Urban-Scale NO2 Prediction with Sensors Aboard Bicycles: A Comparison of Statistical Methods Using Synthetic Observations

Christophe Bertero, Jean-François Léon, Gilles Trédan, Matthieu Roy, Alexandre Armengaud

HAL Id: hal-02998998
https://hal.science/hal-02998998

Submitted on 10 Nov 2020

Urban-Scale NO$_2$ Prediction with Sensors Aboard Bicycles: A Comparison of Statistical Methods Using Synthetic Observations

Christophe Bertero [1], Jean-François Léon [2],*, Gilles Trédan [1], Mathieu Roy [1] and Alexandre Armengaud [3]

1. LAAS-CNRS, Université Toulouse, CNRS, 31400 Toulouse, France; cbertero@laas.fr (C.B.); gtredan@laas.fr (G.T.); Roy@laas.fr (M.R.)
2. Laboratoire d'Aérologie, Université Paul Sabatier, CNRS, 31400 Toulouse, France
3. AtmoSud, 13294 Marseille, France; alexandre.armengaud@atmosud.org
* Correspondence: jean-francois.leon@aero.obs-mip.fr

Abstract: Mobile devices for city-scale air quality monitoring is receiving increasing attention due to the advent of low-cost and miniaturized sensors. Mobility and crowdsensing have emerged as a new means to investigate the ambient air quality in urban areas. However, the design of the network (e.g., number of sensors per unit area) and the scientific interpretation of collected data with an ad hoc method are still challenging. In this paper, we focus on the use of a fleet of private bicycles to monitor NO$_2$ concentrations in the city of Marseille, France. The study is based on synthetic observations generated by means of a regional air quality simulation system at a spatial resolution of 25 m $\times$ 25 m and simulated bike trips that are randomly generated in the city. The bike trips correspond to a maximum of 4500 bike commuters and are generated using a web-based navigation service. Simulated bike tracks are validated using available statistics on bike counts. Each bike track is associated with the along-track corresponding NO$_2$ concentrations collected from the air quality simulations and physical features on the ground collected from Open Street Map. Spatialization of the information collected aboard the bikes is tested by using three different algorithms: kriging, land-use regression (LUR) and neural network (NN). LUR and NN show that the fleet can be limited to below 100 bikes while the performance of kriging is steadily increasing with the number of bikes. Increasing the sample distance above 200 m also impairs the citywide prediction of simulated NO$_2$ concentrations.

Keywords: air quality; mobility; urban-scale; NO$_2$; regression model

1. Introduction

The World Health Organization (WHO) estimates that ambient air pollution contributed to 3.2 million premature deaths worldwide in 2012, including more than 200,000 lung cancer deaths. The WHO provides guidelines for selected harmful chemical species such as PM$_{10}$ (fraction of airborne particles having an aerodynamic diameter less than 10 μm), ozone, nitrogen dioxide and sulfur dioxide [1]. Guidelines are based on estimated thresholds above which there is clear evidence of the effect of pollutants on human health. Nowadays most of the major cities of the world have dedicated air quality monitoring networks. In addition, numerical methods (hereinafter called models) have been intensively used to predict the concentrations of pollutants over time and space. The models can be roughly divided into two classes [2]: deterministic or statistical models. Deterministic models, also called chemical transport models, rely on emission estimates, explicit or parametrized chemical

reactions and physical processes such as advection or deposition (e.g., [3], for a review of European models). The major drawback of deterministic models is the large dependency of the simulated pollutant to the availability and accuracy of the inputs (i.e., emission inventories). Data assimilation practices are able to produce a "best estimate" by using both deterministic predictions and observations but are numerically expensive [4–6]. The second category of models is a statistical approach that aims to link pollutant concentrations with identified predictors, also called features. The statistical approach relies on estimation and prediction algorithms that range from a simple interpolation to a more complex machine learning method. Statistical models can be numerically cost-effective by construction and can adequately handle imprecise or missing measurements thus leading to robust prediction systems.

The advent of low cost and miniaturized sensors combined with pervasive and mobile computing resources paves the way to mobile sensor networks and crowdsensing systems targeting air pollution markers [7]. Such mass produced sensors enjoy low price tag, reduced size and energy requirements, and thus may be embedded in a mobile system carried by citizens. Crowdsensing seems a perfect fit at the city scale, where air pollution shows huge variations in space and time [8], and where many citizens may participate in measurement campaigns by carrying dedicated sensor systems. Statistical methods will clearly benefit from this new collection method. Indeed measurements can be widely distributed in time and space, taking advantage of users' mobility and thus are able to capture fine time and spatial variability in the pollutant concentrations [9]. At least, as pointed out in Hoek et al. [10], a wider spatial covering eases the discovery of extrema in a given area.

In this paper we focus on the ability of a bicycle-based mobile sensor network to predict air quality at the urban scale. Classical interpolation methods are commonly used and possibly adapted to the geolocalized aspect of the problem. These methods build a spatially continuous surface that fits the measurements and represents the phenomenon to predict. Two sub-categories stand out: proximity models and land-use regressions (LUR).

Kriging is a widely used proximity model [11] that minimizes the spatial variance of the residual (estimation error) without bias. Janssen et al. [12] have suggested an improved kriging technique by detrending the local character with regards to land use data, hereinafter also called geographical features. A particular case of kriging over an intrinsic spatial random field is described by Ionescu et al. [13], also called thin plate spline interpolation. Another method, very popular for its simplicity, is the inverse distance weighting. Sivaraman et al. [14] have compared the inverse distance weighting with kriging and have observed more robust results with ordinary kriging, although it is more complex to implement.

LUR models are a set of methods based on geographical features to interpolate the pollutant spatial distribution. Features may include land use type (e.g., vegetation type), human related variables (e.g., population density), meteorology, or distance to identified pollutant emission sources like roads or industries. LUR methods have been widely applied [15,16], compared [10,17] and mixed with other models [9,12]. LUR models extensively rely on Generalized Additive Models (GAM) that are able to capture non-linear phenomena [16,18]. They also have been extended by coupling with principal component analysis (PCA) to decrease the input space dimension [19]. These approaches can be combined. Universal kriging and kriging with an external drift can model the spatial error of the phenomenon and also use geographical features to estimate the deterministic structure of the phenomenon. Mercer et al. [20] have shown that universal Kriging performed as well as or better than analogous LUR for predicting NO_x exposure at given locations.

Finally, machine learning models are also widely used for spatial and temporal interpolation. Neural networks [9,21] are the more commonly used technique. Some specific station-based neural network for air quality [22,23] and a method for neural networks thanks to genetic algorithms [24] have been developed. More elaborate methods have also been tested, for instance, based on the Probabilistic Graph Model (PGM) [2] or based on the fuzzy logic with discretized measurements [25].

The temporal aspect is less studied and often treated as a feature like any other, sometimes discretized [26], or overcome by successive predictions. More complex models predict the temporal behavior of the pollutants from a time series of observations provided by fixed stations [27,28] by means of neural networks [29] or a combination of different techniques [22]. Deep-neural networks are also becoming more popular, being frequently of the spatio-temporal or recurrent kinds.

