

Article

A Nonlinear Land Use Regression Approach for Modelling NO₂ Concentrations in Urban Areas—Using Data from Low-Cost Sensors and Diffusion Tubes

Said Munir ^{1,*} , Martin Mayfield ¹, Daniel Coca ² and Lyudmila S Mihaylova ² 

¹ Department of Civil and Structural Engineering, the University of Sheffield, Sheffield S1 3JD, UK; martin.mayfield@sheffield.ac.uk

² Department of Automatic Control and Systems Engineering, the University of Sheffield, Sheffield S1 3JD, UK; d.coca@sheffield.ac.uk (D.C.); l.s.mihaylova@sheffield.ac.uk (L.S.M.)

* Correspondence: smunir2@sheffield.ac.uk; Tel.: +44-798-600-1328

Received: 8 June 2020; Accepted: 9 July 2020; Published: 11 July 2020



Abstract: Land Use Regression (LUR) based on multiple linear regression model is one of the techniques used most frequently for modelling the spatial variability of air pollution and assessing exposure in urban areas. In this paper, a nonlinear generalised additive model is proposed for LUR and its performance is compared to a linear model in Sheffield, UK for the year 2019. Pollution models were estimated using NO₂ measurements obtained from 188 diffusion tubes and 40 low-cost sensors. Performance of the models was assessed by calculating several statistical metrics including correlation coefficient (R) and root mean square error (RMSE). High resolution (100 m × 100 m) maps demonstrated higher levels of NO₂ in the city centre, eastern side of the city and on major roads. The results showed that the nonlinear model outperformed the linear counterpart and that the model estimated using NO₂ data from diffusion tubes outperformed the models using data from low-cost sensors or both low-cost sensors and diffusion tubes. The proposed method provides a basis for further application of advanced nonlinear modelling approaches to constructing LUR models in urban areas which enable quantifying small scale variability in pollution levels.

Keywords: air quality modelling; land use regression; nonlinear regression; Sheffield; spatial analysis; low-cost sensors; nitrogen dioxide

1. Introduction

Air pollution is one of the serious environmental issues that affects human health and may cause mortality. Landrigan [1] reported that air pollution caused 64 million deaths worldwide in 2015, which shows the significance of the impact of poor air quality on human health. Several air pollutants are shown to have negative impacts on human health, however, among gaseous pollutants nitrogen dioxide (NO₂) is considered the most serious pollutant causing both chronic and acute respiratory diseases including asthma, hospital admission and mortality [2]. NO₂ is considered the most serious gaseous pollutant in urban areas and many air quality management areas (AQMAs) in the UK are based on the exceedances of NO₂ [3]. Therefore, it is important to carry out different monitoring and modelling investigations to analyse its spatial variability, especially micro-level variability. This could be done by developing high resolution maps in urban areas that help in understanding the main drivers of NO₂ levels and quantifying exposure to elevated NO₂ concentrations in urban areas. It is not feasible to capture micro-level spatial variability in NO₂ concentration with the help of monitoring network in a large urban area as this requires a huge number of sensors. Alternatively, a certain

number of sensors can be installed, and the concentrations can be extrapolated to other areas using modelling techniques. This is exactly what this study intends to do.

One of the challenges of air pollution in urban areas is the quantification of small-scale local level exposure by analysing spatial variability in pollutant concentrations. The small scale variability in urban areas is controlled by emission sources, land use features, building density-and-height and geographical characteristics. To characterise spatial variability in pollutant concentrations, three approaches are most commonly used: GIS (Geographical Information Systems) based interpolation methods, Dispersion models, and land use regression (LUR) models.

Hsu et al. [4] reviewed different methods used for interpolation and divided them into two categories: Geostatistical and Non-geostatistical methods. The geostatistical techniques include Ordinary Kriging, Universal Kriging, Simple Kriging, Empirical Bayesian Kriging and Original CoKriging. The non-geostatistical techniques include Splines, Trend Surface Analysis, Inverse Distance Weighting and Natural Neighbour. According to the findings of Hsu et al. [4] the geostatistical interpolations showed better prediction than the non-geostatistical interpolation techniques. Interpolation methods are used to interpolate air pollutant concentrations between various air monitoring stations to provide better spatial coverage. However, when these approaches are applied at a local scale such as intra-city scale, these methodologies are known to produce considerable variations in air pollutant concentrations within a small area and are more effective at large scales, such as national or regional scale [5]. Therefore, at urban scale interpolation approaches should not be the priority for analysing spatial variability of air pollution.

