City and Environment Interactions 28 (2025) 100236

Contents lists available at ScienceDirect

City and Environment Interactions

journal homepage: www.sciencedirect.com/journal/city-and-environment-interactions

Research articles

# Modeling nitrogen dioxide concentrations using citizen science data: The case of the Brussels-Capital Region

Patrick Bogaert [a,*], Noémie Huvelle [a], Axel Briffault [a], Olivier Brasseur [b]

[a] Earth & Life Institute, Université catholique de Louvain (UCLouvain), Belgium
[b] Brussels Environment, Belgium

ARTICLE INFO

*Keywords:*
Nitrogen dioxide
Outdoor air pollution
Geospatial modeling
Land use regression
Multiple linear regression
Random forest regression
SIRANE

ABSTRACT

Air pollution caused by $NO_2$ emissions related to traffic is a major environmental issue in the Brussels-Capital region. Using a large set of measurements collected from a citizen science campaign, this paper shows how such data help us to get an overview of the spatial distribution of $NO_2$ levels over the region. Using two land use regression techniques, these levels were related to spatial proxies collected at the measurement locations. Comparing the proxies selected by each regression method offers deeper insights into the $NO_2$-proxies relationships and helps identify proxies that may have been overlooked in a simpler multilinear regression model. Results show that the multiple linear regression model is able to explain a major part of the variance of the data, while random forest regression performs slightly better, with performances that are on par with those found in the literature. However, both models tend to underestimate high concentrations that are occurring locally. Thanks to a comparison with the prediction results from a physics-based model, this could be related to the quality of the input traffic data, that are expected to play a major role as most of nitrogen oxides emissions in the Brussels-Capital region originate from road traffic.

## 1. Introduction

Air pollution is a major ongoing environmental challenge in urban areas. Among other pollutants, nitrogen dioxide ($NO_2$) that primarily originates from road transport [1] has detrimental human health and environmental effects, either directly or through its role in the formation of secondary pollutants such as ozone and particulate matter [2]. Chronic exposure to elevated $NO_2$ levels is associated with respiratory diseases, cardiovascular issues, and increased mortality [3–5]. Additionally, $NO_2$ contributes to environmental degradation through processes such as acid rain [6], further underscoring the need for mitigation efforts.

In spite of a significant air quality improvement over the last decade thanks to EU regulations and action plans, $NO_2$ emissions are still a concern in Belgium, that ranks fifth among the fourteen EU countries [7] that exceed the European annual limit value for $NO_2$ in several major cities. In particular, the Brussels-Capital Region (BCR) is one of the most problematic area in the country, where road traffic is the main source of $NO_2$ emissions in a region that has no significant industrial activities. The high population density and the urban morphology of the BCR further exacerbate the impact of these emissions on air quality, especially in street canyons roads that trap pollutants [8].

In this context, modeling the spatial distribution of $NO_2$ concentrations is crucial for understanding exposure patterns, estimating $NO_2$ levels in unmonitored areas, and developing effective mitigation strategies. Several modeling approaches have been proposed, each subject to its specific strengths and limitations.

Interpolation methods are relying on spatial correlations between measurement locations for estimating pollutant concentrations in unmonitored areas [9–11]. While this approach is simple and effective over areas where pollutant concentrations are expected to gradually change in space, it faces major challenges in urban environments that are heterogeneous by nature due to the urban morphology. Indeed, urban features such as buildings, roads and distinct land uses create small-scale variations in pollutant concentrations, thus leading to weaker spatial correlations and complex spatial patterns. As a result, interpolation methods may struggle to provide accurate estimates in areas where spatial variations or urban morphology are marked [12].

Land Use Regression (LUR) models aim at collecting spatial information over the region of interest with the goal of linking the pollutant concentrations to spatial proxies measured at the same locations using various statistical regression techniques. These range from traditional Multiple Linear Regression (MLR) models [13–16] to more recent artificial intelligence and machine learning techniques such as, e.g., random

* Corresponding author.
*E-mail addresses:* patrick.bogaert@uclouvain.be (P. Bogaert), noemie.huvelle@uclouvain.be (N. Huvelle).

https://doi.org/10.1016/j.cacint.2025.100236
Received 10 January 2025; Received in revised form 4 August 2025; Accepted 12 August 2025
Available online 21 August 2025
2590-2520/© 2025 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/).

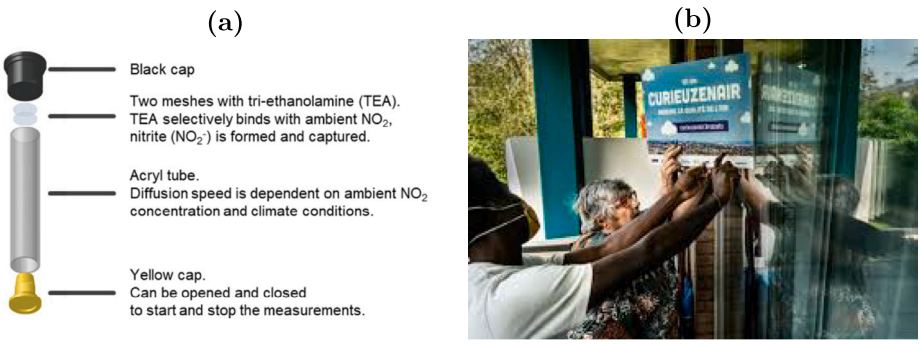


**Fig. 1.** CurieuzenAir measurement setup. (a) Components of a Palmes tube. (b) Installation of the measurement setup on the wall facade of a residence; two tubes are placed on the back of the measurement panels, with openings facing downward.
*Source:* Images are adapted from [32].

forest regression (RFR) [17,18] and neural networks [19,20], with varying data requirements and capabilities for modeling the relationships between pollutant concentrations and proxies.

Alternatively, mechanistic dispersion models rely on the simulation of pollutant emissions, atmospheric transport and chemical transformations with the aim of gaining some insights into the space–time dynamics of pollutant concentrations [21–23]. While these models are built on a strongly grounded chemico-physical framework, they require substantial input data and calibration steps, making their implementation more difficult and complex. Among these models, the SIRANE model [22] is currently used in few major European cities, and its benefits in the BCR are currently investigated for predicting $NO_2$ concentrations on an hourly basis at the street segment level.

Beside the variety of approaches that can be considered for modeling urban air pollution, an important aspect is the amount of data that are at hand for feeding these models. In most European cities, air pollutant monitoring is typically based on a few high-quality automatic measurement stations that are remotely providing data on an hourly basis. In BCR, a network of 13 telemetric stations provides hourly measurements of several air pollutants [24], and this sparse network does not allow us to have a clear view of the spatial distribution of these pollutants over the region. Unfortunately, due to the high cost and maintenance requirements of such stations, it is unlikely that their number will increase in a close future. To address this issue of spatial data scarcity, citizen-based air quality measurement campaigns have gained prominence over the last decades [25–28]. By providing cost-affordable data that are densely covering large areas, they offer valuable insights about pollutants spatial distribution, even if the reliability of these data can be questioned in some instances. Despite the growing use of crowdsourced air quality data, studies explicitly linking crowdfunded $NO_2$ measurements to the development or validation of predictive models are still scarce or relatively recent [29–31]. Our study aims to contribute to this emerging field by exploring the integration of crowdsourced data into air quality modeling thanks to a large data set at hand for the BCR.

In this paper, we leverage an extensive citizen science dataset to train and validate statistical models for estimating mean annual $NO_2$ concentrations over the BCR. Instead of relying on sparse monitoring networks, broad-scale assessments and coarse spatial interpolation, our approach benefits from thousands of measurement locations, enabling a much more precise and spatially detailed assessment of these concentrations. A key contribution of this work is the comparative evaluation of two LUR models using different regression techniques to analyze the relationship between $NO_2$ concentrations and spatial proxies. By contrasting a MLR model with a RFR approach, we provide new insights into the nature of the proxies-$NO_2$ relationships, with a particular focus on how traffic-related variables influence pollution estimates. Our prediction results using these MLR and RFR models are also compared to those provided by the SIRANE model. It is shown how having at hand this extensive citizen science dataset allows us to identify the role that traffic data quality is playing for correctly predicting $NO_2$ concentrations. In general, our findings highlight too the benefit of large-scale, community-driven data collection for enhancing urban air quality modeling and improving the understanding of key pollution drivers.

## 2. Material and methods

### 2.1. Study area

The Brussels-Capital Region (BCR) is located in the center of Belgium. It is the most densely populated region of the country, with a population of 1.25 million inhabitants unevenly spread over an area of 161 $km^2$. This uneven distribution is partly driven by the land use diversity, that ranges from residential (even semirural along the BCR borders) to industrial areas, with a significant amount of green spaces (18.5% of the area) spread throughout the region. As most European cities, it has a dense and complex street network resulting from an urbanization that was historically driven by centripetal population movements. Being the major economic center of the country as well as headquarters of the European institutions, BCR is also subject to an intense commuter traffic. In January 2024, it was considered as one of the top ten most congested cities in the world [33]. Nonetheless, thanks to the ongoing implementation of stricter mobility measures in general [34], major air pollutants concentrations have been declining over time, although $NO_2$ still remains an issue.