The objective of our study is to evaluate the ability of statistical models to spatialize air quality observations acquired aboard bicycles at the urban scale. Bicycles have already been used for air quality monitoring as they have many advantages [30]. First, the distance tracks are longer than by walking and so the tracks provide a larger spatial coverage in both road and pedestrian areas. Second, the embedded energy generator (dynamo) supplies additional power to the system without interfering with the measurements due pollutant emission. However, acquiring a dataset of bike-based pollution measurements large enough to apply geo-statistical methods remains very challenging both from the technical and crowdsensing aspects. So we base the study on synthetic observations that have been generated for a state-of-the-art numerical modeling system. We focus on NO_2, which is one of the main pollutants linked to traffic emissions in the urban atmosphere. NO_2 is a highly reactive gas belonging to the nitrogen oxide group (NO_x). Its average life time is about one day but can be reduced to a few hours within an urban environment with active photochemistry [31], leading to highly variable concentrations at the city scale. NO_2 can also be considered as a proxy for other primary emission like particulate matter as both pollutants are usually correlated [32].

The map issued from the air quality modeling system is used to produce synthetic observations that are generated using a random bike track algorithm. The ability of the selected statistical model to reproduce the spatial variability of the input map is then evaluated. The benefits of our approach, compared to other previous studies [33,34] that compare different methods using a given fixed dataset obtained from observations is twofold: (i) to validate model robustness over the entire studied area rather than over selected measuring points, and (ii) to test the sensitivity of the methods to a spatial perturbation in the observations.

2. Method

The proposed methodology is described in Figure 1. First, we generated the bike tracks and the associated synthetic observations and extracted the GIS features from Open Street Map (OSM). Then, we processed the selected statistical models to get predicted geographical maps of NO_2 concentrations. Finally, we compared the predicted maps to the reference map by means of statistical metrics that are presented in the following subsection. The training data set consists of the synthetic observations, which are then removed from the predicted map for the purpose of model evaluation.
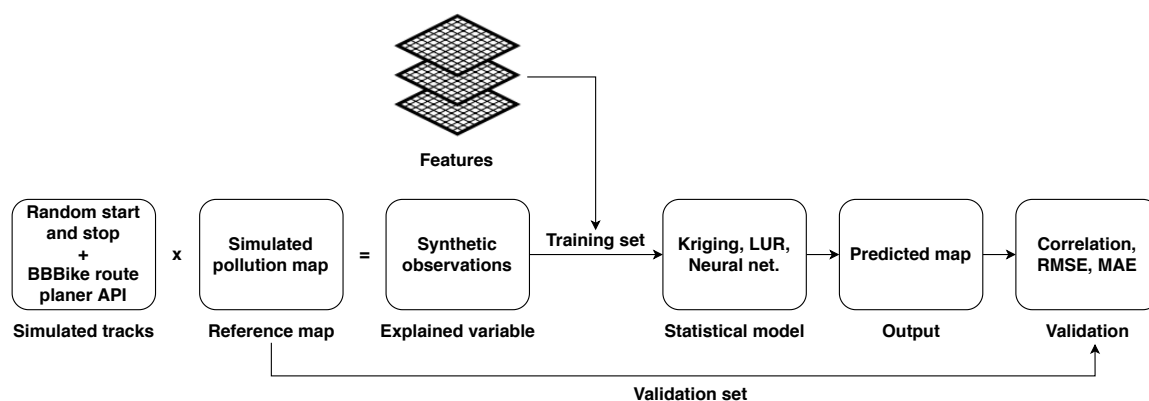

Figure 1. Methodological flowchart used in this study.

2.1. Study Area

Marseille (43°17′47″ N, 5°22′12″ E), is the second largest city in France in terms of population with over 860,000 inhabitants and third in terms of size, with an area of 240 km². Marseille is the city ranked number one for traffic congestion in France and 15th in the world. The city of Marseille is located in southern France on the shore of the Mediterranean sea. Our study area is a rectangle centered on the inner part of the city (downtown) covering an area of 140 km² and ranging from (43°14′9″ N, 5°20′42″ E) to (43°21′0″ N, 5°28′48″ E). Half of the surface of the area is classified as unbuildable natural territory. Marseille hosts the largest port in France and the second in the Mediterranean. Three major highways lead to the city, one from the north, one from the south and one along the coast, and some tunnels. The studied area host several industrial sites, two of them being classified SEVESO (a European directive identifying industrial sites presenting risks of major accidents). The airport is located 25 km westward of the city center.

2.2. Synthetic Observations

2.2.1. Air Quality Simulations

As the aim of the paper is not to discuss the accuracy of air quality simulations, we have relied on state-of-the-art available NO_2 simulations at the urban scale. The synthetic observations are generated from numerical simulations provided by the regional air quality agency AtmoSud. The operational air quality forecast model is part of the Atmospheric Integrated Regional System [35]. Daily forecasts are available at [36]. The spatial resolution is 25 m × 25 m and covers Marseille and its surroundings (see also [37] and references therein). The numerical model uses a combination of the regional chemistry transport model CHIMERE [38,39] and the street-level dispersion model ADMS-URBAN [40,41]. The methodology used for the coupling can be found in Hood et al. [42]. Outputs are reanalyzed by using the time series of local air quality monitoring stations.

NO_2 concentrations are simulated for a period of 14 days from the 12th to the 25th of February 2018. In order to reduce the computation time and to focus on the spatialization rather than on the time fluctuation of the pollutant, we use in the following the predicted daily maximum NO_2 concentration.

2.2.2. Geographical Features

Land use variables (geographical features) are collected from Open Street Map (OSM). OSM is the main collaborative and free project providing an urban cartography service [43]. An OSM file represents a region of the world and is composed of elementary objects of the conceptual model of the physical world: nodes, ways and relations. Nodes are defined by their geographical coordinates (longitude and latitude) and represent points of interest or serve for the definition of ways. Ways represent paths or space areas. Relationships are used to link objects together. Any object can be enriched with specific information using key-value tags. OSM has its own data format designed to facilitate the contribution of users. The contribution of a variety of users results in a large heterogeneity in the feature tags and numerous sub-categories that leads to a plurality of the considered features. The heterogeneity in categories and in spatial and time resolution of the features raises an issue in the comparability of the different studies [2,10]. The use of OSM service ensures a reproducible study, however, at the expense of the introduction of dynamic features, like pollutant emissions, meteorology, or traffic intensity.

Extracting the required features from OSM is not a straightforward process and we have relied on the pre-processed geographical database provided by GeoFabrik [44]. Categories that represent less than 0.5% of the total amount of layers or having less than 10 points are discarded. For instance, "Point of interest" categories (supermarket, restaurant, hospital, pharmacy, etc.) are gathered into nine classes (shopping, leisure, health, etc.). We also compute new layers such as the nearest distance to a main road or the building density. Missing data are filled with the median in case of numerical features and 'NA' in case of categorical features, respectively. The resulting shapefiles are projected into the

Lambert Conformal Conic projection (EPSG 2154) coordinate reference system, scaled, normalized and rasterized at resolution of 25 m × 25 m (see next section). Features extracted are pooled into Table 1. In addition to OSM data we used the topography layer of the Institut National de l'Information Géographique et Forestière (IGN) at a resolution of 75 m. Meteorological features were not included in the study although some studies indicate that meteorological features improve LUR models [45,46].