Dispersion models are probably the most advanced modelling techniques for determining the spatial variability of air pollutants in urban areas. Dispersion models combine the data of pollutant emission from different sources (e.g., point, line and area sources), the geophysical characteristics of the study area and meteorological parameters. Dispersion models have the potential to incorporate both temporal and spatial variations to replace the need for air pollution monitoring. However, high cost of purchase and maintenance, high input demands, and requirement of skilled staff limit their application.

LUR models provide an effective alternative to GIS interpolation and dispersion modelling techniques on urban scale. LUR is a spatial modelling approach used most frequently for analysing the spatial variability and quantifying public exposure to air pollution in urban areas. Different modelling, mapping and data fusion techniques are available for modelling spatial variability of air pollution in urban areas [6]. Several authors have preferred Land Use Regression (LUR) over other approaches. Briggs et al. [7] compared the performance of LUR with spatial interpolation methods (e.g., kriging, TIN-contouring and trend surface analysis) and reported that LUR performed much better than the interpolation techniques. The reason was that in urban areas spatial variability of air pollutants is more controlled by the local emission sources and geographical characteristics, rather than a smoothly varying field which is assumed by the interpolation methods. Hoek et al. [8] reviewed several LUR models and reported that the performance of the LUR model in urban areas was either equivalent or better than dispersion and geostatistical approaches. LUR is a widely employed approach for air pollution exposure estimation using Geographical Information Systems (GIS) and statistical analyses to determine the association between geographic features and measured atmospheric pollutant concentrations [9]. LUR is easy to implement and can provide an effective alternative to geostatistical and dispersion modelling techniques on urban scale. LUR approach was introduced by Briggs et al. [9] and since then has been used in numerous studies around the world [8,10–16]. LUR models associate pollutant concentrations, such as NO₂ to site specific geographical characteristics, e.g., topography, land use, traffic, population density, altitude and meteorological parameters. The use of these variables in the regression model are known to capture small scale variability on city scale [17]. Recent development in GIS technology has also added to the popularity of LUR approaches.

Rahman et al. [13] developed an LUR model in Brisbane, Australia to predict NO₂ and NO_x during 2009–2012. The model was able to explain 64% and 70% variations in NO₂ and NO_x, respectively. Distance to major roads and industrial areas were the common predictor variables for both NO₂ and

NO_x, suggesting an important role of road traffic and industrial emissions. Rahman et al. [13] used the following independent variables in their model: distance to coast (km), distance to port (km), distance to airport (km), distance to nearest major road (km), distance to nearest minor road (km), major road length (km), minor road length (km), population density (person/km²), land use by type (km²), and elevation (m). Muttoo et al. [12] used several geographic predictor variables to predict NO_x levels in Durban, South Africa employing an LUR model. They used length of minor roads within a 1000 m radius, length of major roads within a 300 m radius, and area of open space within a 1000 m radius in the model as independent variables. The LUR model was able to explain 73% variance in NO_x concentrations, however cross validation resulted in R² value of 0.59. Hoek et al. [8] have provided a detailed review on LUR models, identifying 25 studies on the subject. They have identified several significant predictors for LUR models, including various traffic characteristics, population characteristics, land use, physical geography, and climatic conditions. Gillespie et al. [18] developed an LUR model to estimate exposure to NO₂ in Glasgow, Scotland and reported that the use of more than 60 training sites had a considerable beneficial effect on model performance. Mostly, LUR approaches are applied typically at a city scale e.g., [17,19], however, some researchers have applied LUR models to entire countries. Beelen et al. [15] and Stedman et al. [14] applied LUR models in Netherland and UK, respectively, whereas Vienneau et al. [20] developed an LUR model for both UK and Netherland and compared its outputs in both countries.

