset of 22 classes.

The IL, a Copernicus land use product, captures the percentage and change of soil sealing. Built-up areas are characterised by the substitution of the original (semi-) natural land cover or water surface with an artificial, often impervious cover.

CLC, NDVI and IL have been mapped from their original resolutions of 100 m, 300 m and 100 m respectively, to the target one (200 m) by means of GIS spatial joins and intersections procedures.

The predictors related to road traffic emissions are based on the daily average traffic volume estimated by the Open Transport Map project (OPM, 2019) on the Open Street Map road network. Traffic data have been aggregated by summing in each target cell the traffic flow contribution of each road arc; this procedure has been applied separately for main (i.e. motorways and primary) and other roads (secondary and beyond). The RF model has been trained against the daily averages of the target pollutants as measured by the local monitoring networks. The training procedure involves all the six urban areas at the same time, to increase the number of the training data representative of different environments and emission sources. It was performed separately for each pollutant and each solar year from 2013 to 2015 to consider year-to-year variability in full account. A total of 12 RF models (4 pollutants × 3 years) were developed. Predictions by pollutant and year were then obtained for all the six urban areas. Analyses were performed using the "caret" and "ranger" R packages for RF model (Kuhn, 2008; Wright and Ziegler, 2017).

Each RF model has been validated with a two stage approach. In the first stage the random forest was checked by out-of-bag predictions comparing predicted to actual values on the training data. This procedure allows to fine tune the model by choosing the set of model parameters which minimizes the Root Mean Square Error. At the end of this stage the relative importance of the predictors, directly provided by the RF model, has also been checked. All above listed predictors were retained in the final model.

In the second stage, each model has been validated by a left-out monitor, using 10-fold cross-validation procedure by randomly dividing the monitoring stations in 10 groups and the training dataset in 10 corresponding subsets. The model was iteratively trained with 9 groups and prediction was performed on the remaining 10th testing group. The model performance was assessed by comparing predicted values with actual measurements in the testing subgroups: a simple linear model relating predicted values with observations has been fitted and the main statistical parameters such as R^2, RMSE, slope and intercept have been computed for the full dataset data (overall analysis), for yearly averages at each monitoring station (spatial analysis) and for time deviations obtained by subtracting the yearly average from daily values (temporal analysis). Many authors already used this validation by spatial and temporal components (eg. Stafoggia et al., 2017; Di et al., 2019) to test model performance. The former represents the fraction of spatial variation in annual average concentrations across the monitoring stations captured by the model, while the latter the fraction of temporal variation in daily concentrations across all monitoring stations and days captured by the model. In addition, an analysis of model's residual by type of monitoring station has been carried out to test its accuracy across a range of landuse types. Results from the application of RF model are described in Section 3.

2.5. The dynamic urban population data

Dynamic population data are related to the Telecom Italian Mobile (TIM) phone operator subscribers and were provided within the TIM

Table 3
Description of the spatiotemporal and spatial variables used as predictors in the RF model.

Variable	Description	Source	Spatial resolution
Temporal			
Julian day, day of week, month	Day/time characteristics		
Spatial-temporal			
NO_2, PM_{10}, $PM_{2.5}$ and O_3	Model estimated pollutants concentration	FARM	1 km
NDVI	Normalized Difference Vegetation Index	Copernicus (2015 data)	300 m
Spatial			
Administrative areas	Regions, Provinces, Municipalities	ISTAT	Polygons
Population	Resident population from census October 2011	ISTAT	Census blocks
Corine land cover	Land cover characteristics	EEA (2012 data)	100 m
Imperviousness surface areas	An indicator of the spatial distribution of artificial areas. Examples include housing areas, traffic areas (airports, harbours, railway yards, parking lots), roads, industrial and commercial areas, construction sites, etc.	Copernicus (2015 data)	100 m
Elevation	European Digital Elevation Model EU-DEM	EEA - CLMS	~30 m
Vehicular traffic	Daily traffic volumes over two roads classes (main and others)	Open Transport Map	Polygons

BIGDATA Challenge 2015 edition. These data represent a relevant sample of the actual population considering that TIM market penetration is about 32% at national level. The population dataset provides the number of persons located in the studied area, at aggregated level, according to full mobile phone communications types (eg. calls, TXT messages, Internet). The methodology used to derive the population data from mobile phone traffic is described in Gariazzo et al. (2016) and Gariazzo and Pelliccioni (2018). The population data, spatially and temporally resolved, refer to the main Italian cities and related Provinces and have been provided over irregular grids covering the cities with higher spatial resolution in downtown zones (from 0.26×0.34 up to 1.0×1.3 km^2) and coarser resolution in outskirts ones (from 4×5 up to 16×20 km^2). Fig. 1 shows an example of these grids. Table 1 provides a summary of the main characteristics of this dataset. Data were not available for the city of Bologna. The data consist of the number of TIM users located within each cell at a time resolution of 15 min, and span from March to April 2015. Such data were averaged on hourly basis.

2.6. The population-weighted exposure

To estimate population-weighted exposure, population data and RF model results have been matched considering the spatial and temporal resolution of dynamic population data as the target ones. Since the RF models spatial resolution (0.2×0.2 km^2) was higher than the population one, the RF and population grids have been intersected and for each target grid cell the area intersecting RF model cells has been computed. The latter was then used as weight of encompassing RF cells to calculate the air pollution concentration at each target cell. Vice versa, since the temporal resolution of RF model was lower than the dynamic population data, daily RF data were modulated on an hourly basis using corresponding hourly CTM results. Following such an approach, RF model air pollutant concentrations were ready to be matched in space (target population grid) and time (hourly basis) with population data. A similar procedure has also been used to estimate population-weighted exposure with concentration levels predicted by FARM.

To obtain a citywide population exposure we used the Population-Weighted Exposure (PWE), a metric already applied by different authors (Nyhan et al., 2016; Aunan et al., 2018; Chen et al., 2018a). The hourly total PWE of pollutant p at time t, $PWE_p(t)$, is given by following equation:

$$PWE_p(t) = \sum_{i=1}^{ncells} C_{p,i}(t) \times WP_i(t)$$

where *ncells* is the number of target cells intersecting the RF model domain; $C_{p,i}(t)$ is the estimated concentration of pollutant p at time t in the target cell i provided by either RF or FARM models; $WP_i(t)$ is a weight calculated as the ratio of the number of persons located in the target cell i to the total number of persons in the domain at time t. The use of this metric enables a comparison of population exposures among cities. It can be noticed that PWEs values, being a concentration multiplied by a weight, have unit of concentration.

To evaluate the amount of population exposed to specific air pollutant concentration values, cumulative population-weighted exposures were calculated by ordering for increasing values of hourly pollutant concentrations on the target grid, and cumulating the percentage of population involved, as already done by other authors (Gariazzo et al., 2016; Nyhan et al., 2016). This analysis, as PWE, allows evaluating the combined effects on exposure of spatial-temporal variability of population mobility and pollutants concentration.

Since the analysis of population-weighted exposure was limited to March–April 2015, due to dynamic population data availability, the pollution levels during this period are different than the annual ones, and the population-weighted exposure estimates are representative of air quality conditions normally occurring during spring seasons.

3. Results

3.1. RF models concentration results

The results of the RF validation procedure are summarised in Table 4. They show a good agreement of the RF predictions with the observations, picking up the overall, spatial and temporal variability. The poorer R^2 results shown for O_3 in the spatial analysis are related to both its ubiquitous nature, more associated with larger scale photochemical processes than with the considered high-resolution spatial predictors, and the location of monitoring stations that, for this pollutant, are generally sited in remote and background areas. The intercept values, associated with model bias, for the overall analysis are generally close to the ideal value of zero, as well as the slope values (ideal value of one). A performance analysis has been done for FARM and the results are presented in SM (table S2), outlining the improvement obtained by the downscaling procedure. At this regard, it should be considered that the adopted FARM resolution (1 km) does not permit to reproduce adequately the concentration observed at monitoring stations highly influenced by nearby traffic emissions.

Fig. 2 shows scatter plots comparing the daily RF model predicted pollutants concentrations with the observed values at the different cities. Similar results for the FARM model are shown in the SM (Fig. S3). Fig. 3 shows boxplots of the residuals of the RF models by the type of monitoring station for the year 2015. A similar by city analysis is shown in SM (Fig. S4).The median values of the model's residuals are close to the ideal value of zero. As for NO_2, the best performances are observed for urban-traffic stations and urban/suburban background ones, with most of data (25th–75th percentiles) in the range of ± 13 μg/m^3. The other types of station (urban background, urban/suburban industrial and suburban traffic) have a small overestimation tendency (10–15 μg/m^3). This result can be caused by the lower number of monitoring stations located in these areas (see Table 2) and consequently a lower representativeness in the final RF model setup. Better results are obtained for PM_{10} and $PM_{2.5}$, with most of data (25th -75th percentiles) in the range of ± 5 μg/m^3 in both urban-traffic and urban/suburban background stations. In addition the RMSE values shown in Table 4 for the overall analysis, are half of those obtained for the FARM model alone (see Table S2 in SM), supporting the better performance of RF for these pollutants.