2.2.3. Simulated Bike Tracks

Since we neutralized the temporal component, we did not need to simulate a mobility model and the recurrent patterns of displacements. We only had to randomly select fictive bike tracks through the study area, without providing any information on the chronology, and check the spatial consistency of the simulated route set. The tracks were simulated by defining a starting location, a bearing and a distance. Starting locations were selected around the two main areas of interest in the city, namely the beaches and the old port and within an area defined by a normal distribution of variance 1 km. Destination locations were selected through a bearing and a geodesic distance. The bearing is uniformly chosen between 0° and 360°. The geodesic distance, smaller than the actual length of the trip, follows a normal distribution centered on 2.5 km and with a standard deviation of 4.5 km. To ensure the realism of the trajectories, a cyclist route planner application programming interface (API) based on OSM, named BBBike [47] is called for each track. The API provides a series of GPS coordinates for each direction change and sometimes annotated by street name. Figure 2 presents the city map, the probability distribution of a starting location, and the starting locations actually generated.
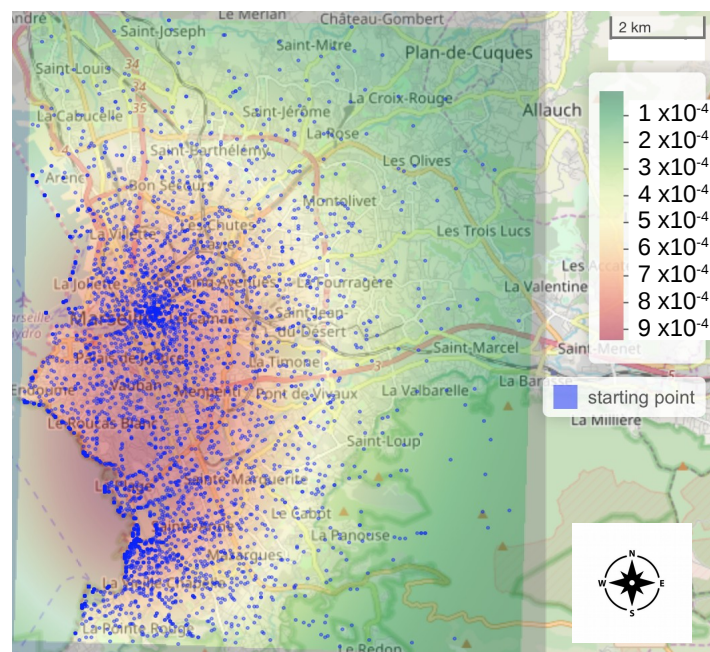

Figure 2. Probability distribution of a starting location (color scale represents probability density) and actual generated locations (blue points).

2.3. Statistical Models

2.3.1. Kriging

Kriging is commonly used to spatialize point measurements [48,49]. Kriging is an estimation method by spatial autocorrelation, i.e., the prediction is written

$$\hat{Y}(s_0) = \sum_{s_i \in V(s_0)} \lambda_i Y(s_i) \tag{1}$$

where s_0 is the position of the prediction, $V(s_0)$ the neighborhood of s_0, s_i a position where the predicted variable Y is known and the λ_i are the unknowns to be determined. The neighborhood $V(s_0)$ can be restricted to a given distance or to a limited number of data points. Kriging minimizes the spatial variance of the residual (error of estimation) without bias as given by Equation (2).

$$min(Var[\hat{Y}(s_0) - Y(s_0)]) \text{ subject to } E[\hat{Y}(s_0) - Y(s_0)] = 0 \tag{2}$$

Moreover, kriging generalizes the classical linear regression by assuming a spatial dependence of the error, i.e.,

$$Y(s) = \mu(s) + \delta(s) \tag{3}$$

where μ represents the deterministic structure of $E[Y]$ and δ a stationary zero-mean random function. Several kinds of Kriging exist depending on the chosen form of μ, but the most common and simple is Ordinary Kriging:

$$\mu(s) = \mu \tag{4}$$

with μ an unknown constant.

Finally, we determine δ thanks to the variographic analysis [50] that fits a smooth surface using a variogram constructed from the data points. Then Equations (1) and (3) are used into Equation (2) to get the final equation system. For computing-time concern, we decided to apply a local kriging limited to 100 neighbors. To implement it in R, we simply use the autoKrige function of the automap package [51], which automatically selects the best set of hypotheses for the error model (i.e., the variographic analysis) and then determines an approximation of Y over the convex hull of the training dataset.

2.3.2. Generalized Additive Model

Generalized Additive Model (GAM) is a combination of Generalized Linearized Models (GLM) and Additive Models (AM). GLM is a generalization of the linear regression to any distributions and using a link function g: $g(\hat{Y}) = \sum_i \beta_i X_i$. AM is a generalization of linear regression to non-linear phenomena: $\hat{Y} = \sum_i f_i(X_i)$ with f_i a smooth function belonging to a given class. Training an AM model is usually by a backfitting algorithm that alternates between fitting Y and fitting the functions f_i. Therefore, GAM is written as $g(\hat{Y}) = \sum_i f_i(X_i)$.

Based on Mueller et al. [18] and Hasenfratz et al. [16], we chose the logarithm function as the link function g and smooth regression splines of up to three degrees for the f_i class. To implement it in R, we use the package mgcv [52]. A set of tracks (see Section 2.4 below for the number of selected tracks) provides independent geolocated NO_2 concentrations along with pre-processed geographical features (see Section 2.2.2). The classes of categorical features that have not been observed in the set of tracks are removed from the model formula and replaced in the validation set (the whole remaining map) by the most observed classes. Model parameters (predictor variables) are around 40 and the training data set always contains more data points than model parameters.

2.3.3. Artificial Neural Network

An artificial neuron is conceptually inspired by a biological neuron and its functioning. In a neural network (NN), a neuron transfers an output according to its inputs, weighted with a synaptic weight that evolves during learning (synaptic plasticity). The whole network defines a function of the features. The most documented and implemented artificial neural network is the multilayer perceptron (MLP) [53]. Adams and Kanaroglou [9] have chosen this architecture for air pollution assessment (for PM_{10}, NO_2) using mobile measurements. The design of our neural network includes an input neuron by feature (see Table 1), one hidden layer with a size ranging from 5 to 25 neurons and the logistic function for the activation function. We used the R package RSNNS [54] for the implementation of MLP and caret [55] for the choice of the hidden layer size. As for the GAM, all unobserved classes of categorical features are removed to run the model.

Table 1. Description of continuous or categorical explanatory features extracted from the GeoFabrik and the Institut National de l'Information Géographique et Forestière (IGN) databases.

Explanatory Variable	Description
Continuous features	
Position	Spatial coordinates
Altitude	Altitude above sea level (IGN base)
Distance to main roads	Distance to 'motorway', 'trunk' or 'primary' roads in OSM calculated from roads layer
Buildings_a	Buildings density calculated from buildings_a layer
Maximum Speed	Speed limit extract from roads layer
Categorical features	
Network	Road and rail network merge of roads and railways layers
Transport	Transport infrastructure (bus stop, ferry terminal, taxi rank,…)
Landuse_a	Land use merged of de water_a, transport_a, traffic_a, natural_a and landuse_a
Traffic	Road network information (traffic lights, signaling, …)
POIs	Points of interest of the city classified by main classes (points), merge of pois and pofw layers
POIs_a	Points of interest of the city classified by main classes (surface), merge of pois_a and pofw_a layers
Tree	Presence of one or more trees in the city calculated from natural layers

2.4. Evaluation Procedure

The predictions from the statistical models are compared to the simulated maps, which is used as a reference. We have defined a set of well-know metrics to evaluate the regression models: Root Mean Squared Error (*RMSE*), Mean Absolute Error (*MAE*) and Pearson correlation coefficient (*R*).

The evaluation procedure aims at defining the optimal number of sensing platforms and then the optimal spatial resolution. A given number of bike tracks are selected at random for each of the 14 days considered as independent. The random draw is done for the following number of tracks: 2, 4, 6, 8, 10, 40, 80, 120, 250, 500, 1500, 3000 and 4500 tracks and a fixed spatial resolution (i.e., the distance between two measurements) of 100 m. For the determined number of tracks, the spatial resolution is set to 25, 50, 75, 100, 200, 300, 400, 500 and 1000 m, respectively. The random experiments are repeated 20 times.