### 2.2. CurieuzenAir sampling campaign

The CurieuzenAir project is a large citizen science project that was conducted from September to October 2021, with the aim of measuring $NO_2$ concentrations over the BCR. About 3,000 pairs of Palmes-type tubes [35] were installed on the street-facing external facades of residences, businesses, and schools (see Fig. 1). These tubes are over 7 cm long and are passive samplers containing a chemical that binds to atmospheric $NO_2$ diffusing in the tube over time. After a month, all pairs of tubes were collected and analyzed in laboratory for $NO_2$ concentrations using methods compliant with European Standards, equivalent to NEN EN 16339 [36]. Data quality was ensured using the pairs of tubes at each location, that were used to detect malfunctions or contaminations. Outliers were identified using the interquartile range method for the difference between pairs of tubes. In parallel, blank measurements from unopened tubes were used as reference for a null concentration. Finally, the measurements obtained from additional tubes placed at the same locations than the telemetric stations were used for calibration. Based on the average of the telemetric measurements, two linear regressions were used to convert the initial Palmes tubes measurements to mean annual $NO_2$ concentrations. More technical details about these data

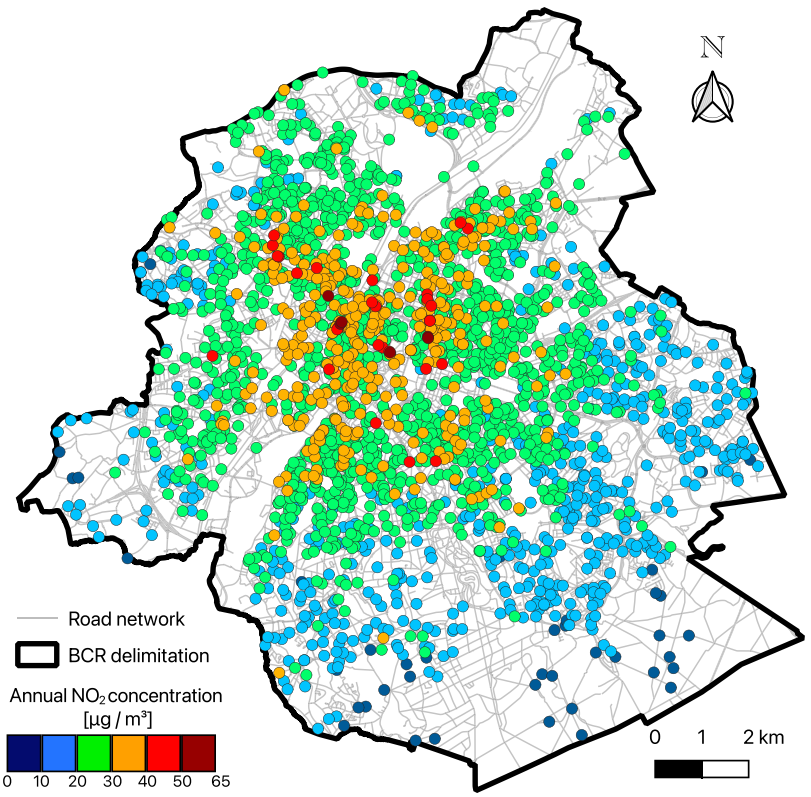


**Fig. 2.** Mean annual $NO_2$ concentrations in 2021 as estimated from the CurieuzenAir campaign ($n = 2,493$).

quality procedures and corrections can be found in [32]. After applying these procedures and corrections, a final set of 2,493 measurements were considered as valid and were converted to mean annual concentrations. The resulting measurements provided an overview of the spatial distribution of the mean annual $NO_2$ concentrations over the BCR, that was part of a subsequent communication campaign (see [32] too).

As seen from Fig. 2, these concentrations reveal a clear concentric pattern across the BCR, with the highest concentrations clustered around the city center and the lowest ones mostly observed in the southern green areas of the region. Local hotspots that are observed outside the central area are consistently located near densely trafficked roads [32]. Based on the distribution of $NO_2$ concentrations (Fig. 3), the estimated annual concentration is about $24.0 \pm 0.1$ µg/m$^3$ across the BCR. Only 1.4% of the measurements ($n = 34$) exceed the annual European regulatory limit of 40 µg/m$^3$ [37].

Although the campaign took place during the COVID-19 pandemic, its impact on traffic levels was minimal. Strict mobility restrictions mainly occurred in 2020 and early 2021, while traffic progressively returned to normal levels afterwards. Data from road-counting cameras in Brussels confirmed this: in October 2021, 1,708,993 vehicles were recorded, a number quite comparable to October 2022 (1,794,971 vehicles) when economic activities fully resumed. In contrast, October 2020 (heavily impacted by COVID-related restrictions) exhibited significantly lower traffic, with only 1,410,647 vehicles. In parallel, the measured mean annual $NO_2$ concentrations for 2020, 2021 and 2022 were respectively equal to 22.0, 24.54 and 24.44 µg/m$^3$. Again, the COVID effect is visible for 2020, but the years 2021 and 2022 are similar. This confirms that the pandemic had no impact on traffic levels and $NO_2$ emissions when the campaign took place.

### 2.3. Spatial proxies

In order to explain the spatial pattern of $NO_2$ concentrations and predict these concentrations at unsampled locations, a set of spatial

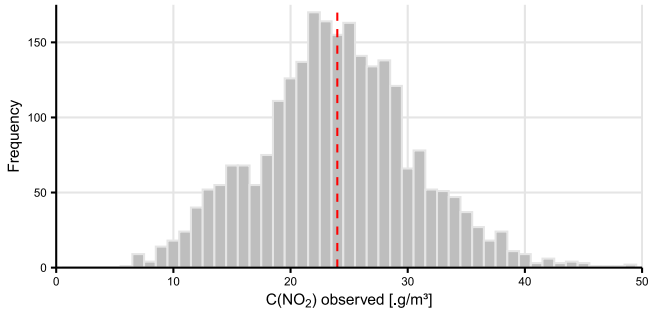


**Fig. 3.** Distribution of $NO_2$ measurements from the CurieuzenAir campaign. The red dotted line indicates the estimated mean. For the sake of readability, 6 measurements (out of the 2.493 measurements) were not displayed as they exceed the maximum axis value of 50 µg/m$^3$.

proxies was gathered at the CurieuzenAir measurement locations. In BCR, $NO_2$ levels are largely driven by traffic emissions, but spatially accurate traffic data for each measurement point during the sampling period were out of reach. A comprehensive set of traffic-related proxies was used instead. The term "proxy" refers here to variables that are not causally linked to these concentrations but that can be used as substitutes as they correlate with the relevant but unobserved causal variables. These include socio-economic variables associated with traffic patterns, urban morphology characteristics, and passive pollutant indicators designed to reflect atmospheric dispersion patterns.

Among these proxies, road traffic deserves a special attention, as it is considered as the major source (about 47%) of nitrogen oxides emissions in the BCR, along with household heating (about 33%). It is thus expected to play a significant role for modeling $NO_2$ concentrations. However, the only official information about traffic comes from apportionment approaches that are based on traffic counts along major roads during rush hours, yielding hourly traffic estimates over any road

**Table 1**
Spatial proxies collected at each CurieuzenAir measurement location. Proxies marked with an asterisk (*) are categorical, while unmarked ones are continuous. Acronyms are BE: Brussels Environment, BM: Brussels Mobility, IBSA: Belgium National Institute for Statistics, UrbIS: IT service of the BCR.