To provide evidence of the effects of the adopted downscaling procedure, Fig. 4 shows the annual averaged NO_2 concentration fields for the year 2015, produced by FARM alone and RF models for Turin and Rome urban areas. Additional maps are reported in the SM (fig. S5-S7). The analysis of this figure shows, as expected, higher NO_2 levels estimated by RF model due to both the increase of horizontal resolution and the consequent more detailed description of road network and related contributions.

Table S3 in the SM shows the relative importance of predictors in the RF model. Specifically, having set to 100% the importance of FARM predictor, it displays the importance of all other predictors relative to it. The relevance of traffic flow predictors, together with other spatial-temporal predictors, as FARM and NDVI, is confirmed for NO_2. As for the other pollutants, the traffic related predictors are less relevant, while the spatial-temporal predictors keep their importance. Table S4 in the SM shows the ranges of mean yearly pollutants concentrations by city predicted by the RF model during the year 2013–2015. Other statistical parameters (standard deviation and percentiles) are also shown. Other than O_3, all remaining pollutants exhibit a clear geographic gradient with higher concentrations for cities with continental climate (Milan, Turin and Bologna) and lower in cities characterised by Mediterranean climate (Bari, Palermo and Rome). Differences in city specific total emissions also contribute to this effect. As for O_3, higher concentrations are estimated for cities located in the Mediterranean climate conditions due to photochemical processes driven by higher solar radiation. It is worth to notice that the different climate conditions existing among

Table 4
Fitting statistics comparing daily observed and 10-fold CV RF predicted NO_2, PM_{10}, $PM_{2.5}$, O_3 concentrations, by pollutant and year: R^2 (percent of explained variability), root mean squared error (RMSE, μg/m^3), intercept (μg/m^3) and slope (μg/m^3), overall and disaggregated by spatial and temporal components.

	Overall				Spatial				Temporal			
	R^2	RMSE	Inter.	Slope	R^2	RMSE	Inter.	Slope	R^2	RMSE	Inter.	Slope
NO_2												
2013	0.59	13.56	−0.63	1.00	0.62	8.54	0.31	0.97	0.57	10.54	0.00	1.03
2014	0.57	12.89	1.07	0.95	0.63	8.54	1.52	0.94	0.50	9.70	0.00	0.98
2015	0.62	13.48	−1.08	1.02	0.64	8.98	0.21	0.99	0.59	10.15	0.00	1.05
PM_{10}												
2013	0.72	10.12	−2.03	1.05	0.68	3.99	−0.29	1.00	0.72	9.30	0.00	1.06
2014	0.69	9.99	−3.33	1.09	0.57	3.92	−1.40	1.02	0.70	9.24	0.00	1.10
2015	0.76	9.32	−1.52	1.03	0.70	3.71	0.53	0.97	0.77	8.54	0.00	1.05
$PM_{2.5}$												
2013	0.77	8.25	−0.79	1.04	0.75	3.33	3.47	0.83	0.78	7.60	0.00	1.05
2014	0.73	7.17	−0.86	1.02	0.60	3.00	0.94	0.93	0.75	6.53	0.00	1.03
2015	0.78	7.92	−0.34	1.01	0.70	3.11	2.72	0.88	0.80	7.25	0.00	1.03
O_3												
2013	0.74	13.79	0.69	0.99	0.14	7.75	21.99	0.56	0.81	11.31	0.00	1.01
2014	0.71	13.30	2.01	0.97	0.12	7.98	26.55	0.46	0.81	10.07	0.00	1.02
2015	0.79	13.25	0.20	1.00	0.31	7.49	12.01	0.77	0.84	10.96	0.00	1.02

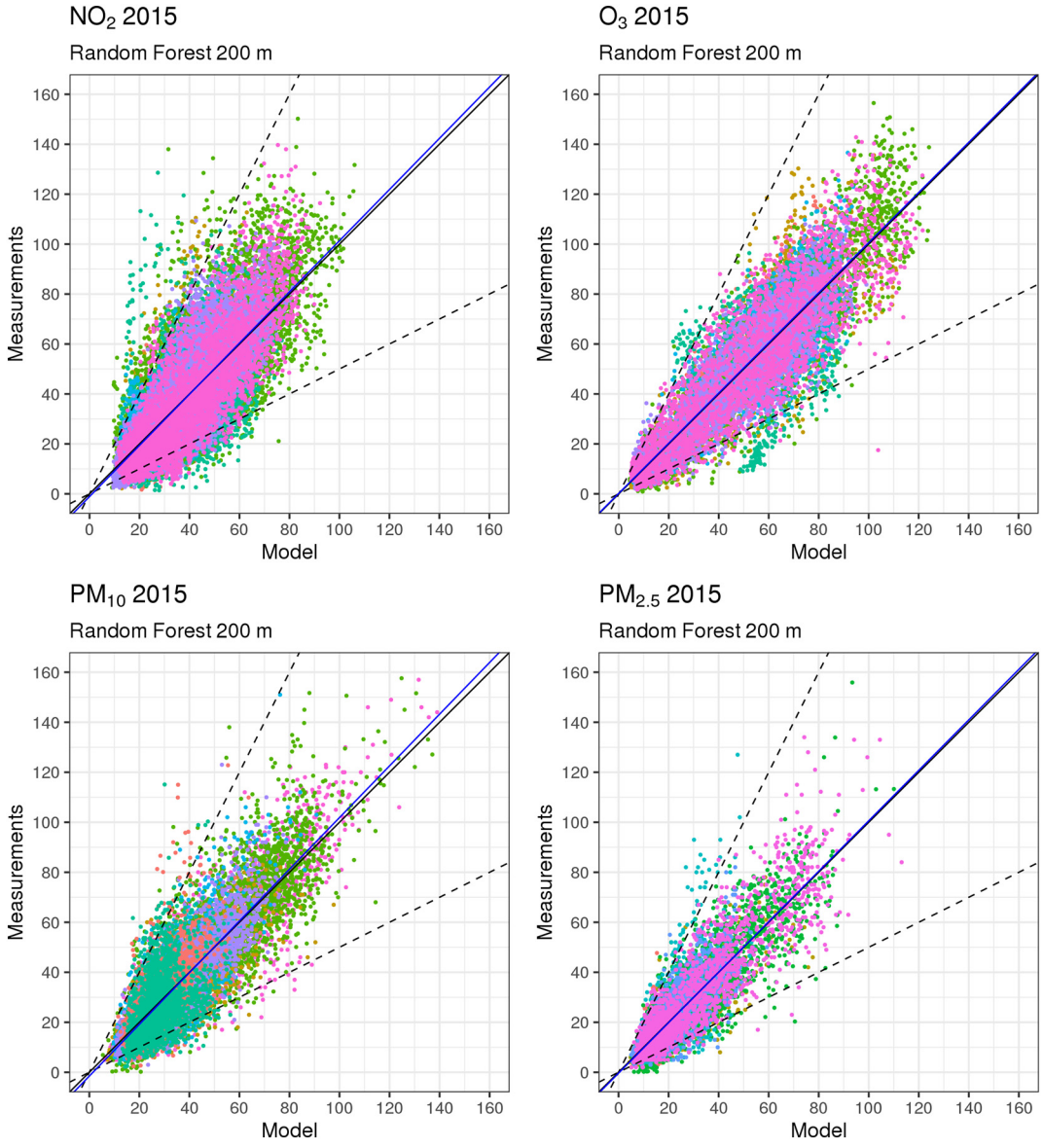


Fig. 2. Scatter plots of daily RF predicted pollutants vs observed concentrations by city. Year 2015. Colours represent the different cities analysed.

cities were taken into account by the WRF analysis, and consequently by both FARM and RF simulations.

3.2. Population exposure results

Fig. 5 shows a box plot of hourly NO_2, O_3, $PM_{2.5}$ and PM_{10} estimations of population-weighted exposure by cities calculated by matching the predictions from the RF model with the population mobility data, as described in Section 2.6. As for NO_2, PWE median values exhibit a clear South-North geographic gradient with the highest exposure in cities located in the Po Valley area (Turin, Milan), followed by the city of Rome. These results are consistent with the amount of emissions in the related areas, which are linked to the total resident population (see Table 1). It is worth noting the boxes size (25th–75th percentiles) that also increase in a similar manner. Extreme values higher than 100 μg/m^3 are observed for Rome and Milan.

A similar South-North geographical gradient is estimated for $PM_{2.5}$ and PM_{10} PWE. People living in more populated cities like Milan, Turin and Rome are affected by higher PWE concentrations as both median value (38, 36 and 25 μg/m^3 for $PM_{2.5}$ respectively) and 25th–75th percentiles interval. $PM_{2.5}$ PWE values higher than 60 μg/m^3 are estimated for Turin and Milan.