Moreover, we have tested the ability of the statistical models to account for a spatial discontinuity in the input dataset, i.e., a NO_2 concentration that is an outlier for the given feature or location. The disturbances are used to evaluate the spatial behavior of the prediction in case of unexpected heterogeneities. The perturbations are of three types: a white noise on the whole map, a zonal and a line perturbation. The white noise corresponds to a random value with a standard deviation of 1% of the maximum value of the map. The zonal perturbation is circle of 2.5 km diameter and with a concentration set at four times the maximum value and exponentially decreasing toward the edge. This concentric disturbance mimics a point source at constant and even levels with no wind. The line perturbation is a constant value fixed at two times the max value along a road of length 2.5 km belonging to the *residential* category. We estimate the impact of the perturbation on the retrieval (i.e., the prediction of the model for the whole map) by evaluating Δ being defined by Equation (5) where δ is the perturbation, *M* the model and *Ref* the reference map.

$$\Delta = M(Ref + \delta) - M(Ref) \tag{5}$$

3. Results

3.1. Synthetic Observations

The bike trajectories were simulated without accounting for the traveling time and temporal evolution of the pollutant. Primary pollutants linked to traffic emission such as the one studied here have a predictable diurnal cycle that is driven by the morning and evening commutes during which the concentrations are the highest. So the simulation exercise is similar to a real-world case of bike commuters equipped with on-board sensors and riding for about 20 min.

We have fixed in our simulation the geodesic distance between the origin and the end of the track. The histogram of the actual track length is given in Figure 3 and compared to nationwide statistics for bike track distances provided by the Institut de la Statistique et des Etudes Economiques (INSEE). The average length is 5.81 km, which corresponds roughly to a duration of 25 min when considering an average bike speed between 12 and 14 km/h. The comparison with the nationwide statistic for France shows a large underestimation of short tracks (below 1 km) that are not accounted for in our simulation as we mainly focus on commuting distances. Consequently we also overestimate the number of tracks that are between 6 and 8 km and largely underestimate the number of tracks with length above 15 km.
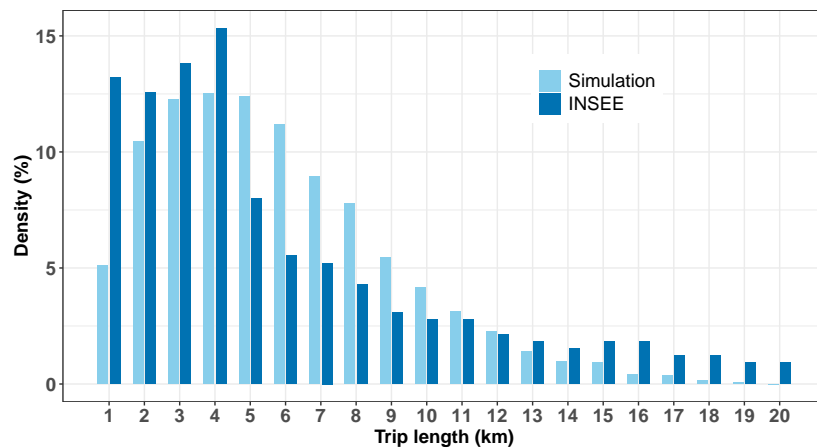

Figure 3. Density histogram of the bike track lengths of our simulations and compared to nationwide statistics provided by the Institut de la Statistique et des Etudes Economiques (INSEE).

Bicycle counts have been conducted in Marseille in 2017 and 2018 at different locations in the city and during rush hours (i.e., 8–9 a.m. and 4:30–6:30 p.m.). The comparison between the simulations and the bike counts (Table 2) shows that we overestimate the total number of bikes by a factor ranging from 1.2 (Michelet/Velodrome) to 5 (Corniche/Hélice). The largest overestimations are observed for roads in the outskirts. Nonetheless, the correlation coefficient between the count and the simulation is high (R = 0.72) showing that the spatial coverage and the number of the tracks are realistic.

Because the bikes are not running on all the ways in the city, there is a possible bias in the sampling of the features (e.g., no sampling on the highways). Figure 4 indicates that the categories *cycleway* and *tertiary* are oversampled as expected. The proportion of *service* and *residential* ways in the bike simulations is less than for the whole map. Categories like *rail*, *subway*, *tram* are not sampled while the category *motorway* is marginally sampled. So most of the road features are well covered.

Table 2. Simulated and observed bike counts at given road intersections in Marseille. The observations are provided by "Vélo en ville" [56].

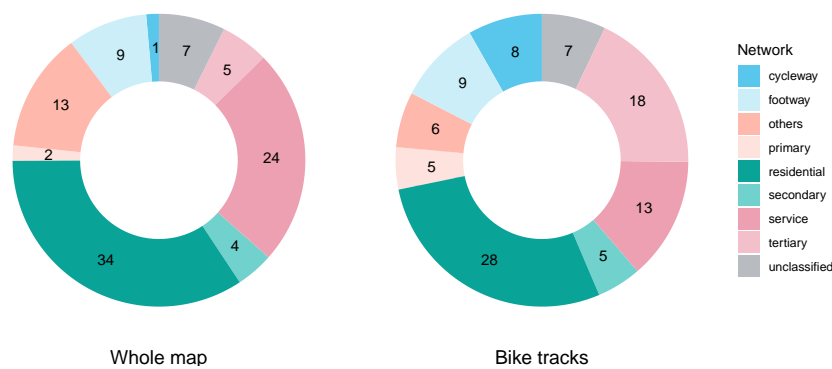| Designation	2017	2018	Simulated
Baille/Lodi	294	-	764
National/Guibal	175	-	487
Prado/Castellane	639	727	1297
Chave/Eugène Pierre	88	-	332
Joliette/République	178	-	623
Rome/Saint Louis	388	-	214
Vieux Port/Canebière	522	667	965
République/Sadi Carnot	262	-	416
Corniche/hélice	157	-	819
Michelet/vélodrome	438	506	638

Figure 4. Road type distribution (%) for the overall map and for our 4500 simulated bike tracks.

The synthetic observations are generated along each bike track from the air quality model output. Each track corresponds to a set of GPS coordinates, the list of associated features and the pollutant concentrations. Figure 5 gives an example of a bike run of 6 km simulated on February 14 for NO_2. The bike is running westward (from right to left) from downtown to the coast. The concentrations decrease from 100 to 20 $\mu g/m^3$ as the bike moves toward the footway on the shoreline. The concentrations do not depend on the road category, however some disruptions can be attributed to a change in the road type, for example, the decrease in the concentrations as the bike moves from the *tertiary* to the *secondary* category. The average concentrations found for each category corresponds to an expected distribution, being lower for the *service* and *residential* categories (27 $\mu g/m^3$ for NO_2) and higher for the *primary*, *secondary* and *tertiary*, increasing in that order. The few samples acquired in the *motorway* category (probably due to the 25 m resolution) give the highest concentration (110 $\mu g/m^3$).
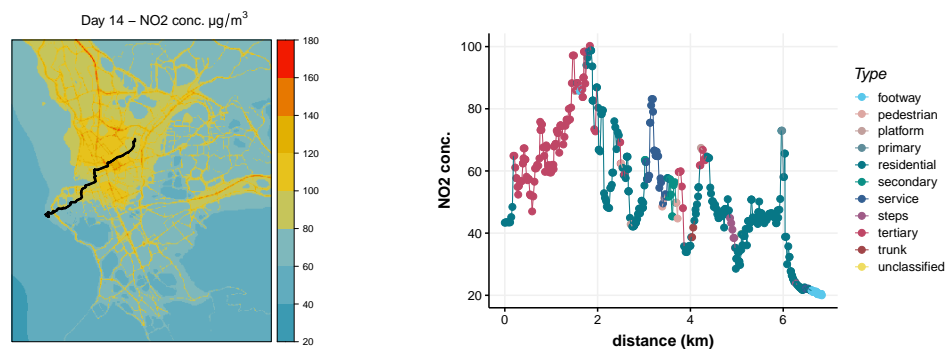

Figure 5. Bike track of 6 km in the city of Marseille on February 14 (**left**) map of NO_2 and (**right**) evolution of NO_2 depending on the distance traveled and the type of lane.