| Proxy | Predictor name | Details/Classes | Unit | Provider |
|---|---|---|---|---|
| Traffic-related proxies | | | | |
| Vehicle count | TRAFFIC_X_Y | Mean or max (Y) on buffer sizes (X) : 10, 100, 500 m at two time intervals : 8–9 am and 5–6 pm | – | Traffic Stats API(Tomtom) |
| Traffic jam index | JAM_X_Y | Mean or max (Y) on buffer sizes (X) : 10, 100, 500 m at two time intervals : 8–9 am and 5–6 pm | – | Traffic Stats API(Tomtom) |
| Road hierarchy* | ROAD_CLASS | Metropolitan road (A1) - Main road (A2) - Inter-district road (A3) - District collector (A4) - Local road (A5) - Paths (B3) | – | BM |
| Speed limit* | SPEEDKMH | 30 – 50 – 70 km/h | – | BM |
| Distance to traffic lights | NEAR_LIGHT | Nearest traffic lights | m | BM |
| Distance to school | NEAR_SCHOOL | Nearest school | m | BM |
| Distance to bike park | NEAR_BIKE | Nearest bike park | m | BM |
| Distance to car park | NEAR_PARK | Nearest car park | m | BM |
| Socioeconomic proxies | | | | |
| Population density | POPDENS21 | By district | – | IBSA |
| Unemployment rate | UNEMPLOY | By district | – | IBSA |
| Cars per household | CAR_HOUSE | By district | – | BM |
| Urban morphology proxies | | | | |
| Distance to the BCR center | NEAR_CENTER | BCR center: Great Market place | km | BM |
| Altitude | ALTITUDE | Meters above sea level | m | UrbIS |
| Land use* | LAND_USE | Green - Residential - Industrial - Mixed areas | – | UrbIS |
| Street canyon index | H_W | Average building height/road width | – | BE |
| Passive pollutant proxies | | | | |
| Black Carbon (BC) | PEAK/BULK_BC | Peak time, bulk time | $\mu g/m^3$ | BE |

segment of the BCR (details not given here), with unassessed precision and accuracy. As this raises concerns about the quality of these data for our modeling purposes, freely available traffic information from TomTom for October 2022 was used instead (as information for October 2021 was not at hand) as an attempt to circumvent this issue. It consists in hourly counts of GPS users over the various road segments of the BCR. Counts during rush hours (8–9 am and 5–6 pm) were averaged during that month and were considered as representative of the mean traffic that should have occurred during October 2021. It is worth noting however that no information is provided by TomTom about the quality of these data and that local differences in road traffic between October 2021 and October 2022 are expected to occur, e.g., due to road repairs, road diversions, etc.

Table 1 gives the list of all proxies considered in this study. Due to differences in spatial resolution and types (categorical vs continuous data), some values were obtained using buffer zones around the measurement locations, while others were assigned based on distance using spatial joining procedure in QGIS (version 3.34.11-Prizren) [38]. For subsequent statistical processing, proxies were standardized (centered and scaled) to facilitate the comparison of their relative impact based on the results coming out from the LUR models.

### 2.4. Statistical procedures

Relating the mean annual $NO_2$ concentrations that have been estimated from the CurieuzenAir campaign with the selected proxies of Table 1 was done using two LUR models that were implemented using RStudio (version 2024.04.2+764) [39]. The simplest one was multiple linear regression (MLR), where each proxy is linearly related to the $NO_2$ concentrations. A more elaborate random forest Regression (RFR) was considered too for capturing potential complex non linear relationships. Both models are viewed as complementary ; the impact of each proxy is directly provided by MLR regression coefficients at the price of a strong assumption of linearity, while RFR is able to handle complex relationships at the price of a more difficult interpretation of the results, thus trading interpretability for flexibility [40]. A comparison of the performance of both models is also a useful way to assess the respective importance of the linear and nonlinear relationships in the modeling process.

#### 2.4.1. Multiple linear regression

Denoting $C_{NO_2}$ as the modeled $NO_2$ concentrations based on a set of $k$ values $x_1, \ldots, x_k$ for the proxies, the MLR model is written as

$$C_{NO_2} = \beta_0 + \sum_{i=1}^{k} \beta_i x_i \quad (1)$$

where the $\beta_i$ regression coefficients are estimated by least squares and are measuring the linear impact of each standardized proxy on the $NO_2$ concentrations. The model was estimated using the standardized ESCAPE (European Study of Cohorts for Air Pollution) methodology [13]. Proxies were first selected through a supervised forward stepwise procedure using the adjusted coefficient of determination $R^2_{adj}$ as measuring the goodness of fit. At each step, the proxy that maximized the increase in $R^2_{adj}$ was added if this increase exceeded 1%. The process continued until no more proxy could be added. At the second stage, proxies with p-values greater than 0.1 were sequentially removed using a backward stepwise procedure, thus leading to a final model where all included proxies are deemed relevant for the prediction. Because collinearity can adversely affect the interpretation of the estimated model, the adjusted generalized variance inflation factors ($GVIF_{adj}$) were computed too. As our proxies consist in a mixture of continuous and categorical variables, these $GVIF_{adj}$ include a correction for the degrees of freedom [41], thus allowing a direct comparison across proxies regardless of their continuous or categorial nature. A $GVIF_{adj}$ that does not exceed 3 is considered as acceptable with respect to collinearity.

#### 2.4.2. Random Forest Regression

Random Forest Regression (RFR) is a commonly-used machine learning algorithm based on the combination of multiple decision trees built from random subsets of data and proxies, where the final predictions is the averaged prediction from these trees. It can handle complex nonlinear relationships and is not prone to overfitting as long as proper care is taken for the selection of the hyperparameters. See, e.g. [40] for a detailed discussion about the algorithm. The model was implemented using the ranger library [42]. The three hyperparameters that were tuned are the number of trees, the tree depth (controlled via the minimum node size), and the size of the predictor subset used at

each split. The tuning process was conducted using a grid search on a 10% subset of the dataset ($n = 247$). Identified optimal values were 500 trees and a minimum node size of 8. These values were used for the proxy selection process as well as for the final model. Meanwhile, the size of the predictor subset at each split was dynamically set to its standard value [43].

While proxies for MLR were chosen through a forward stepwise approach, the RFR model relied on a supervised backward stepwise method, again using $R^2$ as measuring the goodness of fit and where proxies are iteratively removed based on their relative importance [43]. Based on the final set of influential proxies, the model was then trained on a subset of 70% of the data ($n = 1,745$) and applied on the remaining 30% ($n = 748$) to assess its performance.

#### 2.4.3. Performance comparison of MLR and RFR models

For the sake of comparison, the performances of the MLR and RFR models were assessed using the same test dataset that was used for the RFR validation ($n = 748$). Three metrics were considered for measuring the MLR and RFR predictive performances, according to the recommendation by [44] for air quality model predictions. These metrics are :

(i) The coefficient of determination ($R^2$), that corresponds to the fraction of the variance of the original data which is explained by the model. It is worth noting that $R^2$ was used here instead of the adjusted value $R^2_{\text{adj}}$, which cannot be defined for the RFR model as it has no associated degrees of freedom ;
(ii) The root mean square error (RMSE), which is the square root of the variance of the residuals, that accounts for both systemic and random errors ;
(iii) The fractional bias (FB), which only focuses on the systemic error, with

$$\text{FB} = \frac{2\left(\overline{P} - \overline{O}\right)}{\overline{P} + \overline{O}} \tag{2}$$

where $\overline{P}$ is the mean of the predicted concentrations using the model and $\overline{O}$ is the mean of the measured ones.

#### 2.4.4. MLR as benchmark for other LUR models

It is clear that RFR is superior by design to MLR in terms of versatility, as it does not rely on a constraining parametric formula as given by Eq. (1). It can also handle arbitrarily complex nonlinear effects that could be needed to account for obtaining an efficient predictive model. However, as MLR is only accounting for strictly affine relationships between proxies and $NO_2$ concentrations, it can advantageously be used as a benchmark for quantifying the added value of using more complex nonlinear relationships, and in particular for assessing the importance that these non linearities play in the performance of other LUR models.

So far, MLR and RFR were presented as separate regression models, with their dedicated selection procedures for identifying the proxies that each model needs to account for maximizing its predictive performances. However, if the goal is a fair assessment of the benefit of RFR over MLR, it is needed to use to same set of proxies for comparing them on the same ground. For doing so, we considered two additional models in this paper, namely a RFR model that uses the same proxies than those identified as relevant by MLR (denoted RFRC hereafter), and symmetrically a MLR model that uses the same proxies than those identified as relevant by RFR (denoted MLRC hereafter). Both RFRC and MLRC were of course re-estimated using the corresponding new set of proxies, so they can be considered as the most efficient MLR and RFR models conditionally to these imposed proxies.

Comparing now the performances of MLR and RFRC along with the performances of MLRC an RFR allows us to assess the benefit of accounting for non linearities in the random forest regression, even if the proxies that are used might be suboptimal either for the multivariate linear or random forest regression.

#### 2.4.5. Spatial correlation of the residuals

An efficient regression model yields regression residuals that have a low variance by comparison with the variance of the original data. At the same time, these residuals are expected to be spatially uncorrelated, i.e. they correspond to random effects that are spatially unpredictable. Indeed, if a significant spatial correlation still remain in these residuals, it means that there are some spatial effects that have been overlooked during the modeling process, that would have improved the performance of the model if they were accounted for.

Assessing the spatial dependence of the regression residuals can be done either using the spatial covariance function $C(\|\mathbf{h}\|)$ or the variogram $\gamma(\|\mathbf{h}\|)$, with

$$\gamma(\|\mathbf{h}\|) = \frac{1}{2} Var[\varepsilon(\mathbf{x}+\mathbf{h}) - \varepsilon(\mathbf{x})] \tag{3}$$

that measures the dissimilarity between regression residuals $\varepsilon(.)$ at locations $\mathbf{x}$ and $\mathbf{x} + \mathbf{h}$ as a function of the Euclidean distance $\|\mathbf{h}\|$ that separates these locations. In absence of spatial correlation, this variogram corresponds to an horizontal line (i.e., a pure nugget effect) as a function of $\|\mathbf{h}\|$, with $\gamma(\|\mathbf{h}\|) = \sigma^2_\varepsilon \quad \forall \ \|\mathbf{h}\| \neq 0$, where $\sigma^2_\varepsilon$ is the variance of the residuals. More details about variograms, their estimation and their modeling can be found in [45].