Sheffield City Council (SCC) has declared most of the urban area in Sheffield City as AQMA due to the high levels of NO₂. Detailed modelling investigations are required to model NO₂ concentrations and to create high resolution maps in the city for quantifying public exposure to NO₂ and determining small scales NO₂ spatial variability in the city. In this paper LUR models are developed to predict NO₂ concentrations and make high resolution maps (100 m × 100 m) using NO₂ data measured by a network of diffusion tubes (DT) and low-cost sensors (LCS). The literature reviewed above all have employed multiple linear regression models (MLRM) for developing LUR models, which work on the assumption that response and predictor variables have linear association. Here an advanced nonlinear approach Generalised Additive Model (GAM) is proposed, which is more suitable for air quality data analysis. The rest of the paper is structured as follows: Methodology of this paper is presented in Section 2, wherein Section 2.1 describes monitoring sites and predictor variables; Section 2.2 describes LUR model development; Section 2.3 describes model specifications; Section 2.4 describes model validation; and Section 2.5 describes statistical software used in this study. Results and discussion are presented in Section 3 and the main outcomes of this work are summarised in Section 4.

2. Methodology

In this study three LUR models are developed for modelling the spatial variability of NO₂ concentrations and producing high resolution maps (100 m × 100 m) in Sheffield for the year 2019. Sheffield (53°23' N, 1°28' W) is a historical metropolitan borough in South Yorkshire, United Kingdom and has emerged as a green and modern cityscape in the proximity of the Peak District National Park. According to 2011 census Sheffield City had a population of 552,700, however, since then the population has grown and according to more recent estimates has reached about 700,000. Among gaseous pollutants, NO₂ is the pollutant of concern in Sheffield and most of the AQMA in Sheffield is declare due to the elevated levels of NO₂ concentrations mostly emitted by road traffic [21].

2.1. Predictor Variables and NO₂ Monitoring Sites

In this paper NO₂ concentration (µg/m³) is modelled using data from DT and LCS. There were 188 DT and 40 LCS measuring NO₂ concentrations (µg/m³) in Sheffield. The locations of DT and LCS are shown in Figure 1. There are two types of LCS: 13 AQMesh pods and 27 Envirowatch E-motes. For a brief description of these sensors see Munir et al. [21,22]. DT and LCS have relatively high uncertainty (about 25 to 30%) as compared to reference sensors (15%). Generally, DT are exposed for a period of 2–4 weeks (no longer than 5 weeks and no shorter than 1 week). After this period, the old DT are

replaced with new ones. In this way, the monitoring is carried out for the whole year to get annual average. LCS (both AQMesh and Envirowatch E-motes) were installed around the city (Figure 1), providing high resolution temporal data (e.g., 5 min to hourly), which were converted to annual average. A summary of NO_2 concentrations measured by DT and LCS for year 2019 is provided in Table 1.