3.2. Model Prediction and Sensitivity

Figure 6 presents the reference map for day 14 and the associated mean prediction (see Section 2.4) from the three selected models and 40 bike tracks. The road network is noticeable in both the reference and the prediction maps. Obviously the kriging is not able to reproduce the fine scale discontinuities in the map, however the kriging interpolation captures the highest background level in the north-west area. The influence of the features in the LUR and NN predictions can be easily observed. NN has a better capacity to reproduce the elevated background in the downtown and north-west area while LUR smooths out the concentrations outside the traffic axis. The better performance in the NN can be attributed to its better capacity in fitting the numerical predictors, in particular the geographical coordinates, compared to the LUR, which performs better with the categorical features.
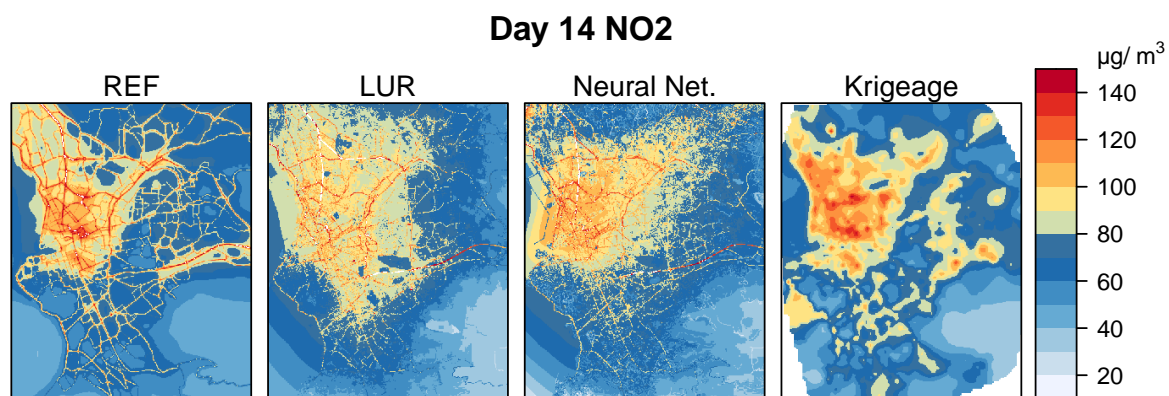
Day 14 NO2

Figure 6. Reference map for the NO_2 for day 14 and corresponding prediction by the different models.

The impact of the number of simulated tracks on the standardized outputs from the predictions is presented in Figure 7. The correlation coefficient increases as a function of the number of tracks while the MAE and RMSE decreases respectively. For a limited number of track (below 100) LUR performs better than the other models. For the maximal number of tracks, both NN and LUR have the same performances. Over 100 tracks, kriging performs better than the predictor-based techniques and has steady increases in performance as the number of tracks increases.

Kriging benefits locally from any new observations while NN and LUR try to fit a global model to the whole set of inputs. The drawback of kriging is the smoothing effect on pollutant concentrations while NN and LUR can reproduce better local heterogeneities based on feature localization.

We have estimated the standard deviation of the simulations for a number of tracks up to 125. For a small number of tracks the simulations may also depend on the draw within the 4500 simulated tracks. Above 125 tracks the probability to draw identical tracks is very low. The relative standard deviation is drastically decreased after 40 tracks (Figure 8), indicating that the choice of the paths (considering the aforementioned track distribution) have little influence on the simulations. LUR performs the best regarding the convergence of simulations with a limited number of tracks.

The sampling distance also has an impact on the retrieved concentrations. Starting with an ideal sampling of 25 m (the resolution of the input raster) we have tested a different distance up to one observation per km. Below 300 m, the sampling distance has no impact on the predictions. For larger sampling distances, the error increases as a function of the sampling distance for all the models. However, as shown in Figure 8, the relative uncertainty of the simulations also depends on the sampling distance. For increasing sampling distance, kriging and NN have an excellent capacity to interpolate between data points and so the relative uncertainty tends to steadily decrease. However, both models diverge after 300 m. LUR performance is degraded as the sampling distance increases and converges to the behavior of the NN after 300 m.
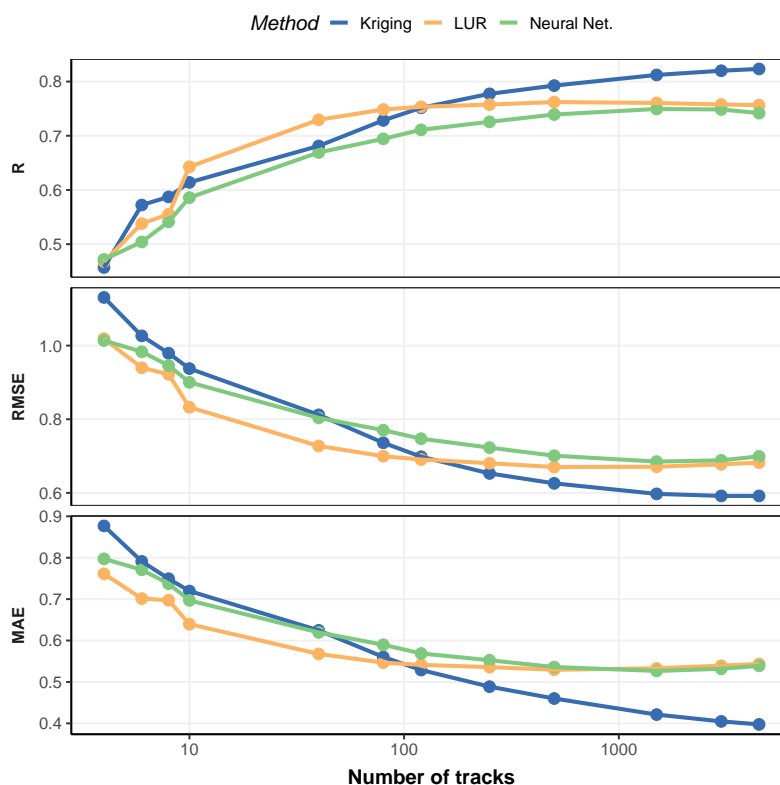
Figure 7. Comparison between the reference map and the predicted map according to the number of trips, to a sampling every 100 m. Logarithmic scale.

NO_2 for the whole map ranges between 14 and 160 µg/m^3. The whole range of NO_2 is encountered even with a small number of tracks (90% of the range for only 1 track) indicating that the heterogeneity of the concentrations are observed at a fine spatial scale. Increasing the sampling distance above 300 m decreases the ability of the system to monitor the concentration range by a few percent starting at 100% at 100 m (40 bikes) to 90% at 1 km.

The spatial coverage is approximated by the total number of cells that have been visited. This number is normalized by the maximum number of tracks (4500) to give a spatial coverage rate. Similarly, the feature coverage rate is estimated as a function of the number of tracks. However, the spatial coverage rate remains low (12.5% for 100 tracks), and the feature coverage rate is high (90% for 100 tracks). The maximum of sensitivity is around 40 tracks and increases steadily after 100 tracks. The rather good performances of the LUR and NN models can be partially explained by the fast increase in the feature coverage rate. The errors are impacted by the choice of the number of tracks and the spatial resolution, however from the point of view of our simulations, there is a set of both parameters that optimizes the predictions. For a given set of tracks around 40, the optimal sampling distance is about 200 m.