Conversely, O_3 PWE exhibits an opposite geographic gradient with higher exposure in cities with Mediterranean climate (Bari, Palermo and Rome) and lower in Turin and Milan characterised by continental climate and higher NO_x levels leading to the well-known ozone titration effect. The higher photochemical activities induced by solar radiation in these Mediterranean cities produces such an effect. The median ozone PWE values span from 50 to 70 μg/m^3, although hourly values higher the 100 μg/m^3 are detected for all cities. Table 5 shows statistics of daily observed pollutants concentrations at the local monitoring networks compared with the RF estimated ambient concentrations and the population weighted exposures calculated using FARM and RF predictions weighted by the amount of population exposed (PWE_FARM and PWE_RF respectively). This comparison allows to verify to what extent the monitoring networks is able to catch the actual population exposure estimated by PWE. At the same time, being the monitoring stations mainly located in populated areas, their observed values can be considered as descriptors of actual population exposure, to be compared with its estimation (PWE). Results show that for O_3, PM_{10} and $PM_{2.5}$, the RF estimated mean PWE (PWE_RF) concentrations are quite close to mean observed concentrations, with some differences in the values of inter quantile range (IQR). As for NO_2, the mean observed and the PWE_RF values are closer for continental cities, like Milan and

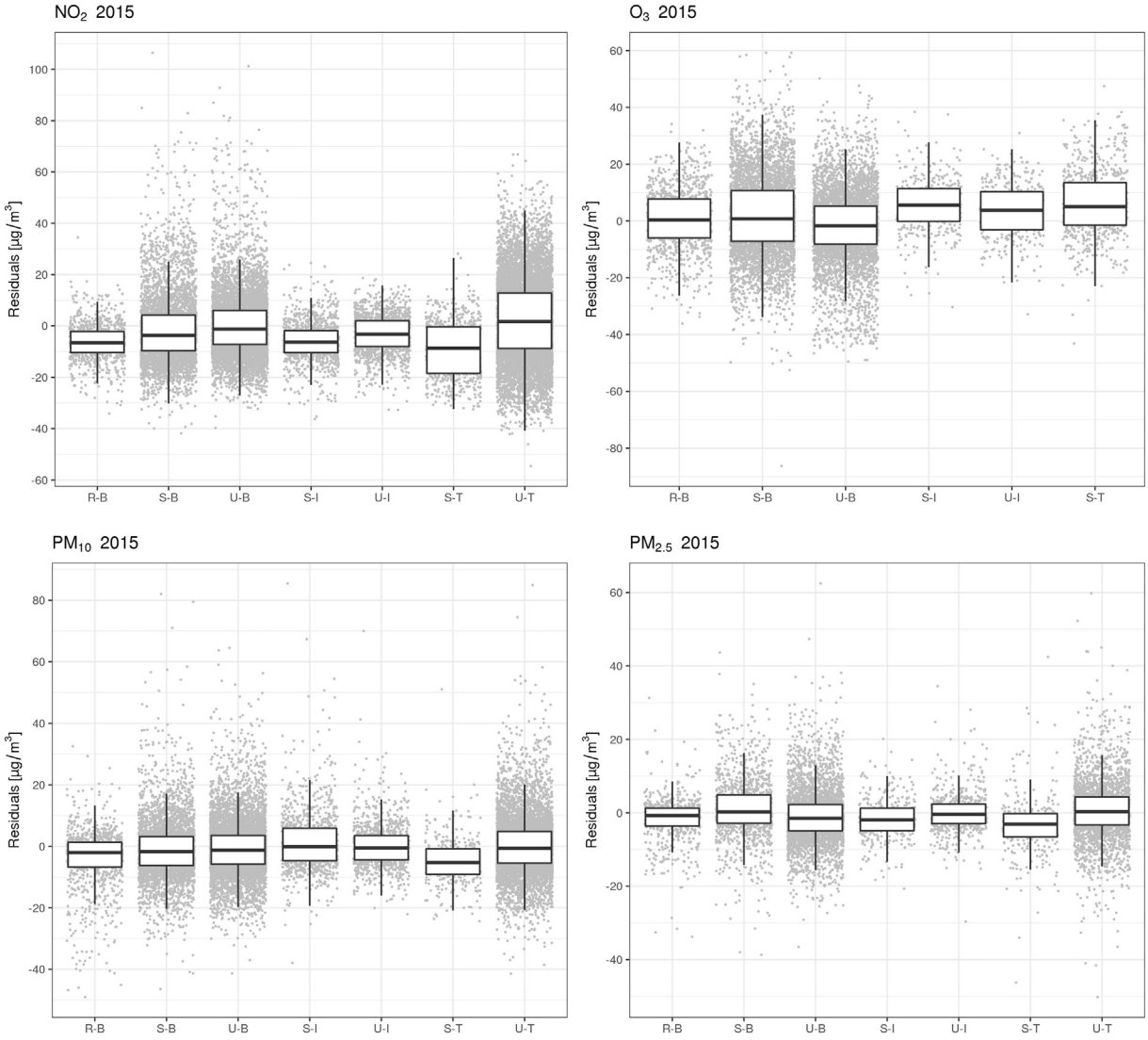


Fig. 3. Boxplots of NO_2, O_3, PM_{10} and $PM_{2.5}$ residuals (as observed minus predicted values) of RF model results by type of monitoring station (R-B Rural Background; S—B Suburban Background; U—B Urban Background; S—I Suburban Industrial; U—I Urban Industrial; S-T Suburban Traffic; U-T Urban Traffic). Grey points represent all data spread around the categorical values. Daily data of year 2015.

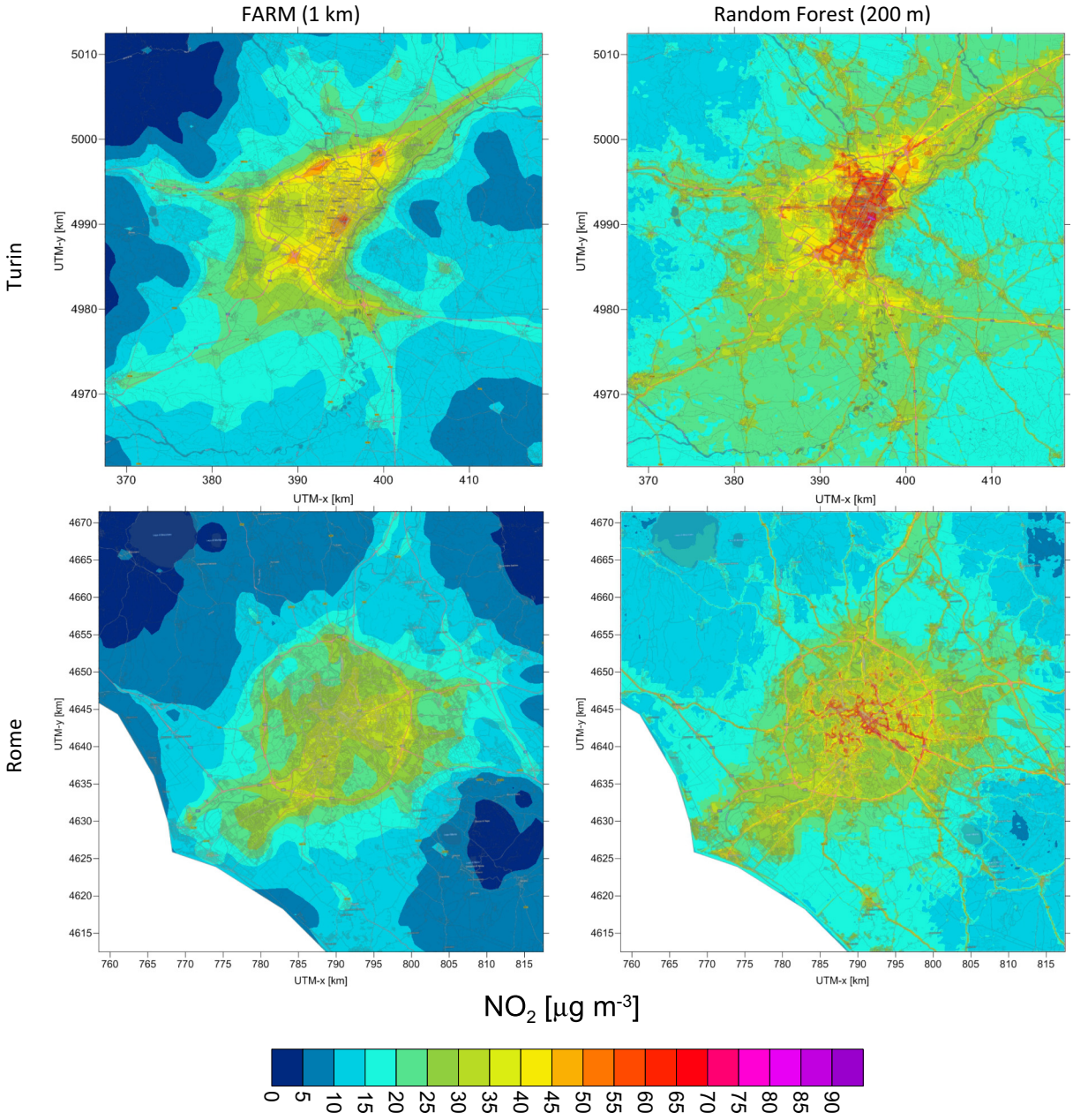


Fig. 4. Comparison of yearly (2015) averaged NO_2 concentrations, for Turin (upper figure) and Rome (bottom figure), computed by FARM (left) and Random Forest models (right).