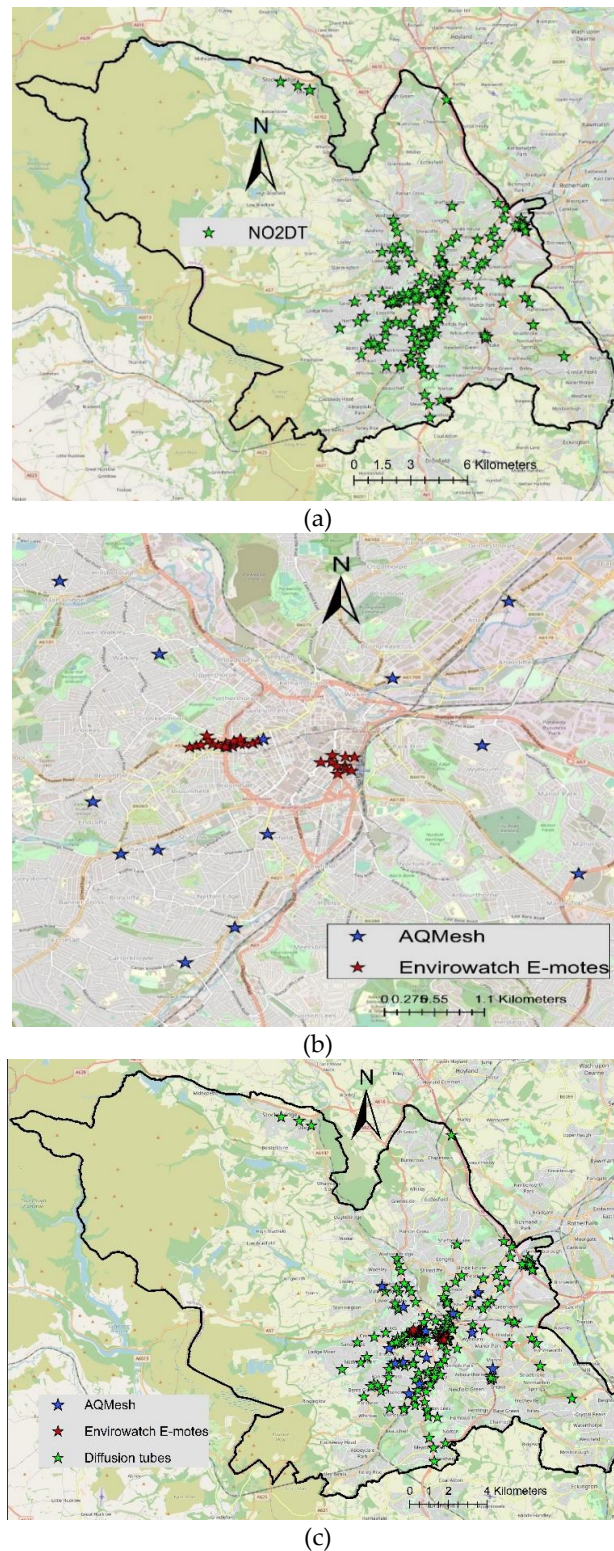


Figure 1. (a) Showing the locations of 188 NO_2 diffusion tubes (DT), (b) 40 low-cost sensors (LCS): 13 AQMesh pods and 27 Envirowatch E-motes, and (c) both LCS and DT in the City of Sheffield.

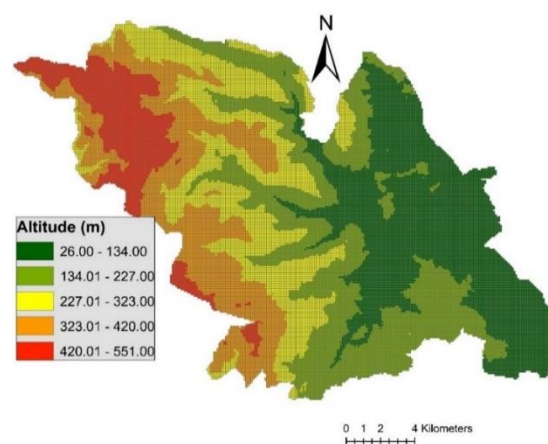
Table 1. Descriptive statistics of NO₂ measured by DT (diffusion tubes) and LCS (low-cost sensors) in Sheffield.

Metrics	DT NO ₂	LCS NO ₂
Minimum	13.58	12.74
1st Quartile (25th percentile)	28.29	25.23
Median	33.77	33.19
Mean	34.23	41.23
3rd Quartile (75th percentile)	40.00	45.28
Maximum	62.03	146.54
Standard Deviation	9.65	27.69

To model NO₂ concentration, data of different land use, traffic and population variables were collected to be used as predictor (independent) variables. Maps of the predictor variables were downloaded from the ordinance survey UK and provided by the Sheffield City Council. ArcGIS version 10.7.1 and its LUR tools were used to extract different values within 100 m × 100 m grid. These variables are given below:

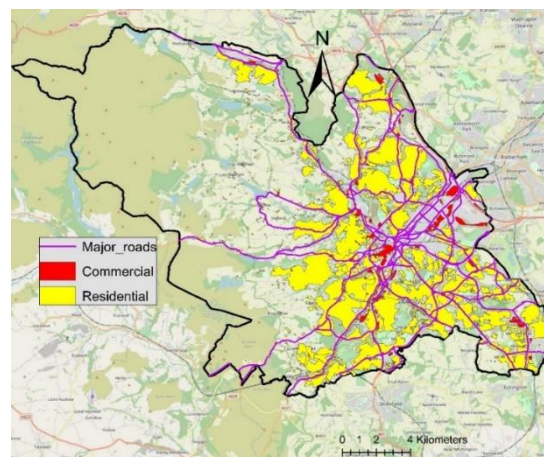
- Area (m²) of industrial land use, residential area, commercial area, parks and green area, and building area;
- Length (m) of motorways, major roads, and minor roads;
- Distance (m) to motorway, major road, minor road, building, industry, bus stop, parks, commercial area, and residential area;
- Population (persons per km²), Altitude (m), number of bus stops, easting (m), northing (m), and street intersection.