Categorical features like POI are well predicted, however the road network has the lowest performance. The prediction is the best for cycle ways due to the good spatial coverage (see Figure 4). The models are not able to reproduce the variability in pollutant concentrations for broad categories as *primary* or *secondary* ways. The error in the prediction is one order of magnitude higher for the *primary* ways than for the cycle ways. Kriging performs poorly around traffic lanes because of the smoothing impact. The concentrations in the western part of the domain above the sea are extrapolated based on model fitting for low altitude and so is overestimated. Some of the categories that have not been sampled (e.g., natural reserves or industrial areas) perform badly. In this case, the increase in the number of tracks enhances the probability of sampling scarce classes of categorical features.
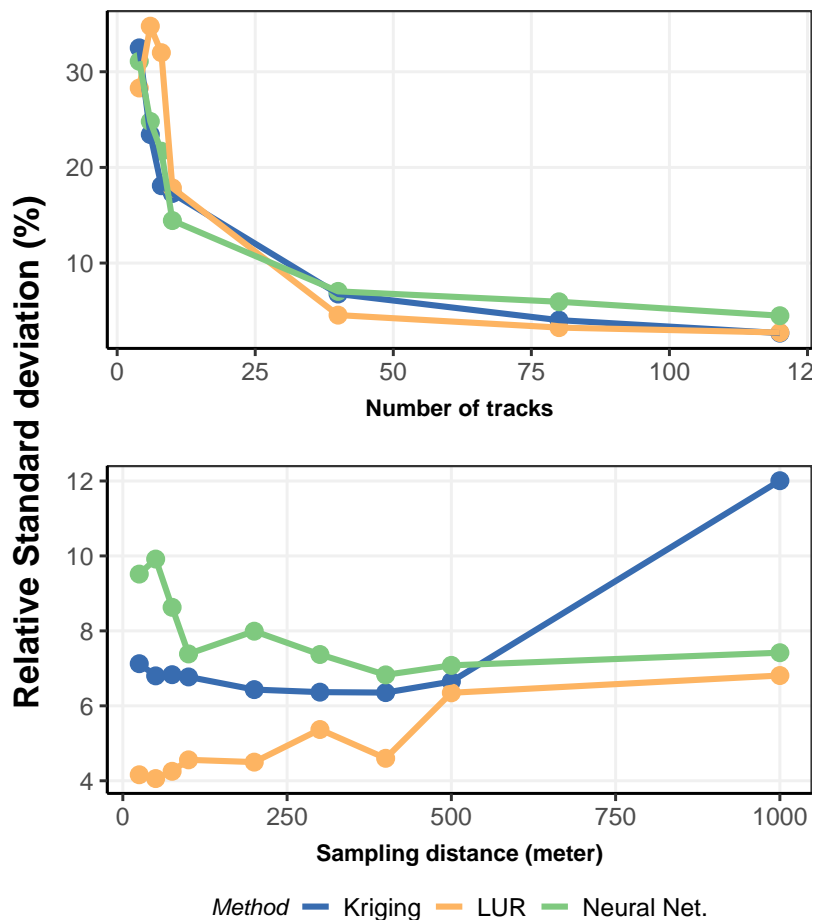
Figure 8. Relative standard deviation of the simulations for the correlation coefficient as a function of the number of (**top**) tracks and (**bottom**) sampling distance.

3.3. Sensitivity of the Methods to a Perturbation

Following the method presented in Section 2.4, the ability of the model to reconstruct a local perturbation is evaluated. First, adding a random noise to the input data set has a negligible impact on the predictions. The kriging technique retrieves the random noise and the methods using the physical features (LUR and NN) distribute the error among the different categorical features. The overall increase in error is limited to 10% of the standard deviation of the white noise.

Second, the line and the concentric perturbations are well localized by the kriging model. LUR attributes the line perturbation to all the roads of the same category but the remaining of the map is not disturbed. Figure 9 represents how the disturbance added to the reference map is estimated by the different models for a zonal perturbation (labeled REF in Figure 9). Such a simulation highlights the impact of heterogeneities in the observations that are not linked to any specific feature. The concentric perturbation has the same impact as the line perturbation on the LUR, the roads affected by the perturbation being increased in the whole map. The prediction of the perturbation by the NN shows that the whole map is affected. The perturbation has a localized impact and NN artificially compensates the perturbation for the whole domain.
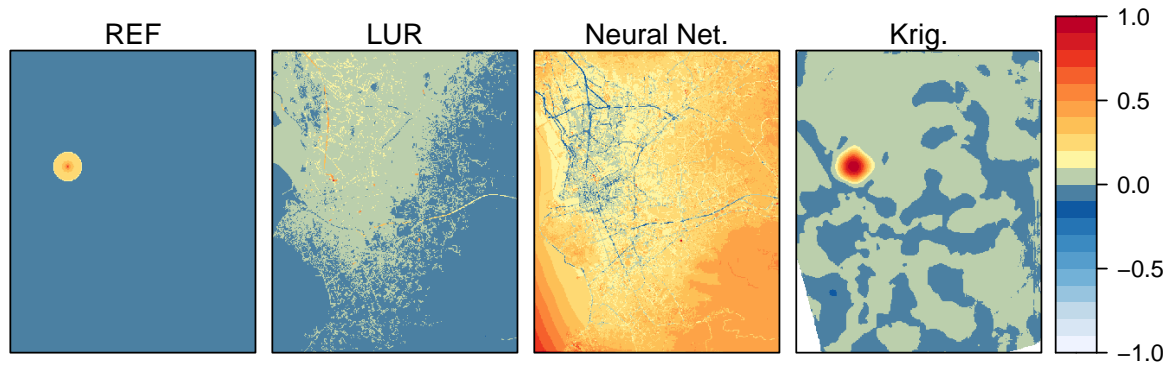
Figure 9. Concentric disturbance introduced on the reference map (REF) and predictions of this disturbance by spatialization methods. Maps are normalized by the maximum value of the predicted values for each map.

4. Conclusions and Outlook

Mobile air quality measurement platforms give a new insight into very fine scale urban pollution. Although low cost mobile air quality devices are now made available for the scientific community, the use of such data for air quality prediction remains a challenge. Our study is based on state-of-the-art numerical air quality simulations to evaluate how a set of realistic mobile observations from bike tracks can be extrapolated to a city-scale map. Different statistical spatialization methods were tested, Kriging, LUR and neural networks on NO_2 simulations. The local heterogeneities linked to spatial features are well reproduced by the LUR and the neural network model. The granularity of the feature description can affect the accuracy of the model in areas having a high variability, as for instance the residential roads. Moreover, models can be biased due to unvisited areas or outlier concentrations for a given category. Kriging is more robust to outliers and its accuracy increases with the size of the training dataset.

For NO_2, we show that a limited number of bikes (less than 100) with a sampling rate of 200 m can be used to retrieve the pollution state coherently with the reference map. As low cost sensors are subject to large measurement errors, significant drifts and possible failures, the approach developed in this paper also provides a framework for studying the impact of measurement uncertainties or device failures on the predictions.

The methodology presented here can be applied to any city provided the synthetic observations and the predictor variables are available. Applying the methodology to other primary pollutants mostly linked to traffic emission would probably result in the same conclusions. A perspective of this work is to account for the traveling time of the bikes. Assessing the time dependence of pollutant emissions would probably impact the results of the need for a larger fleet. Proximity and land-used based models are complementary, analyzing discontinuities that are not related to specific features and should both be used in the processing of actual mobile observations.