Turin, than for Mediterranean cities (eg. Rome, Bari and Palermo). This might be attributed to different factors such as local meteorology, location of monitoring stations and model's ability to represent the actual spatial patterns of pollutants concentrations. It can be noticed that PWE_RF results are in general closer to observed values than RF estimated ambient concentrations, particularly for NO_2, the pollutant with the higher spatial contrast. This means that the use of population mobility data, allows to better describe the actual population exposure represented by the concentrations measured at the monitoring stations. Significant differences are detected between PWE results calculated using RF data (PWF_RF) and those obtained with FARM one (PWE_FARM), particularly for NO_2 and PM. The use of RF model estimated concentrations significantly increases the agreement with observed values, out of Ozone results in which computed PWEs from RF and FARM are quite closer. These results support the use of RF model for population exposure studies.

To evaluate the effect of population mobility on PWE, we compared the RF based PWE obtained using the dynamic population data with those derived by static population data based on place of residence (census 2011). Results are shown in SM (Fig. S8) as box plots. No large differences are detected between the two approaches. However, it should be considered that PWE is an urban spatially averaged variable and consequently hotspots exposure might be smoothed during the spatial averaging.

To evaluate the amount of population exposed to a specific pollution concentration value, hourly cumulative population exposures were calculated according to the method described in Section 2.6.

Fig. 6 shows the results for NO_2 and PM_{10} by city. The presented pollutants concentration values are not weighted for population as PWE, but just ordered for increasing concentrations and related to cumulated involved population. Overall, the geographic gradients already seen for PWE are confirmed. The daily cycle of pollutants concentrations is clearly detected in these figures (yellow bands followed by red ones in Fig. 6). For most of the time a large part of population is estimated to be exposed to concentrations lower than 50 and 40 µg/m^3 for NO_2 and PM_{10} respectively (yellow bands in Fig. 6) regardless of the city. Some

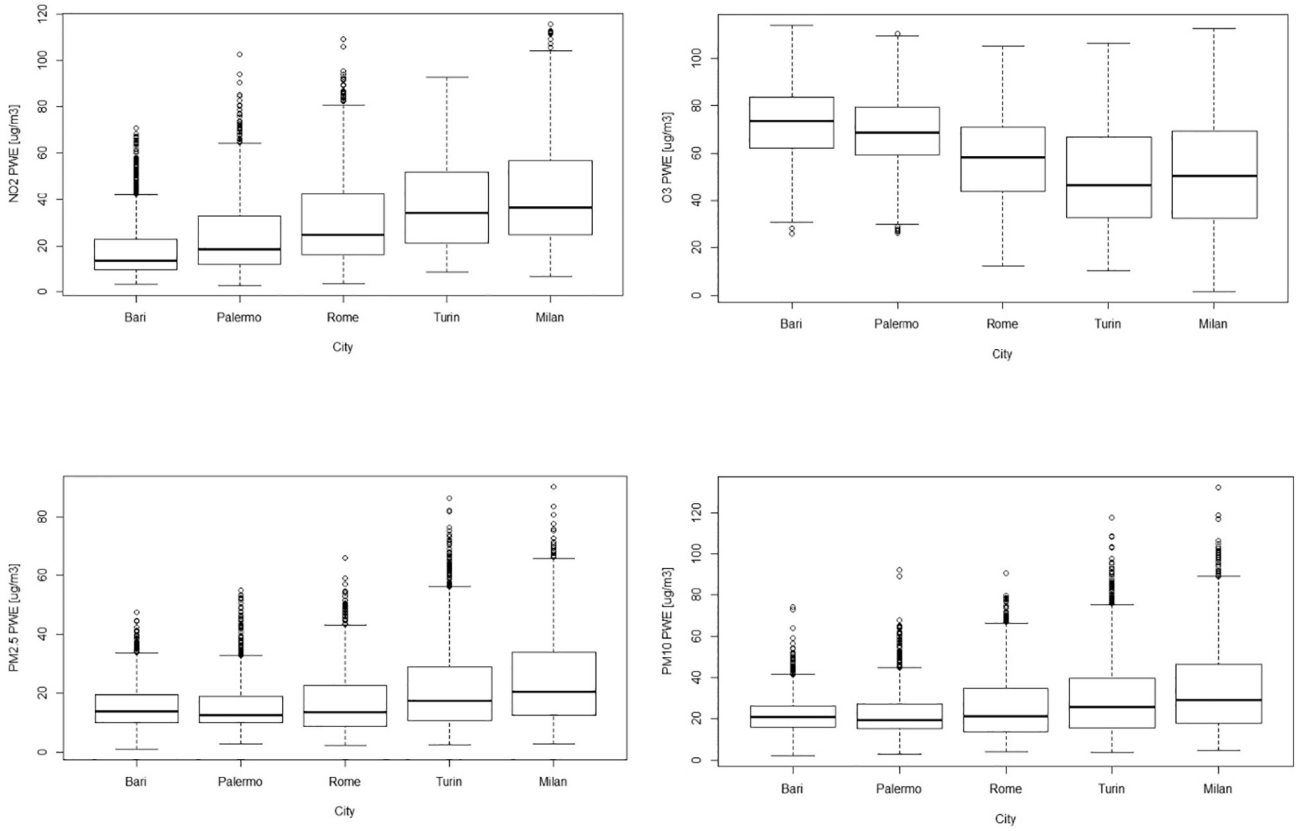


Fig. 5. Boxplots of population weighted exposure calculated using RF model results and dynamic population data derived by mobile phone traffic, by pollutants and city (NO_2 top left; O_3 top right; $PM_{2.5}$ bottom left; PM_{10} bottom right). Hourly data from March to April 2015. PWE expressed in terms of $\mu g/m^3$.

Table 5
Statistics of observed pollutants concentrations, RF model estimated ambient concentrations, PWEs calculated using either FARM (PWE_FARM) or RF (PWE_RF) by city ($\mu g/m^3$). Daily data from March to April 2015.

	NO_2			O_3			$PM_{2.5}$			PM_{10}		
	Mean	Median	IQR	Mean	Median	IQR	Mean	Median	IQR	Mean	Median	IQR
Milan												
Observed	43.93	41.40	29.50	50.58	48.90	28.35	26.40	24.30	18.50	36.32	33.60	24.23
RF	31.19	29.86	16.31	53.20	54.94	24.34	23.07	19.72	15.63	33.01	28.08	21.21
PWE_FARM	32.64	30.43	15.97	52.67	57.72	27.33	15.96	12.16	8.40	18.49	14.31	9.88
PWE_RF	42.40	41.29	14.24	50.41	54.32	20.91	24.79	21.94	13.30	34.99	30.15	17.12
Turin												
Observed	39.62	36.70	29.98	50.20	48.80	22.70	23.72	18.00	20.00	32.42	27.00	26.00
RF	23.18	20.53	11.44	53.11	54.29	13.39	19.52	14.65	12.03	25.93	20.96	16.65
PWE_FARM	26.16	26.34	8.84	51.70	51.77	17.98	12.81	10.06	6.82	15.31	12.48	7.61
PWE_RF	38.25	38.02	10.11	49.31	50.26	13.98	22.07	17.52	10.58	30.65	24.70	13.05
Rome												
Observed	40.17	39.30	28.10	54.65	54.00	23.50	16.71	15.50	10.00	27.34	27.00	14.00
RF	17.65	14.30	11.15	65.26	65.89	12.45	14.57	13.27	8.03	21.47	21.33	8.24
PWE_FARM	21.77	22.42	12.01	61.71	63.18	19.52	13.98	12.76	9.06	16.82	16.35	10.36
PWE_RF	31.53	31.93	9.49	56.98	59.26	13.64	16.75	16.24	7.41	25.95	26.97	7.13
Bari												
Observed	26.12	22.10	14.55	72.77	72.80	24.80	17.29	16.70	10.80	25.74	23.20	12.33
RF	13.04	11.38	3.55	74.11	74.05	10.28	14.71	12.82	8.08	20.99	20.52	5.48
PWE_FARM	9.03	8.43	5.10	72.47	71.96	14.49	7.68	6.35	4.45	9.21	8.42	4.43
PWE_RF	18.33	17.93	3.78	73.05	75.76	13.31	15.70	13.98	8.68	22.23	21.38	5.65
Palermo												
Observed	42.15	41.15	27.33	64.73	70.00	33.35	–	–	–	27.42	25.20	16.65
RF	12.13	10.86	2.98	73.41	73.37	8.70	11.94	11.11	3.06	18.67	18.52	4.08
PWE_FARM	14.44	13.68	9.14	71.51	71.52	11.71	9.22	8.88	5.80	11.98	11.26	4.42
PWE_RF	24.60	24.04	5.58	68.83	69.43	12.47	15.93	15.88	5.44	23.10	23.39	3.30

pollution episodes are identified, particularly for PM_{10} and for the cities of Turin, Milan and Rome (red hot spots bands in Fig. 6). During them, the whole population seems to be affected by PM_{10} concentrations higher than 80 μg/m^3 (red bands spanning up to values lower than 20% of population). Fig. 6 also shows the range of NO_2 and PM_{10} hourly concentrations to which at least 50% of population is exposed. Differences among the cities are observed, with 50% of population of Milan experiencing the highest concentrations (25–55 μg/m^3 and 20–50 μg/m^3 for NO_2 and PM_{10} respectively) and Bari the lowest ones (10–20 μg/m^3 and 15–25 μg/m^3 for NO_2 and PM_{10} respectively). NO_2 peak values higher than 100 μg/m^3 are also predicted for some cities.