The variograms of the residuals were estimated both for MLR and RFR, and they were modeled using an exponential model given by

$$\gamma(||\mathbf{h}||) = \alpha_1 + \alpha_2\left(1 - \exp\left(-3\frac{||\mathbf{h}||}{r}\right)\right) \tag{4}$$

where $\sigma^2_\varepsilon = \alpha_1 + \alpha_2$ and where the second term is measuring the spatial dependence of the residuals. In absence of spatial dependence, we thus have $\sigma^2_\varepsilon = \alpha_1$ and $\alpha_2 = 0$. The ratio $\alpha_2/(\alpha_1 + \alpha_2) \in [0, 1]$ can be used as a measure of the fraction of spatial information still left in the residuals, while the $r$ parameter is measuring the spatial extent of the dependence between residuals.

### 2.5. Performance comparison with the SIRANE model

Thanks to the development of computational capacities, spatially distributed physics-based air quality models are becoming more and more accessible for operational use. The relevance of using a large citizen science dataset for air quality modeling based on a LUR approach could thus be questioned. In this context, we decided to compare our results to those obtained from the SIRANE model [46], that simulates the dispersion of pollutants within and above the urban canopy and that provides hourly average concentration of pollutants for each street segment in the BCR. SIRANE is making use of a street network approach by handling separately point, linear and surface emission sources. The model also includes a meteorological preprocessor along with a photochemical module for $NO_x/O_3$ reactions. SIRANE has been validated both in wind tunnel [47] and for real case studies [48–50] and was shown to adequately predict pollutant concentrations in urban domains, provided that input data (especially traffic emissions) are valid [46,48].

When using SIRANE in the BCR, $NO_2$ emissions are first obtained from the COPERT model [51], commonly used in Europe for annual pollutant emissions inventory, with a subsequent apportionment of these emissions over the road segments in the BCR using a gravitational model that accounts for traffic counts along major roads during rush hours (details not given here). Road traffic on other road segments and for other hours of the day are then extrapolated from additional periodic counting and socio-economic data. As a result, it is expected that the performance of the SIRANE model is impacted by the quality of these extrapolations for the traffic data. So far, the quality of the SIRANE predictions was not really assessed, as the only comparison between predictions and measurements is at locations that correspond to the few telemetric stations that were used for calibrating the model. It is thus interesting to compare our MLR/RFR prediction results to those obtained from SIRANE when it comes to mapping the spatial

**Table 2**
Relevant proxies for the MLR and RFR models (two last columns) based on their respective selection procedure, along with the determination coefficients ($R^2$) for the univariate linear regression models relating $NO_2$ concentrations to each proxy given in Table 1. Proxies are sorted in decreasing order of $R^2$ values. The $GVIF_{adj}$ values are measuring collinearity between proxies and are adjusted for the degrees of freedom.

| Proxy | $R^2$ | $GVIF_{adj}$ | MLR $GVIF_{adj}$ | RFR $GVIF_{adj}$ |
|---|---|---|---|---|
| NEAR_CENTER | 0.52 | 3.30 | 1.70 | 2.94 |
| NEAR_PARK | 0.47 | 2.61 | | 2.47 |
| CAR_HOUSE | 0.38 | 1.72 | | 1.62 |
| POPDENS21 | 0.37 | 2.00 | 1.56 | 1.70 |
| UNEMPLOY | 0.28 | 1.85 | | 1.69 |
| LAND_USE | 0.27 | 1.51 | | |
| ALTITUDE | 0.24 | 1.56 | 1.16 | 1.50 |
| ROAD_CLASS | 0.15 | 1.25 | 1.04 | 1.09 |
| NEAR_BIKE | 0.14 | 1.66 | | |
| NEAR_TRAIN | 0.12 | 1.26 | | |
| NEAR_LIGHT | 0.12 | 1.26 | | |
| TRAFFIC_100_MEAN | 0.11 | 1.37 | 1.19 | 1.24 |
| TRAFFIC_10_MAX | 0.09 | 1.51 | | 1.46 |
| NEAR_SCHOOL | 0.07 | 1.36 | | |
| H_W | 0.07 | 1.31 | | |
| PEAK_BC | 0.07 | 5.59 | | |
| BULK_BC | 0.06 | 5.49 | | |
| TRAFFIC_500_MEAN | 0.06 | 1.32 | | |
| JAM_100_MEAN | 0.007 | 1.28 | | |
| JAM_10_MAX | 0.007 | 1.21 | | |
| SPEEDKMH | 0.006 | 1.08 | | |
| JAM_500_MEAN | 0.007 | 1.44 | | |

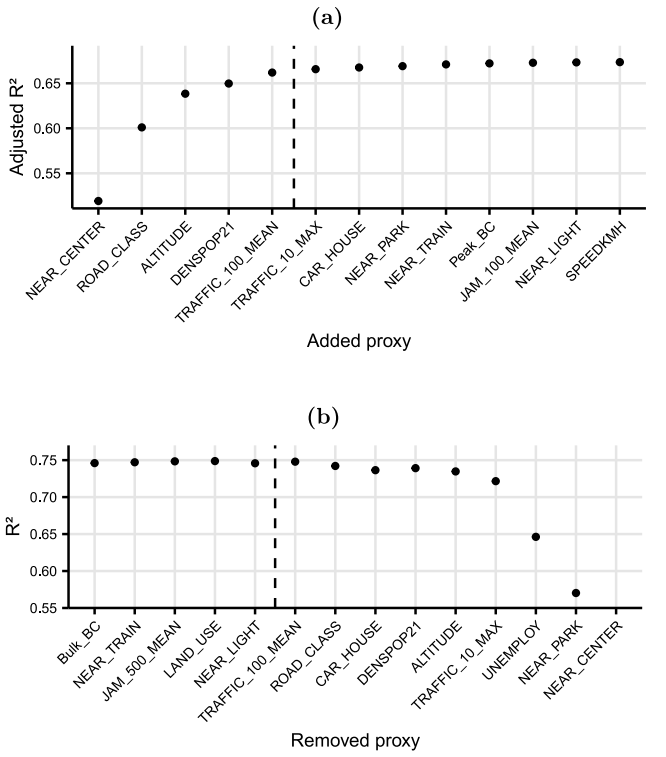


**Fig. 4.** Proxies selection (a) for the MLR model using a supervised forward selection and (b) for the RFR model using a supervised backward selection. The $R^2_{adj}$ and $R^2$ values were used as measure of goodness-of-fit, respectively. The dashed black line represents the cutoff, so that relevant proxies are on the left for (a) and on the right for (b).

distribution of the mean annual $NO_2$ concentrations over the BCR. Moreover, as the quality of traffic data is known to be a potential issue when predicting $NO_2$ concentrations, its impact can also be assessed thanks to the large dataset used in our study.

## 3. Results

### 3.1. Proxies selection

Based on their respective selection procedures, Table 2 presents the proxies that were considered as relevant for the MLR and RFR models. The individual predictive capacity of each proxy was also assessed through the $R^2$ values for the univariate linear regressions (continuous proxies) and one-way ANOVA (categorical proxies) that relate $NO_2$ concentrations to one proxy at a time.

Out of the 22 proxies at hand that are listed in Table 2, it can be seen that MLR and RFR select a subset of 5 and 9 of them, respectively, and that all proxies selected for MLR were also selected for RFR. This emphasizes the consistency of our results, where RFR is able to account for more complex relationships and thus includes additional relevant proxies, without discarding those that were identified as relevant by MLR. As shown from the ordering of the proxies with respect to $R^2$ in Table 2, it can also be seen that the selected proxies are not necessarily those that are individually the most strongly related to $NO_2$ concentrations. This can be explained by collinearities between some proxies. For instance, NEAR_CENTER (the most relevant proxy for MLR as seen from Fig. 4a) is positively correlated both with NEAR_PARK ($R^2 = 0.91$) and CAR_HOUSE ($R^2 = 0.78$), while ALTITUDE (third proxy selected for MLR) is negatively correlated with UNEMPLOY ($R^2 = -0.70$). When it comes to the RFR model, it can been seen that NEAR_PARK (ranked second with respect to $R^2$) that was not selected by MLR is now accounted for. Similarly, UNEMPLOY that was linearly related to ALTITUDE (as lowest altitudes correspond to areas along the Charleroi canal, which are the poorest in the BCR) is now included too, along with other additional proxies. As RFR is able to handle nonlinear relationships as well, these are expected results as more subtle effects can be evidenced. The results of the proxies selection for RFR are presented in Fig. 4b. With respect to collinearity, results in Table 2 do not indicate major issues for the MLR and RFR models (the case of NEAR_CENTER that has the highest $GVIF_{adj}$ value is discussed later on).