As the impact distance varies among different variables, buffers of multiple radii (10, 50, 100, 200, 300, 500, and 1000 m) were created for industrial area, commercial area, park and green area, residential area, major roads, minor roads, motorways, and bus stops. Figure 2 shows the spatial distribution of different predictor variables in the city.

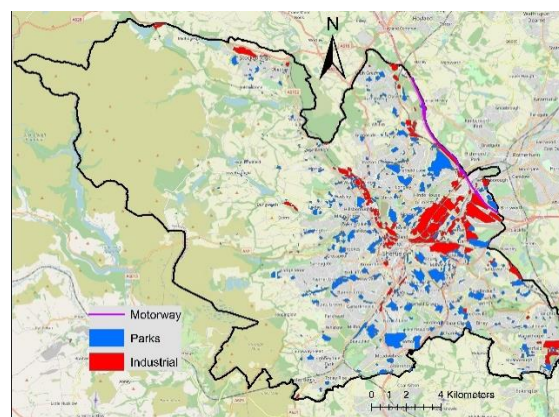


(a)

Figure 2. Cont.



(b)



(c)

Figure 2. (a) Altitude, (b) major roads, commercial and residential area (c) and motorway, green parks and industrial area in Sheffield.

2.2. LUR Model Development

In this paper LUR models were developed for modelling NO_2 concentrations in Sheffield. NO_2 concentration was regressed against the predictor variables. Both linear and non-linear LUR models were developed using three NO_2 datasets:

- (1) Measurements of NO_2 obtained from 188 DT;
- (2) Measurements of NO_2 obtained from 40 LCS;
- (3) Combined NO_2 measurements obtained from both LCS and DT (228).

It should be remember that in each case 75% randomly selected data were used for model training (fitting) and 25% data were hold-out for model testing (cross validation). The novelty of this study is that in addition to a Multiple Linear Regression Model (MLRM), a nonlinear Generalised Additive Model (GAM) is proposed for developing LUR model. The performance of MLRM and GAM is compared. Secondly, in addition to DT measurements, NO_2 data from a network of 40 LCS are also used in this paper. It should be noted that both MLRM and GAM were developed and validated in R programming language [23]. However, to extrapolate predicted NO_2 concentration to the entire Sheffield city to produce continuous heat maps, MLRM and GAM used different approaches (see Section 2.5).

MLRM and GAM are shown in Equations (1) and (2) (these are just examples, total 72 predictor variables were used in the initial model, which were minimised by stepwise regression model).

i. MLRM

$$\begin{aligned} \text{NO}_2\text{Conc} \sim & \beta_0 + \beta_1 (\text{Population}) + \beta_2 (\text{BusStops}) + \beta_3 (\text{Altitude}) + \beta_4 (\text{IndustrialArea}) + \\ & \beta_5 (\text{ResidentialArea}) + \beta_6 (\text{BuildingArea}) + \beta_7 (\text{ParksArea}) + \beta_8 (\text{Easting}) + \beta_9 (\text{Northing}) + \\ & \beta_{10} (\text{MajorRd}) + \beta_{11} (\text{MinorRd}) + \beta_{12} (\text{St_Intersect}) + \beta_{13} (\text{CommercialArea}) + \\ & \beta_{14} (\text{Dist_MajorRd}) + \beta_{15} (\text{Dist_MinorRd}) + \beta_{16} (\text{Dist_Building}) + \beta_{17} (\text{Dist_Industry}) + \\ & \beta_{18} (\text{Dist_Parks}) + \beta_{19} (\text{Dist_Motorway}) + \beta_{20} (\text{Dist_commercialArea}) + \\ & \beta_{21} (\text{Dist_ResidentialArea}) + \beta_{22} (\text{Dist_BusStop}) + \beta_{23} (\text{Motorway}) + \varepsilon \end{aligned} \quad (1)$$

In Equation (1), β_0 is the intercept, β_1 to β_{23} are the coefficients (slopes) of the predictor variables and ε is the error term (the difference between modelled and measured concentrations).