Mean values of cumulative population exposed vs NO_2 and PM_{10} concentrations are shown in SM (Fig. S9). The hourly cumulative population exposed vs O_3 and $PM_{2.5}$ concentrations are also shown in SM (Fig. S10). As for ozone, results indicate that about 75% of population is exposed to concentrations up to 60 μg/m^3. Peak values up to 100 μg/m^3 may sometimes affect the remaining part of population.

4. Discussion and conclusions

The accuracy of the estimates of atmospheric pollution concentrations is relevant for the assessment of health risk, in particular in urban areas where errors of prediction and incorrect classification can occur for pollutants that are spatially uneven.

Different methods were used, based on different approaches, such as ambient measurements, regression analysis and deterministic numerical models, achieving different level of accuracy (Hoek, 2017; Hoek et al., 2008; Cesaroni et al., 2012; Zhang et al., 2012; Kukkonen et al., 2012, 2016; Gariazzo et al., 2007; Parvez and Wagstrom, 2019). Recently, statistical machine learning methods were used to predict particulate matter at national or continental wide scales (Chen et al., 2018a, 2018b; Stafoggia et al., 2019). Other authors (de Hoogh et al., 2019) applied the same methods to predict NO_2 at high resolution across Switzerland using OMI satellite data. There are very few studies (eg. Araki et al., 2018) dealing with machine learning methods to estimate pollutant concentrations in urban areas with enough accuracy. The availability of good quality predictors able to describe the relevant processes occurring in such areas, and the number of enough monitoring stations capable to provide information on the actual spatial heterogeneity of pollutants concentration, limited the use of these methods.

In the present study, we applied a random forest machine learning method to predict daily urban concentrations of the main pollutants (NO_2, O_3, PM_{10} and $PM_{2.5}$) at high spatial resolution (200 m) in six Italian metropolitan areas, chosen for their different climate conditions. The former machine learning applications used AOD/OMI satellite and other spatial-temporal data to predict nation-wide PM and NO_2 concentrations (Chen et al., 2018a, 2018b; Stafoggia et al., 2019; de Hoogh et al., 2019). We extended for the first time the application to O_3 and applied the method at urban scale using a novel approach in which relevant urban scale phenomena (eg. emission, dispersion, transformation and deposition) were simulated by a CTM (FARM) run at 1 km horizontal resolution. FARM computed concentrations were used as a predictor for the RF model. To consider sub-grid effects not considered by FARM simulations (eg. road structure and street effects), other spatial predictors (eg. population, daily traffic flows on two road classes, etc.) available at higher spatial resolution were considered. This combined approach allowed to improve the spatial resolution (200 m) of predicted urban pollutants concentrations. In addition, the RF training procedure by involving all the six urban areas at the same time, increased the number of the training data (observed values) with respect to a single city approach. As stated by Harrison (2018), this improvement cannot be achieved by simply increasing the CTM spatial resolution since many relevant processes differ from rural to street-level scale (eg. heat fluxes, turbulence and mixing effects, emission resolution, chemical and street processes).

We were able to capture 57–62%, 72–76%, 73–78% and 71–79% of the overall variability in left-out monitors of NO_2, PM_{10}, $PM_{2.5}$ and O_3 respectively without large differences among the three years. Improvements in model performance were also observed with respect to FARM results when used without combination with RF. Good performance was achieved in predicting day-to-day variability as well as spatial contrasts in annual averages (except for ozone), justifying the use of predictions for the analysis of short-term and long-term health effects. Small biases and underestimations of daily observations were also detected in estimated concentrations, particularly for NO_2 at urban-traffic stations, addressing for further improvements at this spatial resolution, if better predictors will be identified and made available. An analysis of model residuals by different type of monitoring stations demonstrated that the model is accurate across a range of land-use types, particularly for largely represented type of stations like urban-traffic stations and urban/suburban background ones. These results are comparable with LUR results achieved for the city of Rome in former studies. Cesaroni et al. (2012) obtained a R^2 of 0.61 and RMSE of 5.38 when urban NO_2 exposure was estimated. Recently, Cattani et al. (2017) achieved an R^2 of 0.64 in a leave one out cross-validation, in estimating ultrafine particles in the same city. Weissert et al. (2018) developed a micro-scale LUR model to estimate NO_2 in a portion of the city of Auckland (NZ), obtaining a R^2 of 0.66 and RMSE of 3.88 μg/m^3.

According to a review on LUR model studies (Hoek et al., 2008), for NO_2 the percentage of explained variation from the prediction models is typically about 60–70%, depending on the variability in the measured concentrations, quality of the predictor variables, the modelling approach and the complexity of the city. According to the same review, the performance on variability of $PM_{2.5}$ contrasts between 0.17 and 0.82 depending from the limited variability of measured $PM_{2.5}$ concentrations in the small study areas. However, it should be considered that LUR applications are part of dedicated studies, like the ESCAPE (Cesaroni et al., 2014), in which proper time limited field campaigns are carried out to collect data to be used for model reconstruction of pollutant concentrations. The present study achieved the same performance using conventional monitoring station data used by the machine learning model as target values. Multi-year studies can be performed providing concentrations data for short- and long-term health effects studies, overcoming the lack of temporal variability of LURs models.

Another advancement of this study is the multi-pollutants approach. The availability of gaseous and aerosol pollutants concentrations provided by the CTM at urban scale (1 km), gave the possibility to use them as predictors for pollutant specific random forest models properly customized to reproduce observed concentrations. To our knowledge, there are no other studies approaching a high-resolution multi-pollutants exposure analysis at urban scale using a combined CTM/RF approach. Ozone exposure at urban scale was never estimated using LUR or RF models, although CTM study provided such data but at lower spatial resolution (Gariazzo et al., 2007). In addition, the multi-city approach allowed training the random forest models using more representative data, getting the different peculiarities among monitoring stations, as location, type (eg. rural, urban, suburban, and industrial) and climate conditions. Finally, the multi-city method provided pollutants maps that are used for comparison.

As far as the number and quality of spatial-temporal parameters used in RF models are concerned, the FARM model results were the most important predictors. This was expected as it embedded all relevant processes involved in the determination of pollutants concentrations. The FARM model was found to underestimate the observed concentrations, particularly for $PM_{2.5}$ and PM_{10}. In fact, it is well known that aerosol models are unable to reproduce correctly PM size fractions due to the limited number of aerosol processes included in the models, particularly the secondary process driven by organic compounds (Kukkonen et al., 2012). RF model corrected this underestimation. The daily traffic volumes data provided by Open Transport Map were also one of the best predictors, particularly for spatially