References

1.   World Health Organization. *Evolution of WHO Air Quality Guidelines Past, Present and Future*; OCLC: 1075973767; World Health Organization: Geneva, Switzerland, 2017.

2.   Marjovi, A.; Arfire, A.; Martinoli, A. High Resolution Air Pollution Maps in Urban Environments Using Mobile Sensor Networks. In Proceedings of the International Conference on Distributed Computing in Sensor Systems, Fortaleza, Brazil, 10–12 June 2015. [CrossRef]

3.   Thunis, P.; Miranda, A.; Baldasano, J.M.; Blond, N.; Douros, J.; Graff, A.; Janssen, S.; Juda-Rezler, K.; Karvosenoja, N.; Maffeis, G.; et al. Overview of Current Regional and Local Scale Air Quality Modelling Practices: Assessment and Planning Tools in the EU. *Environ. Sci. Policy* **2016**. [CrossRef]

4.   Benedetti, A.; Morcrette, J.J.; Boucher, O.; Dethof, A.; Engelen, R.J.; Fisher, M.; Flentje, H.; Huneeus, N.; Jones, L.; Kaiser, J.W.; et al. Aerosol Analysis and Forecast in the European Centre for Medium-Range Weather Forecasts Integrated Forecast System: 2. Data Assimilation. *J. Geophys. Res.* **2009**. [CrossRef]

5.   Tilloy, A.; Mallet, V.; Poulet, D.; Pesin, C.; Brocheton, F. BLUE-Based NO 2 Data Assimilation at Urban Scale. *J. Geophys. Res. Atmos.* **2013**. [CrossRef]

6.   Menut, L.; Bessagnet, B. What Can We Expect from Data Assimilation for Air Quality Forecast? Part I: Quantification with Academic Test Cases. *J. Atmos. Ocean. Technol.* **2019**, *36*, 269–279. [CrossRef]

7.   Gressent, A.; Malherbe, L.; Colette, A.; Rollin, H.; Scimia, R. Data Fusion for Air Quality Mapping Using Low-Cost Sensor Observations: Feasibility and Added-Value. *Environ. Int.* **2020**, *143*, 105965. [CrossRef]

8.   Deville Cavellin, L.; Weichenthal, S.; Tack, R.; Ragettli, M.S.; Smargiassi, A.; Hatzopoulou, M. Investigating the Use of Portable Air Pollution Sensors to Capture the Spatial Variability of Traffic-Related Air Pollution. *Environ. Sci. Technol.* **2016**. [CrossRef]

9.   Adams, M.D.; Kanaroglou, P.S. Mapping real-time air pollution health risk for environmental management: Combining mobile and stationary air pollution monitoring with neural network models. *J. Environ. Manag.* **2016**. [CrossRef]

10.   Hoek, G.; Beelen, R.; Hoogh, K.D.; Vienneau, D.; Gulliver, J.; Fischer, P.; Briggs, D. A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmos. Environ.* **2008**. [CrossRef]

11.   Jerrett, M.; Arain, A.; Kanaroglou, P.; Beckerman, B.; Potoglou, D.; Sahsuvaroglu, T.; Morrison, J.; Giovis, C. A review and evaluation of intraurban air pollution exposure models. *J. Expo. Anal. Environ. Epidemiol.* **2004**. [CrossRef]

12.   Janssen, S.; Dumont, G.; Fierens, F.; Mensink, C. Spatial interpolation of air pollution measurements using CORINE land cover data. *Atmos. Environ.* **2008**. [CrossRef]

13.   Ionescu, A.; Candau, Y.; Mayer, E.; Colda, I. Analytical determination and classification of pollutant concentration fields using air pollution monitoring network data: Methodology and application in the Paris area, during episodes with peak nitrogen dioxide levels. *Environ. Model. Softw.* **2000**. [CrossRef]

14.   Sivaraman, V.; Carrapetta, J.; Hu, K.; Luxan, B.G. HazeWatch: A participatory sensor system for monitoring air pollution in Sydney. In Proceedings of the 38th Annual IEEE Conference on Local Computer Networks-Workshops, Sydney, Australia, 21–24 October 2013. [CrossRef]

15.   Su, J.G.; Jerrett, M.; Beckerman, B.; Wilhelm, M.; Ghosh, J.K.; Ritz, B. Predicting traffic-related air pollution in Los Angeles using a distance decay regression selection strategy. *Environ. Res.* **2009**. [CrossRef] [PubMed]

16.   Hasenfratz, D.; Saukh, O.; Walser, C.; Hueglin, C.; Fierz, M.; Arn, T.; Beutel, J.; Thiele, L. Deriving high-resolution urban air pollution maps using mobile sensor nodes. *Pervasive Mob. Comput.* **2015**. [CrossRef]

17.   Ghassoun, Y.; Ruths, M.; Löwner, M.O.; Weber, S. Intra-urban variation of ultrafine particles as evaluated by process related land use and pollutant driven regression modelling. *Sci. Total. Environ.* **2015**. [CrossRef] [PubMed]

18.   Mueller, M.D.; Hasenfratz, D.; Saukh, O.; Fierz, M.; Hueglin, C. Statistical modelling of particle number concentration in Zurich at high spatio-temporal resolution utilizing data from a mobile sensor network. *Atmos. Environ.* **2016**. [CrossRef]

19.   Li, S.; Zhai, L.; Zou, B.; Sang, H.; Fang, X. A Generalized Additive Model Combining Principal Component Analysis for PM2.5 Concentration Estimation. *ISPRS Int. J. Geo-Inf.* **2017**, *6*, 248. [CrossRef]

20.   Mercer, L.D.; Szpiro, A.A.; Sheppard, L.; Lindstrom, J.; Adar, S.D.; Allen, R.W.; Avol, E.L.; Oron, A.P.; Larson, T.; Liu, L.J.S.; et al. Comparing universal kriging and land-use regression for predicting concentrations of gaseous oxides of nitrogen (NOx) for the Multi-Ethnic Study of Atherosclerosis and Air Pollution (MESA Air). *Atmos. Environ.* **2011**. [CrossRef]

21.   Kurt, A.; Gulbagci, B.; Karaca, F.; Alagha, O. An online air pollution forecasting system using neural networks. *Environ. Int.* **2008**. [CrossRef] [PubMed]

22. Zheng, Y.; Yi, X.; Li, M.; Li, R.; Shan, Z.; Chang, E.; Li, T. Forecasting Fine-Grained Air Quality Based on Big Data. In Proceedings of the 21th SIGKDD Conference on Knowledge Discovery and Data Mining, Sydney, Australia, 10–13 August 2015. [CrossRef]

23. Hsieh, H.P.; Lin, S.D.; Zheng, Y. Inferring Air Quality for Station Location Recommendation Based on Urban Big Data. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, Australia, 10–13 August 2015. [CrossRef]

24. Niska, H.; Hiltunen, T.; Karppinen, A.; Ruuskanen, J.; Kolehmainen, M. Evolving the neural network model for forecasting air pollution time series. *Eng. Appl. Artif. Intell.* **2004**. [CrossRef]

25. Onkal-Engin, G.; Demir, I.; Hiz, H. Assessment of urban air quality in Istanbul using fuzzy synthetic evaluation. *Atmos. Environ.* **2004**. [CrossRef]

26. Dons, E.; Poppel, M.V.; Kochan, B.; Wets, G.; Panis, L.I. Modeling temporal and spatial variability of traffic-related air pollution: Hourly land use regression models for black carbon. *Atmos. Environ.* **2013**. [CrossRef]