An unexpected result for MLR is that traffic-related proxies are not highly relevant for the model. Given the fact that road traffic is the major anthropogenic source of $NO_2$ in the BCR, this is surprising. This could be linked to potential inaccuracies in the TomTom data that were used for building the various TRAFFIC and JAM proxies presented in Table 1, as data for October 2022 where used as an approximation of those for October 2021. This specific hypothesis will be investigated later on by comparing our results with those obtained from the SIRANE model. Another explanation is the fact that the NEAR_CENTER proxy already conveys part of the traffic-related information, as discussed above for its correlation with the NEAR_PARK and CAR_HOUSE proxies. Finally, potential nonlinear relationships between $NO_2$ concentrations and traffic-related proxies cannot be accounted for through the simple expression of the MLR model given by Eq. (1), as already mentioned. Previous studies support the idea that $NO_2$ levels and vehicle counts are not necessarily linearly related [52] or that proxies involving the ratio of traffic volume to meteorological variables better relate to these levels [53], though we were not able to evidenced similar effects based on our data set.

Beside this issue, from a comparison with the literature, our MLR results are in agreement with the most commonly used proxies for predicting air pollution [13], e.g., by evidencing among others the effect of population density and altitude. The use of NEAR_CENTER as a proxy is less common, probably as it does not convey any causal explanation about the observed $NO_2$ concentrations. However, it is statistically the most useful one for predicting them using MLR, as seen from Fig. 4a. The idea of using proxies based on distances can be found in [14,15], were distances to influential features (bridges,

**Table 3**
Regression coefficients estimates for the MLR model, along with their p-values. All proxies are standardized and are unitless. Descriptions of road class categories can be found in Table 1, and A1 was used as the reference category using a dummy coding.

| Proxy | Coefficient | p_value |
|---|---|---|
| Intercept | 28.7 | <2e−16 |
| ROAD_CLASS: A2 | −1.29 | 0.064 |
| ROAD_CLASS: A3 | −1.94 | 1e−03 |
| ROAD_CLASS: A4 | −3.64 | 2e−09 |
| ROAD_CLASS: A5 | −5.80 | <2e−16 |
| ROAD_CLASS: B3 | −6.25 | 5e−07 |
| NEAR_CENTER | −2.81 | <2e−16 |
| ALTITUDE | −1.63 | <2e−16 |
| POPDENS21 | 1.34 | <2e−16 |
| TRAFFIC_100_MEAN | 0.93 | <2e−16 |
| $R^2$ [%] | | 66.2 |
| $R^2_{adj}$ [%] | | 65.6 |

**Table 4**
Relative importance of each proxy selected for the RFR model, as computed from the test dataset ($n$ = 748).

| Proxy | Importance (%) |
|---|---|
| NEAR_CENTER | 15.6 |
| NEAR_PARK | 8.53 |
| ALTITUDE | 5.89 |
| POPDENS21 | 4.77 |
| UNEMPLOY | 4.76 |
| CAR_HOUSE | 4.32 |
| TRAFFIC_10_MAX | 4.15 |
| ROAD_CLASS | 3.60 |
| TRAFFIC_100_MEAN | 2.26 |
| $R^2$ [%] | 74.6 |

**Table 5**
Performance metrics for the various regression models as estimated from the test data set ($n$ = 748), where MLR and RFR refer to the multiple linear and random forest regression models using their specific set of selected proxies, while MLRC and RFRC refer to these models with constrained set of proxies as explained in Section 2.4.4.

| Model | # proxies | $R^2$ | RMSE | FB |
|---|---|---|---|---|
| MLR | 5 | 66.2 | 3.87 | −0.00406 |
| RFR | 9 | 74.6 | 3.37 | −0.00588 |
| MLRC | 9 | 66.6 | 3.85 | −0.00282 |
| RFRC | 5 | 69.9 | 3.65 | −0.00701 |

borders, etc.) have advantageously been used. In our study, the reason of including it is directly linked to the urban morphology. The BCR has a densely populated inner area characterized by intense economic activities and heavy traffic, where narrow canyon-like streets are common [16] and limit atmospheric dispersion, thus further elevating $NO_2$ concentrations. As a result, NEAR_CENTER summarizes all these effects that are acting jointly and that are thus hard to disentangle. This is also clear from its highest $GVIF_{adj}$ value among all proxies in the MLR and RFR models.

### 3.2. Comparison of MLR and RFR results

The estimated regression coefficients for the proxies selected by the MLR model are given in Table 3 and are in close agreement with what is expected from our understanding of the processes. It can be seen that NEAR_CENTER has a negative regression coefficient on $NO_2$ concentrations, indicating higher levels in the city center, while POPDENS21 and TRAFFIC_10_MEAN have a positive one. Concentrations tend to decrease with ALTITUDE, mainly due to the fact that high altitude are located in the greener and less populated southeast part of the region. For ROAD_CLASS, one can see that a lower road hierarchy corresponds to lower concentrations, as expected too.

The spatial distribution of the mean annual $NO_2$ concentrations predicted from the MLR model at the street segments level across the BCR is shown in Fig. 5. Higher concentration close to the center of the region and along the major traffic axes are easily identified and are the result of the combined effect of the various proxies listed in Table 3.

For RFR, Table 4 provides the relative importance of each proxy included in the model. As for MLR, the most relevant one is NEAR_CENTER, and the first five proxies having the highest $R^2$ values in Table 2 are now selected. While TRAFFIC_10_MAX was not included in the MLR model, it is now part of the RFR model (see Figs. 4a and b).

As expected, a comparison of the predictive performances of both models (first two lines of Table 5) shows that RFR outperforms MLR, but the improvement is not significant. For MLR, the $R^2$ value is within the range of 56% to 90% as reported by [13] for predicting air pollutant concentrations, while it agrees with the range of 60% to 85% as reported by [17,18] for $NO_2$ predictions using RFR. It is worth noting too that using RFR requires information about additional proxies, and is thus substantially more demanding with respect to data collection. Indeed, if the goal is to make predictions using the RFR model at arbitrary locations in the BCR, the values of these proxies need to be collected for all road segments, and not only for those associated with the measurements that are used for estimating the model.

If neither MLR nor RFR exhibit a systematic bias as indicated by their low FB values, both tend to underestimate high $NO_2$ concentrations, as seen from Fig. 6. This underestimation might be severe above 30 μg/m$^3$ and put at stake the ability of the model to accurately identify hotspots at a high spatial resolution. As EU regulation currently requires that mean annual $NO_2$ concentrations do not exceed 40 μg/m$^3$ [37], this raise concerns about the use of such models for assessing compliance with air quality standards at the very local scale. However, these underestimations are likely due to local effects that are by essence difficult to model based on the spatial resolution of the proxies at hand. Since the $NO_2$ data set consists of point measurements at street level, unpredictable variations occurring over few meters are expected due to very local effects. This limitation has been highlighted in studies on land-use regression models, where peak $NO_2$ concentrations in high-traffic areas were often underestimated [54]. This requires further work for a better characterization of high concentrations and extreme pollution events. As suggested by [17], a possibility to address this issue would be to build distinct models when handling low-to-moderate and high concentrations, respectively. However, from an human exposure viewpoint, mapping $NO_2$ concentrations at such a high resolution does not make much sense, as individuals are expected to move around in the city instead of staying at these hotspots locations for extended period of time.

As explained in Section 2.4.4, it is also useful to compare MLR and RFR on the same ground, i.e. by using both models with the same set of proxies. The corresponding results are given in Table 5, where MLRC refers to the multiple linear model that uses the optimal set of proxies for RFR, while RFRC refers to the random forest model using the optimal set of proxies for MLR.