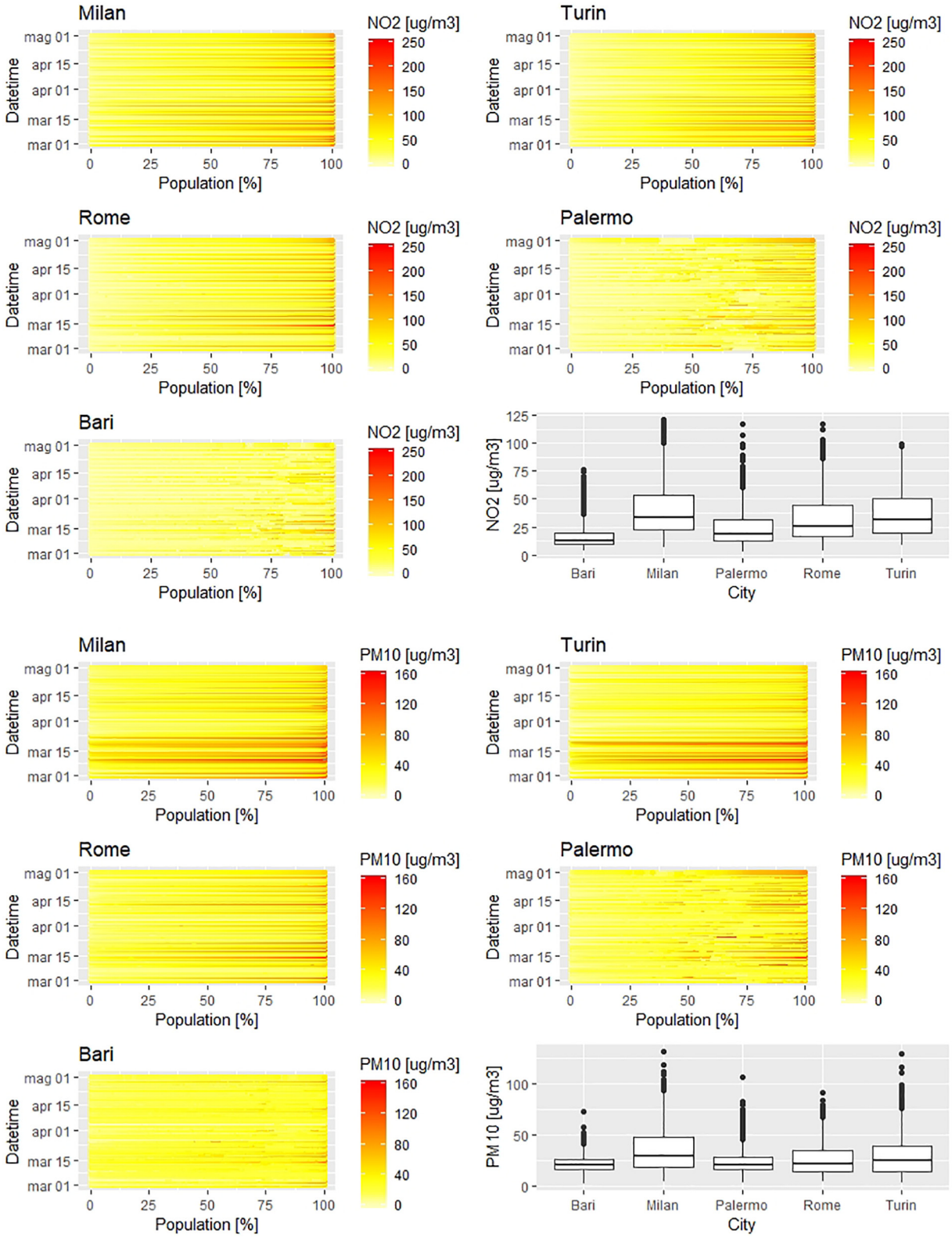


Fig. 6. Hourly cumulative percentage of population exposed vs RF NO_2 (upper figure) and PM_{10} (bottom figure) concentrations by city based on dynamic population data derived by mobile phone traffic. Lower right boxplots refer to hourly concentrations to which at least 50% of population is exposed. Data from March to April 2015.

inhomogeneous pollutants as NO_2. They were effective in adding spatial details to the predicted maps, particularly for NO_2. When traffic volumes data were verified with a direct authors' knowledge in selected roads, they were found not highly accurate in estimating the actual values. However, the relative differences in volumes among roads were described quite well, addressing to their use as a proxy of traffic volumes to add spatial contrast in the final maps. Other spatial parameters were less important as predictors of pollutants concentrations. The remaining amount of variability not explained by the model addresses for use of additional spatial parameters able to describe small-scale

urban effects such as hotspots, busy intersection roads and canyon effects. Proxy of such effects are hard to be obtained for large metropolitan areas and their use remain challenging.

Cities located in the Po valley were estimated to be highly exposed to both PM and NO_2. This is consistent with a large body of literature (Pernigotti et al., 2012; Bigi and Ghermandi, 2014; Perrino et al., 2014). Ozone was found to increase in cities located in the Mediterranean area and this result was expected due to the higher photochemical activity.

The availability of dynamic population mobility data, derived by mobile phone traffic, in five of six studied cities, allowed to obtain dynamic population weighted exposure by coupling, in space and time, population mobility with RF model estimated pollutants concentration. With respect to the former study conducted for the city of Rome by Gariazzo et al. (2016), the current study extends the analysis to other important Italian metropolitan areas having different geographical location and climate, and improves the accuracy of exposure estimations by using more accurate and detailed air quality data. Recent studies have incorporated mobility of population for dynamic exposure assessment (Gariazzo et al., 2016; Nyhan et al., 2016; Dewulf et al., 2016; Chen et al., 2018a, 2018b; Yu et al., 2018; Picornell et al., 2019). Nevertheless, the short-term availability of mobility data and the lack of information on individual mobility patterns, limited their application to epidemiological studies.

This study found that, although on average the population was exposed to values of pollutants concentrations lower than limit values (even though people were exposed to values above the WHO guidelines published in the year 2005), for some cities, short time-periods episodes were identified with very high concentrations where the whole population is involved. The comparison of PWEs with observed pollutants concentrations, results in the ability of the local monitoring networks to represent the population exposure, particularly for urban diffuse pollutants like $PM_{2.5}$, PM_{10} and O_3, but in a lesser extent for spatial inhomogeneous pollutants as NO_2. With respect to RF estimated ambient concentrations, the use of population mobility data in evaluating exposure allows to better describe the actual population exposure represented by the concentrations measured at the monitoring stations. Significant differences are detected between PWE results calculated using RF (PWF_RF) and FARM data (PWE_FARM), outlining a better agreement with observed NO_2 and PM concentrations using the former. This result can be ascribed to the better estimations of ambient concentrations obtained by RF approach. The hourly cumulative population exposure results allow to assess the amount of population exposed to specific concentration values. A large part of population was found to be exposed to concentrations lower than 50 and 40 $\mu g/m^3$ for NO_2 and PM_{10} respectively regardless of the city. In particular, the 50% of population is exposed to NO_2 concentrations between 12 and 38 $\mu g/m^3$ and between 20 and 35 $\mu g/m^3$ for PM_{10}, as median value, depending on the city. Pollution episodes were identified in which the whole population is exposed to PM_{10} concentrations higher than 80 $\mu g/m^3$. As population data were provided in aggregate form, it was not possible to track individual mobility and assess individual exposure. Availability of such data, provided the protection of privacy with proper anonymization procedures, would be very useful for population exposure assessment and health effects studies.

In conclusion, the adopted approach combining CTM and ML models improved CTM's estimations adding spatial details in the final air quality maps. Machine learning methods in combination with multiple parameters data, can be a valid tool for predicting level of air pollutants concentrations at fine spatial and temporal resolution. Further details on population exposure were provided by coupling RF model results with dynamic population data.

However, more research is needed to catch urban effects not still resolved, like hotspots and accumulation of pollutants in urban canyon. New parameters and proxy data, able to describe such urban effects should be identified and collected.

Finally, the air pollutants predictions made available from this study will provide novel evidence on the short-term and long-term health effects in main Italian metropolitan areas.

CRediT authorship contribution statement

Claudio Gariazzo: Conceptualization, Resources, Methodology, Writing - original draft. **Giuseppe Carlino:** Software, Formal analysis, Validation, Writing - original draft. **Camillo Silibello:** Software, Formal analysis, Validation, Writing - original draft. **Matteo Renzi:** Resources, Software. **Sandro Finardi:** Resources, Software, Formal analysis. **Nicola Pepe:** Resources, Software, Formal analysis. **Paola Radice:** Resources, Software, Formal analysis. **Francesco Forastiere:** Writing - review & editing. **Paola Michelozzi:** Writing - review & editing. **Giovanni Viegi:** Project administration, Writing - review & editing. **Massimo Stafoggia:** Conceptualization, Methodology, Supervision, Writing - review & editing.

Acknowledgements

The TIM Big Data Challenge 2015 is acknowledged for the provision of the mobile phone derived population data.

Funding

This work has been partially funded by the Italian Workers' Compensation Authority (INAIL), Italy, within the project "BEEP" (project code B72F17000180005).

Declaration of competing interest

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

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.scitotenv.2020.138102.

References

Araki, S., Shima, M., Yamamoto, K., 2018. Spatiotemporal land use random forest model for estimating metropolitan NO2 exposure in Japan. Sci. Total Environ. 634 (2018), 1269–1277. https://doi.org/10.1016/j.scitotenv.2018.03.324.

Arpa Sicilia, 2012. L'inventario delle emissioni in atmosfera della regione Sicilia. http://www.arpa.sicilia.it/wp-content/uploads/2015/08/Relazione-Inventario-Emissioni.pdf (in Italian). Accessed on March 13th 2020.

Aunan, K., Ma, Q., Lund, M.T., Wang, S., 2018. Population-weighted exposure to $PM_{2.5}$ pollution in China: an integrated approach. Environ. Int. 120, 111–120.

Baccini, M., Mattei, A., Mealli, F., Bertazzi, P.A., Carugno, M., 2017. Assessing the short-term impact of air pollution on mortality: a matching approach. Environ. Health 16, 7 (2017). https://doi.org/10.1186/s12940-017-0215-7.