27. Romanowicz, R.; Young, P.; Brown, P.; Diggle, P. A recursive estimation approach to the spatio-temporal analysis and modelling of air quality data. *Environ. Model. Softw.* **2006**. [CrossRef]

28. Qi Gan, W.; Koehoorn, M.; Davies, H.W.; Demers, P.A.; Tamburic, L.; Brauer, M. Long-Term Exposure to Traffic-Related Air Pollution and the Risk of Coronary Heart Disease Hospitalization and Mortality. *Environ. Health Perspect.* **2011**. [CrossRef]

29. Russo, A.; Raischel, F.; Lind, P.G. Air quality prediction using optimal neural networks with stochastic variables. *Atmos. Environ.* **2013**. [CrossRef]

30. Elen, B.; Peters, J.; Poppel, M.V.; Bleux, N.; Theunis, J.; Reggente, M.; Standaert, A. The Aeroflex: A Bicycle for Mobile Air Quality Measurements. *Sensors* **2013**, *13*, 221–240. [CrossRef] [PubMed]

31. Beirle, S.; Boersma, K.F.; Platt, U.; Lawrence, M.G.; Wagner, T. Megacity Emissions and Lifetimes of Nitrogen Oxides Probed from Space. *Science* **2011**, *333*, 1737–1739. [CrossRef] [PubMed]

32. Madhavi Latha, K.; Highwood, E.J. Studies on Particulate Matter (PM10) and Its Precursors over Urban Environment of Reading, UK. *J. Quant. Spectrosc. Radiat. Transf.* **2006**, *101*, 367–379. [CrossRef]

33. Deligiorgi, D.; Philippopoulos, K. Spatial Interpolation Methodologies in Urban Air Pollution Modeling: Application for the Greater Area of Metropolitan Athens, Greece. *Adv. Air Pollut.* **2011**. [CrossRef]

34. Wong, D.W.; Yuan, L.; Perlin, S.A. Comparison of spatial interpolation methods for the estimation of air quality data. *J. Expo. Sci. Environ. Epidemiol.* **2004**. [CrossRef]

35. AIRES Méditerranée. Available online: http://www.aires-mediterranee.org/ (accessed on 21 September 2020).

36. AtmoSud. Available online: https://www.atmosud.org/ (accessed on 21 September 2020).

37. Riviere, E.; Bernard, J.; Hulin, A.; Virga, J.; Dugay, F.; Charles, M.A.; Cheminat, M.; Cortinovis, J.; Ducroz, F.; Laborie, A.; et al. Air Pollution Modeling and Exposure Assessment during Pregnancy in the French Longitudinal Study of Children (ELFE). *Atmos. Environ.* **2019**, *205*, 103–114. [CrossRef]

38. Menut, L.; Bessagnet, B.; Khvorostyanov, D.; Beekmann, M.; Blond, N.; Colette, A.; Coll, I.; Curci, G.; Foret, G.; Hodzic, A.; et al. CHIMERE 2013: A Model for Regional Atmospheric Composition Modelling. *Geosci. Model Dev.* **2013**. [CrossRef]

39. Mailler, S.; Menut, L.; Khvorostyanov, D.; Valari, M.; Couvidat, F.; Siour, G.; Turquety, S.; Briant, R.; Tuccella, P.; Bessagnet, B.; et al. CHIMERE-2017: From Urban to Hemispheric Chemistry-Transport Modeling. *Geosci. Model Dev.* **2017**, *10*, 2397–2423. [CrossRef]

40. Carruthers, D.J.; Holroyd, R.J.; Hunt, J.C.R.; Weng, W.S.; Robins, A.G.; Apsley, D.D.; Thompson, D.J.; Smith, F.B. UK-ADMS: A New Approach to Modelling Dispersion in the Earth's Atmospheric Boundary Layer. *J. Wind Eng. Ind. Aerodyn.* **1994**, *52*, 139–153. [CrossRef]

41. Carruthers, D.; Stidworthy, A.; Clarke, D.; Dicks, J. Urban Emission Inventory Optimisation Using Sensor Data, an Urban Air Quality Model and Inversion Techniques. *Int. J. Environ. Pollut.* **2019**, *4*, 15. [CrossRef]

42. Hood, C.; MacKenzie, I.; Stocker, J.; Johnson, K.; Carruthers, D.; Vieno, M.; Doherty, R. Air Quality Simulations for London Using a Coupled Regional-to-Local Modelling System. *Atmos. Chem. Phys.* **2018**, *18*, 11221–11245. [CrossRef]

43. Haklay, M.; Weber, P. OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Comput.* **2008**, *7*, 12–18. [CrossRef]

44. GEOFABRIK. Available online: https://www.geofabrik.de/ (accessed on 21 September 2020).

45. Arain, M.A.; Blair, R.; Finkelstein, N.; Brook, J.R.; Sahsuvaroglu, T.; Beckerman, B.; Zhang, L.; Jerrett, M. The Use of Wind Fields in a Land Use Regression Model to Predict Air Pollution Concentrations for Health Exposure Studies. *Atmos. Environ.* **2007**, *41*, 3453–3464. [CrossRef]

46. Wilton, D.; Szpiro, A.; Gould, T.; Larson, T. Improving Spatial Concentration Estimates for Nitrogen Oxides Using a Hybrid Meteorological Dispersion/Land Use Regression Model in Los Angeles, CA and Seattle, WA. *Sci. Total Environ.* **2010**, *408*, 1120–1130. [CrossRef]

47. BBBike. Available online: https://www.bbbike.org/ (accessed on 21 September 2020).

48. Kim, S.Y.; Yi, S.J.; Eum, Y.S.; Choi, H.J.; Shin, H.; Ryou, H.G.; Kim, H. Ordinary Kriging Approach to Predicting Long-Term Particulate Matter Concentrations in Seven Major Korean Cities. *Environ. Health Toxicol.* **2014**, *29*. [CrossRef]

49. Li, J.; Zhang, H.; Chao, C.Y.; Chien, C.H.; Wu, C.Y.; Luo, C.H.; Chen, L.J.; Biswas, P. Integrating Low-Cost Air Quality Sensor Networks with Fixed and Satellite Monitoring Systems to Study Ground-Level PM2.5. *Atmos. Environ.* **2020**, *223*, 117293. [CrossRef]

50. Oliver, M.A.; Webster, R. *Basic Steps in Geostatistics: The Variogram and Kriging*; SpringerBriefs in Agriculture; Springer International Publishing: Cham, Switzerland, 2015. [CrossRef]

51. Hiemstra, P.H.; Pebesma, E.J.; Twenhöfel, C.J.W.; Heuvelink, G.B.M. Real-time automatic interpolation of ambient gamma dose rates from the Dutch radioactivity monitoring network. *Comput. Geosci.* **2009**. [CrossRef]

52. Wood, S.N. *Generalized Additive Models: An Introduction with R*; Chapman & Hall: Boca Raton, FL, USA, 2017.

53. Jain, A.K.; Mao, J.; Moidin Mohiuddin, K. Artificial Neural Networks: A Tutorial. *Computer* **1996**. [CrossRef]

54. Bergmeir, C.; Benítez, J.M. Neural Networks in R Using the Stuttgart Neural Network Simulator: RSNNS. *J. Stat. Softw.* **2012**. [CrossRef]

55. Kuhn, M. Building Predictive Models in *R* Using the caret Package. *J. Stat. Softw.* **2008**. [CrossRef]

56. Collectif Vélos En Ville. Available online: http://www.velosenville.org (accessed on 21 September 2020).