A comparison of MLR with MLRC shows that the proxies selection procedure that identified five of them as relevant did a good job, as adding four additional ones (i.e., those identified as relevant for RFR) does not improve the performance of the model, with almost unchanged value for $R^2$, RMSE and FB. However, these four proxies convoyed useful nonlinear information, as shows by a comparison of RFRC with RFR, where the improvement is clear for $R^2$ and RMSE. In parallel, it can be seen that the performance of RFRC is only marginally better than for MLR. This seems to indicate that the five proxies selected for MLR mainly relate linearly to $NO_2$ levels (as using RFRC does not

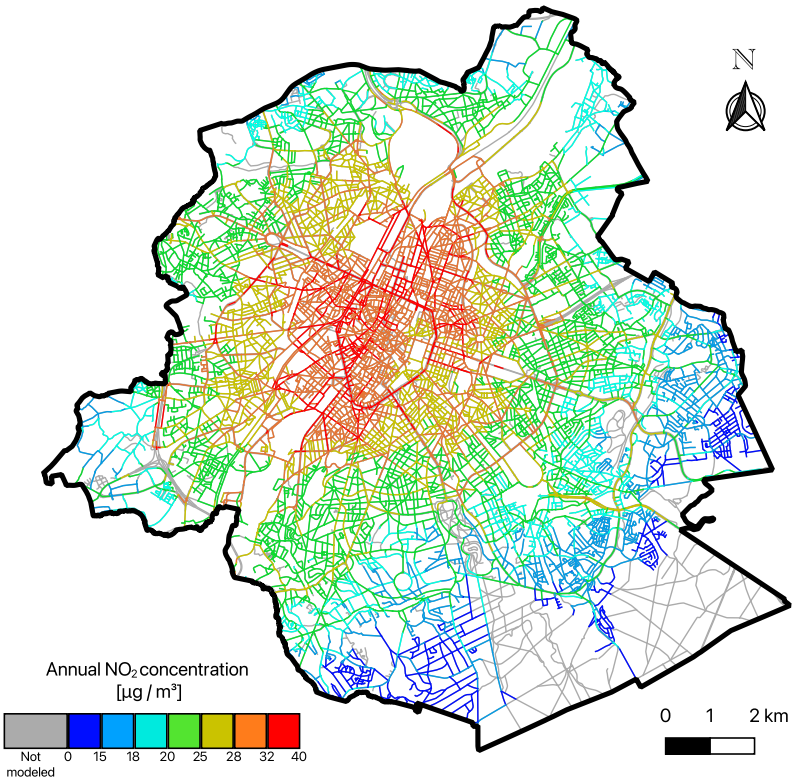


**Fig. 5.** Map of the predicted mean annual $NO_2$ concentrations at the street segments level using the MLR model.

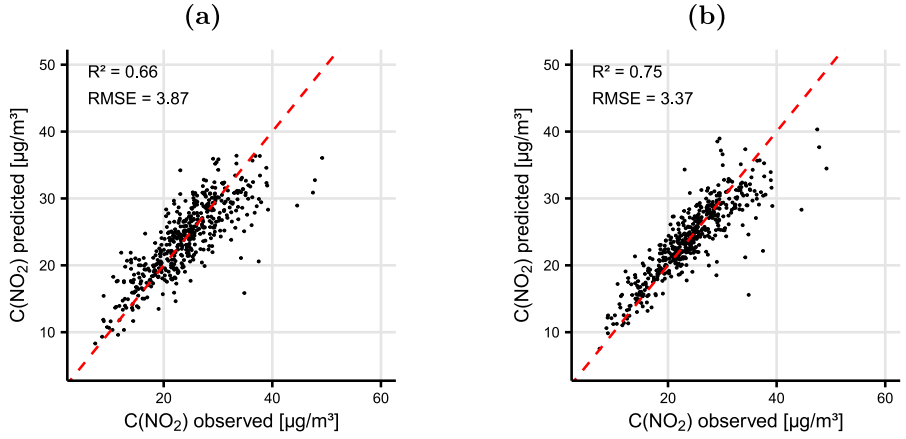


**Fig. 6.** Comparison between observed and predicted annual $NO_2$ concentrations for the test dataset ($n = 748$) using (a) the MLR model and (b) the RFR model. Values on the 1:1 red dashed line correspond to perfect predictions.

improve much the performance), while the additional four proxies for RFR mainly refer to nonlinear relationships (as using MLRC instead leads to lower performances, that are close to MLR). Interestingly, three of these four proxies are related to traffic, as seen from Table 2. This suggests that even if $NO_2$ concentrations in BCR are known to be traffic-driven, evidencing this is uneasy and requires to account for nonlinear relationships with traffic-related proxies.

### 3.3. Spatial correlation of the residuals

The spatial dependence of regression residuals for MLR and RFR models was estimated using variograms. For both models, the spatial extent of this dependence is close to $r = 850$ m as seen from Fig. 7. However, the variance $\sigma_\varepsilon^2 = \alpha_1 + \alpha_2$ of RFR residuals is lower than for MLR, in agreement with the higher $R^2$ of RFR compared to MLR. In addition, it can be seen than the spatially correlated part of this residual variance as measured by $\alpha_2/(\alpha_1 + \alpha_2)$ is lower for RFR (56.7%) than for MLR (81.5%). This suggests that RFR extracts in a better way the spatial information conveyed by the proxies, thanks to its ability to capture nonlinear relationships. However, even if RFR is doing a better job as emphasized by a higher $R^2$, a lower $\sigma_\varepsilon^2$ and a lower $\alpha_2/(\alpha_1 + \alpha_2)$, it remains true that the model is far from precisely predicting $NO_2$ concentrations, with accuracy issues for high concentrations that are largely underestimated. This highlights the need for better predictions using more relevant proxies, possibly obtained at a finer spatial resolution.

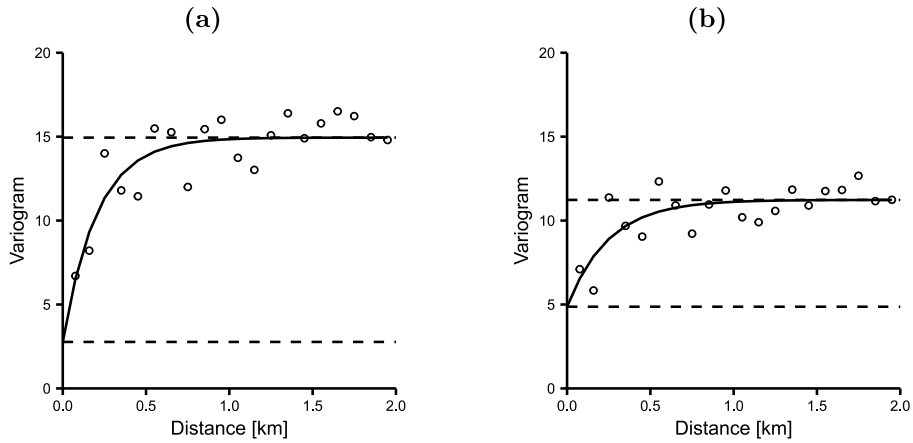


**Fig. 7.** Variograms for the regression residuals (a) for the MLR model and (b) for the RFR model using the test data set ($n$ = 748). The lower dashed line corresponds to $\alpha_1$, i.e., the nugget effect. The upper dashed line corresponds to $\alpha_1 + \alpha_2$, i.e. the variance of the residuals $\sigma_\varepsilon^2$.

### *3.4. Comparing with SIRANE results*

In order to assess the performance of our MLR/RFR models in comparison with those of a near-real time physics-based model, SIRANE was run for October 2021 and its predictions were averaged over time at the same locations than those of the CuriuzenAir campaign. It is worth remembering that SIRANE is calibrated using the telemetric stations as well, but that traffic data used in the model differ from those we used in our study.

The comparison between CurieuzenAir data and time-averaged SIRANE predictions is shown in Fig. 8a. Although the quality of these predictions are comparable to those obtained using SIRANE in Torino [55], it can be seen that the SIRANE performances are lower than for our MLR/RFR models as presented in Figs. 6a and b. It can be concluded that, as far as one is interested by mapping the average $NO_2$ concentrations over the BCR, the MLR/RFR approach is relevant compared to more physics-based ones, as illustred here using SIRANE.

Thanks to the large size of the CurieuzenAir dataset, it is also interesting to investigate the discrepancies between measured and predicted values in Fig. 8a. One hypothesis is that the quality of the traffic data could partly be the reason. To check this, Fig. 8a was split in two parts, with Fig. 8b corresponding to locations where traffic data are considered as reliable because they are close to main roads, while Fig. 8c corresponds to locations in the secondary road network, with lower quality traffic information that mostly comes from extrapolations (details not given here). It can be seen that the main part of the locations where largest discrepancies occur are indeed associated with low quality traffic data, even if this does not explain all of them.

As TomTom traffic data are used in this study, it is of course difficult to further compare these results with ours, but the few large discrepancies between observed and predicted concentrations in Figs. 6a and 6b could be linked to the use of traffic data for October 2022, while data for October 2021 should have been ideally used. As traffic is known to be a major source of $NO_2$ emissions and as the quality of traffic data does indeed impact the predictions (as seen above from the SIRANE results), this could also explain the conterintuitive results obtained for the TRAFFIC proxies, that were not identified as particularly important in our LUR models (see Table 4).

## 4. Conclusions

Based on a large set of $NO_2$ concentrations collected during a citizen science campaign over the BCR, this study investigated the prediction of these concentrations by using two LUR models, namely MLR and RFR. Results have shown better performances for RFR thanks to its ability to handle nonlinear relationships between $NO_2$ concentrations and the spatial proxies we were able to collect over the BCR. As a consequence of the urban morphology of the BCR, the distance to the city center appears to be the most useful proxy for both models, as it can be viewed as a synthesis of several effects that are acting jointly. On the opposite, the benefit of using traffic-related proxies is rather limited, despite the fact that traffic is known to be the main source of $NO_2$ emissions in the region.