Bigi, A., Ghermandi, G., 2014. Long-term trend and variability of atmospheric PM_{10} concentration in the Po Valley. Atmos. Chem. Phys. 14 (10), 4895–4907 (2014).

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

Brook, R.D., Newby, D.E., Rajagopalan, S., 2017. Air pollution and cardiometabolic disease: an update and call for clinical trials. Am. J. Hypertens., 1–10 https://doi.org/10.1093/ajh/hpx109.

Cattani, G., Gaeta, A., Di Menno di Bucchianico, A., De Santis, A., Gaddi, R., Cusano, M., Ancona, C., Badaloni, C., Forastiere, F., Gariazzo, C., Sozzi, R., Inglessis, M., Silibello, C., Salvatori, E., Manes, F., Cesaroni, G., 2017. Development of land-use regression models for exposure assessment to ultrafine particles in Rome. Italy. Atmos. Environ. 156, 52–60. https://doi.org/10.1016/j.atmosenv.2017.02.028.

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. Environ. Health 11, 48–69–11–48.

Cesaroni, G., Forastiere, F., Stafoggia, M., Andersen, Z.J., Badaloni, C., Beelen, R., Caracciolo, B., de Faire, U., Erbel, R., Eriksen, K.T., Fratiglioni, L., Galassi, C., Hampel, R., Heier, M., Hennig, F., Hilding, A., Hoffmann, B., Houthuijs, D., Jöckel, K.-H., Korek, M., Lanki, T., Leander, K., Magnusson, P.K.E., Migliore, E., Ostenson, C.-G., Overvad, K., Pedersen, N.L., J, J.P., Penell, J., Pershagen, G., Pyko, A., Raaschou-Nielsen, O., Ranzi, A., Ricceri,

F., Sacerdote, C., Salomaa, V., Swart, W., Turunen, A.W., Vineis, P., Weinmayr, G., Wolf, K., de Hoogh, K., Hoek, G., Brunekreef, B., Peters, A., 2014. Long term exposure to ambient air pollution and incidence of acute coronary events: prospective cohort study and meta-analysis in 11 European cohorts from the ESCAPE project. BMJ 348, f7412. https://doi.org/10.1136/bmj.f7412.

Chen, H., Kwong, J.C., Copes, R., Hystad, P., van Donkelaar, A., Tu, K., Brook, J.R., Goldberg, M.S., Martin, R.V., Murray, B.J., Wilton, A.S., Kopp, A., Burnett, R.T., 2017. Exposure to ambient air pollution and the incidence of dementia: a population-based cohort study. Environ. Int. 108, 271–277. https://doi.org/10.1016/j.envint.2017.08.020.

Chen, Bin, Song, Yimeng, Jiang, Tingting, Chen, Ziyue, Huang, Bo, Xu, Bing, 2018a. Real-time estimation of population exposure to $PM_{2.5}$ using mobile- and station-based big data. Int. J. Environ. Res. Public Health 15, 573. https://doi.org/10.3390/ijerph15040573.

Chen, B., Song, Y., Kwan, M.P., Huang, B., Xu, B., 2018b. How do people in different places experience different levels of air pollution? Using worldwide Chinese as a lens. Environ. Pollut. 238 (2018), 874–883. https://doi.org/10.1016/j.envpol.2018.03.093.

Chen, G., Li, S., Knibbs, L.D., et al., 2018a. A machine learning method to estimate PM2.5 concentrations across China with remote sensing, meteorological and land use information. Sci. Total Environ. 636, 52–60.

Chen, G., Wang, Y., Li, S., et al., 2018b. Spatiotemporal patterns of PM10 concentrations over China during 2005–2016: a satellite-based estimation using the random forests approach. Environ. Pollut. 242, 605–613.

Chiusolo, M., Cadum, E., Stafoggia, M., Galassi, C., Berti, G., Faustini, A., Bisanti, L., Vigotti, M.A., Dessì, M.P., Cernigliaro, A., Mallone, S., Pacelli, B., Minerba, S., Simonato, L., Forastiere, F., on behalf of the EpiAir Collaborative Group, 2011. Short-term effects of nitrogen dioxide on mortality and susceptibility factors in 10 Italian cities: the EpiAir study. Environ. Health Perspect. 119, 9 CID. https://doi.org/10.1289/ehp.1002904.

Cohen, A.J., Brauer, M., Burnett, R., 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. Lancet 389, 1907–1918. https://doi.org/10.1016/S0140-6736(17)30505-6.

de Hoogh, K., Saucy, A., Shtein, A., Schwartz, J., West, E.A., Strassmann, A., Puhan, M., Röösli, M., Stafoggia, M., Kloog, Itai, 2019. Predicting fine-scale daily NO2 for 2005–2016 incorporating OMI satellite data across Switzerland. Environmental Science & Technology 53 (17), 10279–10287. https://doi.org/10.1021/acs.est.9b03107 2019.

Dewulf, B., Neutens, T., Lefebvre, W., Seynaeve, G., Vanpoucke, C., Beckx, C., Van de Weghe, N., 2016. Dynamic assessment of exposure to air pollution using mobile phone data. Int. J. Health Geogr. 15 (2016), 14. https://doi.org/10.1186/s12942-016-0042-z.

Di, Q., Amini, H., Shi, L., Kloog, I., Silvern, R., Kelly, J., Sabath, M.B., Choirat, C., Koutrakis, P., Lyapustin, A., Wang, Y., Mickley, L.J., Schwartz, J., 2019. An ensemble-based model of PM2.5 concentration across the contiguous United States with high spatiotemporal resolution. Environ. Int. 130, 104909. https://doi.org/10.1016/j.envint.2019.104909.

Forehead, H., Huynh, N., 2018. Review of modelling air pollution from traffic at street-level - the state of the science. Environ. Pollut. 241, 775–786. https://doi.org/10.1016/j.envpol.2018.06.019.

Gariazzo, C., Pelliccioni, A., 2018. A multi-city urban population mobility study using mobile phone traffic data. Appl. Spatial Analysis https://doi.org/10.1007/s12061-018-9268-4.

Gariazzo, C., Silibello, C., Finardi, S., Radice, P., Piersanti, A., Calori, G., Cecinato, A., Perrino, C., Nussio, F., Cagnoli, M., Pelliccioni, A., Gobbi, G.P., Di Filippo, P., 2007. A gas/aerosol air pollutants study over the urban area of Rome using a comprehensive chemical transport model. Atmos. Environ. 41, 7286–7303. https://doi.org/10.1016/j.atmosenv.2007.05.018.

Gariazzo, C., Pelliccioni, A., Bolignano, A., 2016. A dynamic urban air pollution population exposure assessment study using model and population density data derived by mobile phone traffic. Atmos. Environ. 131 (2016), 289–300. https://doi.org/10.1016/j.atmosenv.2016.02.011.

Harrison, R.M., 2018. Urban atmospheric chemistry: a very special case for study. npj Climate and Atmospheric Science 1, 20175. https://doi.org/10.1038/s41612-017-0010-8.

Health Effects Institute, 2009. Traffic-Related Air Pollution: A Critical Review of the Literature on Emissions, Exposure, and Health Effects. Special Report #17, 2009-05-04.

Hoek, G., 2017. Methods for assessing long-term exposures to outdoor air pollutants. Curr. Envir. Health Rpt. Topical Collection on Air Pollution and Health. Springer. https://doi.org/10.1007/s40572-017-0169-5.

Hoek, G., Beelen, R., de Hoogh, K., Vienneau, D., Gulliver, J., Fischer, P., Briggs, D., 2008. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmos. Environ. 42 (33), 7561–7578. https://doi.org/10.1016/j.atmosenv.2008.05.057.

Kuhn, M., 2008. Building predictive models in R using the caret package. J. Stat. Softw. 28 (5), 1–26. https://doi.org/10.18637/jss.v028.i05.

Kukkonen, J., Olsson, T., Schultz, D.M., Baklanov, A., Klein, T., Miranda, A.I., et al., 2012. A review of operational, regional-scale, chemical weather forecasting models in Europe. Atmos. Chem. Phys. 12 (1), 1–87. https://doi.org/10.5194/acp-12-1-2012.

Kukkonen, J., Karl, M., Keuken, M.P., Denier Van Der Gon, H.A.C., Denby, B.R., Singh, V., et al., 2016. Modelling the dispersion of particle numbers in five European cities. Geoscientific Model Dev 9 (2), 451–478.

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

Martilli, A., Clappier, A., Rotach, M.W., 2002. An urban surface exchange parameterisation for Mesoscale models. Bound.-Layer Meteorol. 104 (2), 261–304. https://doi.org/10.1023/A:1016099921195.

Murray, N.L., Holmes, H.A., Liu, Y., Chang, H., H., H., 2019. A Bayesian ensemble approach to combine $PM_{2.5}$ estimates from statistical models using satellite imagery and numerical model simulation. Environ. Res. 178, 108601. https://doi.org/10.1016/j.envres.2019.108601.