Even if the performances of MLR and RFR models are globally on par with the results that are found in the literature, both models tend to underestimate high concentrations. At the very local scale, this raises issues about the use of these models for assessing compliance with air quality standards as set by EU regulations. However, from an human exposure viewpoint, mapping street pollutant concentrations at a very high spatial resolution does not really make sense, as exposure is the combination of largely varying exposures associated with the various locations that an individual is visiting in the region (see, e.g., [56]).

Nevertheless, it remains true that further work would be helpful for handling in a better way these high concentrations and hotspot effects, possibly by considering a specific model for them. Indeed, differences in $NO_2$ concentration of 25 $\mu g/m^3$ over an approximate distance of only 30 m can be observed, thus underscoring that local effects have a substantial impact and strongly depend on the intensity of both local emissions and local ventilation patterns. It is unlikely that these effects can be captured by proxies that are defined at a broader spatial scale. Based on our preliminary investigations in the field (results not detailed in this paper), examples of identified local emission effects are, e.g., (i) the proximity of a tunnel exit, where it had been previously verified that concentrations are peaking, or (ii) the proximity of a traffic light causing a long line of stationary vehicles, especially during rush hours [57]. Such local effects are not captured by the proxies that were at hand for the LUR models, but even the SIRANE model that accounts for some of them (like tunnel exits) has a hard time predicting the correct concentrations at those locations. For other locations, however, no obvious reasons could be identified from a simple visual examination of the street, and it cannot be excluded that poor estimates of traffic could be the reason too. So far, no general conclusion can be drawn, and the best strategy would be to investigate first about these causes (that are likely to be multifactorial) with the aim of including the corresponding high-resolution proxies in a refined version of the LUR models, as long as such a detailed spatial information is at hand over the BCR.

Our results show that MLR and RFR models outperform a physics-based model like SIRANE when it comes to spatially predicting the mean $NO_2$ concentration over the BCR, such predictions being relevant for long-term exposure studies. However, for predicting these concentrations over time, physics-based models like SIRANE are the way to

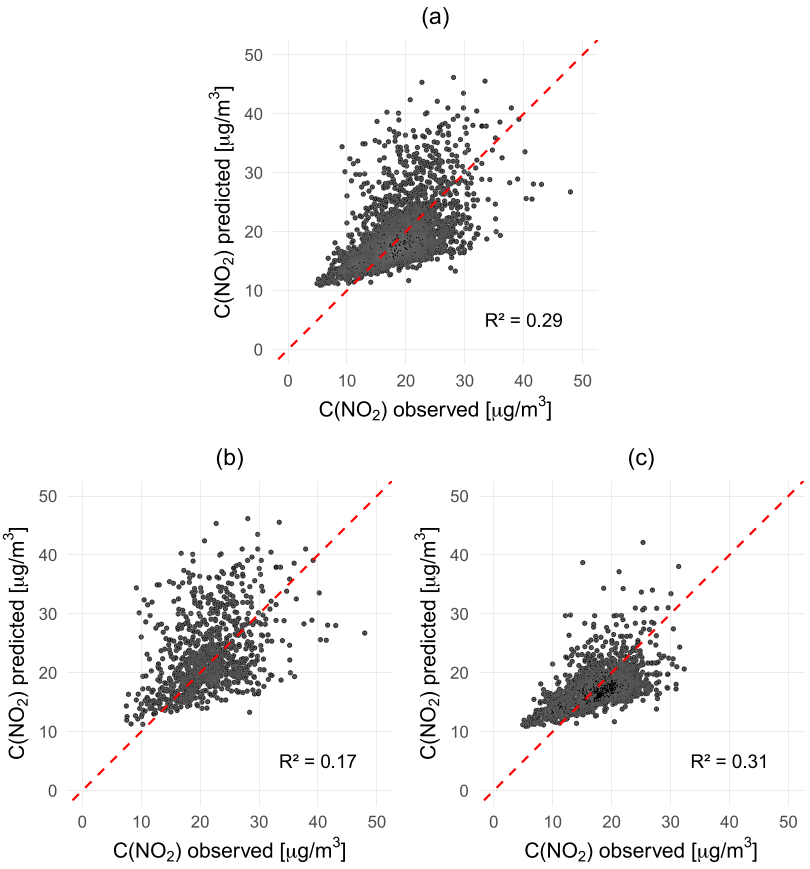


**Fig. 8.** Comparison between measured $NO_2$ concentrations from CurieuzenAir and predicted mean $NO_2$ concentrations from SIRANE model over the same period. Part (a) is using the whole CurieuzenAir data set ($n = 2,493$), while (b) and (c) correspond to locations with low quality and high quality traffic input data for SIRANE, respectively. Values on the 1:1 red dashed line correspond to perfect predictions.

go, as they account for dynamic variables (i.e., weather conditions and traffic) that are known to strongly impacting $NO_2$ concentrations over time. So, even if our MLR and RFR models proved to be relevant for getting a spatial overview of the mean $NO_2$ concentrations over the BCR, they are clearly not a substitute for dynamic models. Indeed, the SIRANE and LUR models are fundamentally different. SIRANE is a dynamic model that aims at predicting NO2 concentrations on an hourly basis, by making use of both dynamic and static inputs. Its goal when used in the BCR is to make near real-time predictions and to assess the impact that various scenarios have on these concentrations (e.g., when implementing a low emission zone). On the other hand, the LUR model is a static model that only makes use of static inputs that are proxies. Its goal in this paper was to identify the most useful proxies for predicting the average annual concentrations (and not the near real-time concentrations) and to evaluate the respective importance of these proxies.

Despite these limitations, this study emphasizes the benefit of using a cost-affordable citizen-based campaign for modeling the spatial distribution of air pollutants in a city. By providing a large amount of data, it enhances our understanding of this distribution and it provides valuable insights about the relationships of pollutant concentrations with proxies that can be collected over the city. Though our study was specifically conducted for $NO_2$ in the BCR, it is believed that the methodology can be used for other cities and possibly for other air pollutants as well.

## CRediT authorship contribution statement

**Patrick Bogaert:** Writing – review & editing, Validation, Supervision, Funding acquisition. **Noémie Huvelle:** Writing – original draft, Methodology, Formal analysis. **Axel Briffault:** Writing – review & editing, Supervision, Conceptualization. **Olivier Brasseur:** Writing – review & editing, Supervision, Project administration, Conceptualization.

## 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.

## Acknowledgment

The authors are grateful to Pr. Filip Meysman for providing them with the CurieuzenAir data set, as well as to Gauthier Koten for an earlier processing of this data set. This study and the CurieuzenAir project were supported by fundings from Bloomberg Philanthropies.

## Data availability

The authors do not have permission to share data.

## References

[1] European Environment Agency. Health impacts of air pollution in Europe. 2022.
[2] Jarvis Debbie J, Adamkiewicz Gary, Heroux Marie-Eve, Rapp Regula, Kelly Frank J. Nitrogen dioxide. In: WHO guidelines for indoor air quality: selected pollutants. World Health Organization; 2010.
[3] Huang Shiwen, Li Haomin, Wang Mingrui, Qian Yaoyao, Steenland Kyle, Caudle William Michael, et al. Long-term exposure to nitrogen dioxide and mortality: A systematic review and meta-analysis. Sci Total Environ 2021;776:145968.

[4] Stieb David M, Berjawi Rania, Emode Monica, Zheng Carine, Salama Dina, Hocking Robyn, et al. Systematic review and meta-analysis of cohort studies of long term outdoor nitrogen dioxide exposure and mortality. PLoS One 2021;16(2):e0246451.

[5] Huangfu Peijue, Atkinson Richard. Long-term exposure to $NO_2$ and $O_3$ and all-cause and respiratory mortality: A systematic review and meta-analysis. Environ Int 2020;144:105998.

[6] Menezes Filipe, Popowicz Grzegorz Maria. Acid rain and flue gas: Quantum chemical hydrolysis of $NO_2$. ChemPhysChem 2022;23(21):e202200395.

[7] Sciensano. Determinants of health: Air quality. 2024, Health Status Report.

[8] Voordeckers D, Meysman FJR, Billen P, Tytgat T, Van Acker M. The impact of street canyon morphology and traffic volume on $NO_2$ values in the street canyons of Antwerp. Build Environ 2021;197:107825.

[9] Goutham Priya M, Jayalakshmi S. Assessment of air pollution modelling using spatial interpolation techniques. Asian Rev Civ Eng 2018;7(2):1–6.

[10] Janssen Stijn, Dumont Gerwin, Fierens Frans, Mensink Clemens. Spatial interpolation of air pollution measurements using CORINE land cover data. Atmos Environ 2008;42(20):4884–903.

[11] Lotrecchiano Nicoletta, Barletta Diego, Poletto Massimo, Sofia Daniele. Comparison of spatial interpolation techniques for innovative air quality monitoring systems. Results Eng 2023;17.

[12] Munir Said, Mayfield Martin, Coca Daniel. Understanding spatial variability of $NO_2$ in urban areas using spatial modelling and data fusion approaches. Atmosphere 2021;12(2):179.