Nyhan, M., Grauwin, S., Britter, R., Misstear, B., McNabola, A., Laden, F., Barrett, S.R.H., Ratti, C., 2016. "Exposure track"—the impact of mobile-device-based mobility patterns on quantifying population exposure to air pollution. Environmental Science & Technology 50 (17), 9671–9681. https://doi.org/10.1021/acs.est.6b02385.

Open Transport MAP (OPM), 2019. http://opentransportmap.info/ accessed on October 2019.

Ostro, B., Spadaro, J.V., Gumy, S., Mudu, P., Awe, Y., Forastiere, F., Peters, A., 2018. Assessing the recent estimates of the global burden of disease for ambient air pollution: methodological changes and implications for low- and middle income countries. Environ. Res. 166 (2018), 713–725. https://doi.org/10.1016/j.envres.2018.03.001.

Özkaynak, H., Baxter, L.K., Dionisio, K.L., Burke, J., 2013. Air pollution exposure prediction approaches used in air pollution epidemiology studies. Journal of Exposure Science and Environmental Epidemiology 23, 566–572. https://doi.org/10.1038/jes.2013.15.

Parvez, F., Wagstrom, K., 2019. A hybrid modeling framework to estimate pollutant concentrations and exposures in near road environments. Science of Total Environment 663, 144–153. https://doi.org/10.1016/j.scitotenv.2019.01.218.

Pelliccioni, A., Tirabassi, T., 2006. Air dispersion model and neural network: a new perspective for integrated models in the simulation of complex situations. Environ. Model. Softw. 21 (4), 539–546.

Pelliccioni, A., Gariazzo, C., Tirabassi, T., 2003. Coupling of neural network and dispersion models: a novel methodology for air pollution models. Int. J. Environ. Pollut. 20 (1–6), 136–146.

Pernigotti, D., Georgieva, E., Thunis, P., Bessagnet, B., 2012. Impact of meteorology on air quality modeling over the Po valley in northern Italy. Atmos. Environ. 51, 303–310. https://doi.org/10.1016/j.atmosenv.2011.12.059.

Perrino, C., Catrambone, M., Dalla Torre, S., Rantica, E., Sargolini, T., Canepari, S., 2014. Seasonal variations in the chemical composition of particulate matter: a case study in the Po Valley. Part I: macro-components and mass closure. Environ. Sci. Pollut. Res. 21 (6), 3999–4009. https://doi.org/10.1007/s11356-013-2067-1.

Picornell, M., Ruiz, T., Borge, R., García-Albertos, P., de la Paz, D., Lumbreras, J., 2019. Population dynamics based on mobile phone data to improve air pollution exposure assessments. Journal of Exposure Science & Environmental Epidemiology 29, 278–291. https://doi.org/10.1038/s41370-018-0058-5.

Renzi, M., Cerza, F., Gariazzo, C., Agabiti, N., Cascini, S., Di Domenicantonio, R., Davoli, M., Forastiere, F., Cesaroni, G., 2018. Air pollution and occurrence of type 2 diabetes in a large cohort study. Environ. Int. 112 (2018), 68–76. https://doi.org/10.1016/j.envint.2017.12.007.

Scheers, H., Jacobs, L., Casas, L., Nemery, B., Nawrot, T.S., 2015. Long-term exposure to particulate matter air pollution is a risk factor for stroke: meta-analytical evidence. Stroke 46, 3058–3066. https://doi.org/10.1161/STROKEAHA.115.009913.

Schikowski, T., Adam, M., Marcon, A., Cai, Y., Vierkotter, A., Carsin, A.E., Jacquemin, B., Al Kanani, Z., Beelen, R., Birk, M., Bridevaux, P.O., Brunekeef, B., Burney, P., Cirach, M., Cyrys, J., De Hoogh, K., De Marco, R., De Nazelle, A., Declercq, C., Forsberg, B., Hardy, R., Heinrich, J., Hoek, G., Jarvis, D., Keidel, D., Kuh, D., Kuhlbusch, T., Migliore, E., Mosler, G., Nieuwenhuijsen, M.J., Phuleria, H., Rochat, T., Schindler, C., Villani, S., Tsai, M.Y., Zemp, E., Hansell, A., Kauffmann, F., Sunyer, J., Probst-Hensch, N., Kramer, U., Kunzli, N., 2014. Association of ambient air pollution with the prevalence and incidence of COPD. Eur. Respir. J. 44, 614–626. https://doi.org/10.1183/09031936.00132213.

Shtein, A., Kloog, I., Schwartz, J., Silibello, C., Michelozzi, P., Gariazzo, C., Viegi, G., Forastiere, F., Karnieli, A., Just, A.C., Stafoggia, M., 2020. Estimating daily PM2.5 and PM10 over Italy using an ensemble model. Environmental Science & Technology 54 (1), 120–128. https://doi.org/10.1021/acs.est.9b04279.

Silibello, C., Calori, G., Brusasca, G., Giudici, A., Angelino, E., Fossati, G., Peroni, E., Buganza, E., 2008. Modelling of PM_{10} concentrations over Milano urban area using two aerosol modules. Environ. Model. Softw. 23, 333–343.

Silibello, C., Calori, G., Finardi, S., Radice, P., Uboldi, F., Stafoggia, M., Gariazzo, C., Viegi, G., 2019. Three years simulation of meteorological parameters and airborne pollutants over Italy for exposure assessment of population. Proceedings of 19th International Conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes, 3–6 June 2019, Bruges, Belgium.

Skamarock, W.C., Klemp, J.B., Dudhia, J., Gill, D.O., Barker, D.M., Duda, M.G., Huang, X.Y., Wang, W., Powers, J.G., 2008. A Description of the Advanced Research WRF Version 3. *NCAR Tech. Note NCAR/TN-475+STR.* 113 pp. https://doi.org/10.5065/D68S4MVH.

Stafoggia, M., Cesaroni, G., Peters, A., Andersen, Z.J., Badaloni, C., Beelen, R., Caracciolo, B., Cyrys, J., de Faire, U., Gigante, B., Havulinna, A.S., Hennig, F., Hilding, A., Hoek, G., Hoffmann, B., Houthuijs, D., Korek, M., Lanki, T., Leander, K., Magnusson, P.K., Meisinger, C., Migliore, E., Overvad, K., Östenson, C., Pedersen, N.L., Pekkanen, J., Penell, J., Ranzi, A., Ricceri, F., Sacerdote, C., Swart, W.J.R., Turunen, A.W., 2014. Long-term exposure to ambient air pollution and incidence of cerebrovascular events: results from 11 European cohorts within the ESCAPE project. Environ. Health Perspect. 122, 919–925. https://doi.org/10.1289/ehp.1307301.

Stafoggia, M., Schwartz, J., Badaloni, C., Bellander, T., Alessandrini, E., Cattani, G., de' Donato, F., Gaeta, A., Leone, G., Lyapustin, A., Sorek-Hamer, M., de Hoogh, K., Di, Q., Forastiere, F., Kloog, I., 2017. Estimation of daily PM10 concentrations in Italy (2006–2012) using finely resolved satellite data, land use variables and meteorology. Environ. Int. 99, 234–244. https://doi.org/10.1016/j.envint.2016.11.024.

Stafoggia, M., Bellander, T., Bucci, S., Davoli, M., de Hoogh, K., de' Donato, F., Gariazzo, C., et al., 2019. Estimation of daily PM10 and PM2.5 concentrations in Italy, 2013–2015, using a spatiotemporal land-use random-forest model. Environ. Int. 124, 170–179. https://doi.org/10.1016/j.envint.2019.01.016.

Tramuto, F., Cusimano, R., Cerame, G., Vultaggio, M., Calamusa, G., Maida, C.M., Vitale, F., 2011. Urban air pollution and emergency room admissions for respiratory symptoms: a case-crossover study in Palermo, Italy. Environ. Health 10 (31). https://doi.org/10.1186/1476-069X-10-31 (2011).

Weissert, L.F., Salmond, J.A., Miskell, G., Alavi-Shoshtari, M., Williams, D.E., 2018. Development of a microscale land use regression model for predicting NO2 concentrations at a heavy trafficked suburban area in Auckland, NZ. Sci. Total Environ. 619–620, 112–119. https://doi.org/10.1016/j.scitotenv.2017.11.028.

Wright, M.N., Ziegler, A., 2017. Ranger: a fast implementation of random forests for high dimensional data in C++ and R. J. Stat. Softw. 77, 1–17. https://doi.org/10.18637/jss.v077.i01.

Yu, H., Russell, A., Mulholland, J., Huang, Z., 2018. Using cell phone location to assess misclassification errors in air pollution exposure estimation. Environ. Pollut. 233 (2018), 261–266. https://doi.org/10.1016/j.envpol.2017.10.077.

Zhang, Y., Bocquet, M., Mallet, V., Seigneur, C., Baklanov, A., 2012. Real-time air quality forecasting, part I: history, techniques, and current status. Atmos. Environ. 60 (2012), 632–655. https://doi.org/10.1016/j.atmosenv.2012.06.031.