[13] Hoek Gerard, Beelen Rob, de Hoogh Kees, Vienneau Danielle, Gulliver John, Fischer Paul, et al. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmos Environ 2008;42(33):7561–78.

[14] Wheeler Amanda J, Smith-Doiron Marc, Xu Xiaohong, Gilbert Nicolas L, Brook Jeffrey R. Intra-urban variability of air pollution in Windsor, Ontario—Measurement and modeling for human exposure assessment. Environ Res 2008;106(1):7–16.

[15] Gonzales M, Qualls C, Hudgens EE, Neas L. Characterization of a spatial gradient of nitrogen dioxide across a United States-Mexico border city during winter. Sci Total Environ 2005;337(1–3):163–73.

[16] Guyot Madeleine, Araldi Alessandro, Fusco Giovanni, Thomas Isabelle. The urban form of Brussels from the street perspective: The role of vegetation in the definition of the urban fabric. Landsc Urban Plan 2021;205:103947.

[17] Kamińska Joanna A. A random forest partition model for predicting $NO_2$ concentrations from traffic flow and meteorological conditions. Sci Total Environ 2019;651:475–83.

[18] Chen Jiaxin, et al. Predicting particulate matter, nitrogen dioxide, and ozone across Great Britain with high spatiotemporal resolution based on random forest models. Sci Total Environ 2024;926:171831.

[19] Bhardwaj R, Pruthi D. Development of model for sustainable nitrogen dioxide prediction using neuronal networks. Int J Environ Sci Technol 2020;17(5):2783–92.

[20] Dragomir Carmelia Mariana, Voiculescu Mirela, Constantin Daniel-Eduard, Georgescu Lucian Puiu. Prediction of the $NO_2$ concentration data in an urban area using multiple regression and neuronal networks. 2015.

[21] Dedele A, Miškinyte A. The statistical evaluation and comparison of ADMS-urban model for the prediction of nitrogen dioxide with air quality monitoring network. Environ Monit Assess 2015;187(9):578.

[22] Soulhac L, Nguyen CV, Volta P, Salizzoni P. The model SIRANE for atmospheric urban pollutant dispersion. PART III: Validation against $NO_2$ yearly concentration measurements in a large urban agglomeration. Atmos Environ 2017;167:377–88.

[23] Johansson Lasse, Karppinen Ari, Kurppa Mona, Kousa Anu, Niemi Jarkko V, Kukkonen Jaakko. An operational urban air quality model ENFUSER, based on dispersion modelling and data assimilation. Environ Model Softw 2022;156:105460.

[24] Brussels Environment. La qualité de l'air en Région de Bruxelles-Capitale – Rapport annuel 2023. 2024.

[25] De Craemer Sam, Vercauteren Jordy, Fierens Frans, Lefebvre Wouter, Meysman Filip JR. Using large-scale $NO_2$ data from citizen science for air-quality compliance and policy support. Environ Sci Technol 2020;54(18):11070–8.

[26] Hooyberghs H, De Craemer S, Lefebvre W, Vranckx S, Maiheu B, Trimpeneers E, et al. Validation and optimization of the ATMO-street air quality model chain by means of a large-scale citizen-science dataset. Atmos Environ 2022;272:118946.

[27] Dale Ethan R, Kremser Stefanie, Tradowsky Jordis S, Bodeker Greg E, Bird Leroy J, Olivares Gustavo, et al. The winter 2019 air pollution ($PM_{2.5}$) measurement campaign in Christchurch, New Zealand. Earth Syst Sci Data 2021;13(5):2053–75.

[28] Directorate-General for Research and Innovation (European Commission). Tackling air pollution with citizen science. 2024.

[29] Hooyberghs H, De Craemer S, Lefebvre W, Vranckx S, Maiheu B, Trimpeneers E, et al. Validation and optimization of the ATMO-street air quality model chain by means of a large-scale citizen-science dataset. Atmos Environ 2022;272:118946.

[30] O'Regan Anna C, Grythe Henrik, Hellebust Stig, Lopez-Aparicio Susana, O'Dowd Colin, Hamer Paul D, et al. Data fusion for enhancing urban air quality modeling using large-scale citizen science data. Sustain Cities Soc 2024;116:105896.

[31] Lu Tianjun, Garcia Dulce A, Garcia Armando, Liu Yisi. Leveraging crowd-sourced environmental data to assess air pollution exposure disparity: A case of Los Angeles county. Int J Appl Earth Obs Geoinf 2023;125:103599.

[32] Lauriks F, Jacobs D, Meysman FJR. CurieuzenAir: data collection, data analysis and results. Technical report, University of Antwerp; 2022.

[33] TomTom International BV. Traffic index ranking | TomTom traffic index. 2023.

[34] ITF. Developing innovative mobility solutions in the Brussels-capital region. Technical Report 97, Paris: OECD Publishing; 2021.

[35] Palmes ED, Gunnison AF, DiMattio J, Tomczyk C. Personal sampler for nitrogen dioxide. Am Ind Hyg Assoc J 1976;37(10):570–7.

[36] Netherlands Standardization Institute (NEN). NEN EN 16339: Ambient air - method for the determination of the concentration of nitrogen dioxide by diffusive sampling. 2023.

[37] European Parliament and Council of the European Union. Directive 2008/50/EC of the European parliament and of the council of 21 2008 on ambient air quality and cleaner air for Europe. 2008.

[38] QGIS Development Team. QGIS geographic information system. Technical report, 2024.

[39] Posit team. RStudio: integrated development environment for R. posit software. Technical report, boston, MA: PBC; 2024, URL http://www.posit.co/.

[40] Breiman Leo. Random forests. Mach Learn 2001;45:5–32.

[41] Fox John, Monette Georges. Generalized collinearity diagnostics. J Amer Statist Assoc 1992;87(417):178–83.

[42] Wright Marvin N, Ziegler Andreas. ranger: A fast implementation of random forests. 2015.

[43] Janitza Silke, Celik Ender, Boulesteix Anne-Laure. A computationally fast variable importance test for random forests for high-dimensional data. Adv Data Anal Classif 2018;12(4):885–915.

[44] Hanna Steven. Uncertainties in air quality model predictions. Bound-Layer Meteorol 1993;62:3–20.

[45] Cressie Noel. Geostatistics. John Wiley & Sons, Ltd; 1993.

[46] Soulhac L, Salizzoni P, Cierco F-X, Perkins R. The model SIRANE for atmospheric urban pollutant dispersion, PART I: presentation of the model. Atmos Environ 2011;45:7379–95.

[47] Carpentieri M, Salizzoni P, Robins A, Soulhac L. Evaluation of a neighborhood scale, street network dispersion model through comparison with wind tunnel data. Environ Model Softw 2012;37:110–24.

[48] Soulhac L, Salizzoni P, Mejean P, Didier D, Rios I. The model SIRANE for atmospheric urban pollutant dispersion; PART II: validation of the model on a real case study. Atmos Environ 2012;49:320–37.

[49] Soulhac L, Nguyen C-V, Volta P, Salizzoni P. The model SIRANE for atmospheric urban pollutant dispersion. PART III: validation against $NO_2$ yearly concentration measurements in a large urban agglomeration. Atmos Environ 2017.

[50] Grylls T, Le Cornec CMA, Salizzoni P, Soulhac L, Stettler MEJ, van Reeuwijk M. Evaluation of an operational air quality model using large-eddy simulation. Atmos Environ: X 2019.

[51] Ntziachristos L, Gkatzoflias D, Kouridis C, Samaras Zissis. COPERT: A European road transport emission inventory model. Environ Sci Eng 2009.

[52] Agudelo-Castañeda Dayana, De Paoli Fabrício, Morgado-Gamero Wendy B, Mendoza Martha, Parody Alexander, Maturana Aymer Y, et al. Assessment of the $NO_2$ distribution and relationship with traffic load in the Caribbean coastal city. Sci Total Environ 2020;720:137675.

[53] Jenkins N, Legassick W, Sadler L, Sokhi RS. Correlation between NO and $NO_2$, roadside concentrations, traffic volumes and local meteorology at a major London route. WIT Trans Ecol Environ 1995;6:405–12.

[54] Zhan Yang, Luo Nan, Wang Qian, Zhang Hong, Yu Qiang, Gu Xiaohong, et al. A satellite-based land use regression model of ambient NO with high spatiotemporal resolution. Remote Sens 2021;13(3):397.

[55] Bo M, Salizzoni P, Pognant F, Mezzalama R, Clerico M. A combined citizen science - modelling approach for $NO_2$ assessment in torino urban agglomeration. Atmosphere 2020.

[56] Testi I, Wang A, Paul S, Mora S, Walker E, Nyhan M, et al. Big mobility data reveals hyperlocal air pollution exposure disparities in the Bronx, New York. Nat Cities 2024. http://dx.doi.org/10.1038/s44284-024-00093-x.

[57] Beckwith M, Bates E, Gillah A, Carslaw N. $NO_2$ hotspots: Are we measuring in the right places? Atmos Environ: X 2019;2:100025.