ii. GAM

$$\begin{aligned} \text{NO}_2\text{Conc} \sim & \alpha + s_1 (\text{Population}) + s_2 (\text{BusStops}) + s_3 (\text{Altitude}) + s_4 (\text{IndustrialArea}) + \\ & s_5 (\text{ResidentialArea}) + s_6 (\text{BuildingArea}) + s_7 (\text{ParksArea}) + s_8 (\text{Easting}) + s_9 (\text{Northing}) + \\ & s_{10} (\text{MajorRd}) + s_{11} (\text{MinorRd}) + s_{12} (\text{St_Intersect}) + s_{13} (\text{CommercialArea}) + \\ & s_{14} (\text{Dist_MajorRd}) + s_{15} (\text{Dist_MinorRd}) + s_{16} (\text{Dist_Building}) + s_{17} (\text{Dist_Industry}) + \\ & s_{18} (\text{Dist_Parks}) + s_{19} (\text{Dist_Motorway}) + s_{20} (\text{Dist_commercialArea}) + \\ & s_{21} (\text{Dist_ResidentialArea}) + s_{22} (\text{Dist_BusStop}) + s_{23} (\text{Motorway}) + \varepsilon \end{aligned} \quad (2)$$

In Equation (2), α is the intercept, 's' term is the smoothing function of the covariates and ε is the residual or error term, the difference between measured and predicted values. For smoothing term the degree of smoothing was automatically assigned by the generalised cross validation (GCV) method described by Wood and Augustin [24]. For more details on smoothing functions see Wood [25,26]. MLRM explicitly assume normality of the error term and linearity of the relationship between response variable and predictor variables. GAM is an advanced model and relaxes such restrictions. Therefore, GAM is able to successfully handle nonlinearities in the association between response and predictor variables, which is important for air quality data as the relationship is not always linear. For more details on GAM see Hastie and Tibshirani [27], Wood [25] and Wood [26].

2.3. Model Specification

Model specification (also refer to as model selection) is the process of determining which predictor variable to include or exclude from the model. In this study we employed stepwise regression algorithm (both forward and backward) for model selection. The aim was to find the best performing model (minimising prediction error) with minimum number of predictors (Parsimonious model) by selecting only those predictors whose contribution was significant in controlling the variations of NO_2 concentrations. For this purpose MASS-package [28] in R-programming language [23] was used.

2.4. Model Validation

Model validation is testing the goodness of fit of the fitted model. In this process we compare the predicted concentrations with measured one. Here we used cross validation process which is a generalisation of the model to an independent dataset, not used in the model fitting. Randomly selected 75% data was used for model fitting and 25% for model validation.

2.5. Mapping Modelled NO_2 Concentration

After both MLRM and GAM were fitted and validated, NO_2 concentration was predicted for the entire Sheffield City for 37,605 square grids each with size 100 m $\times$ 100 m. To do this MLRM and GAM used different techniques. In case of MLRM to predict NO_2 concentrations the coefficients of

predictor variables were used in ArcGIS using ‘field calculator’ function in attribute table. However, GAM being a nonlinear model, doesn’t produce a single coefficient (slope) for each predictor variable. Therefore, it was not possible to use the same approach. Instead, NO₂ concentration was predicted in mgcv-package [25] in R-programming language [23] and then exported to ArcGIS. Once in ArcGIS, the layer containing the predicted NO₂ concentration was joined with the polygon layer having the squared grids. For this purpose the whole study area was divided into 37,605 square grids, each 100 m × 100 m resolution using ArcGIS 10.7.1 software.

2.6. Statistical Software

In this study mainly two statistical and mapping software were used: (a) ArcGIS version 10.7.1 and its LUR tools. LUR Tools is an ArcGIS toolbox having some important functions for constructing land use predictor variables for developing LUR model. (b) R programming language [23] and several of its packages including ‘MASS’ [28], ‘openair’ [29] and ‘mgcv’ [25]. The ‘mgcv-package’ was used for running GAM, ‘MASS’ was used for running stepwise regression and ‘openair-package’ was used for general data analysis and developing different plots.

3. Results and Discussion

In this study both MLRM and GAM are employed to model the spatial variability of NO₂ concentrations in Sheffield. NO₂ data were collected from two main sources: DT and LCS. Firstly, MLRM and GAM were developed using data from 188 DT (Section 3.1), followed by the models using NO₂ data from LCS (Section 3.2), and both DT&LCS (Section 3.3). In all three case the data were divided into training dataset (randomly selected 75% data, used in the fitted model) and hold-out testing dataset (randomly selected 25% data, used in the cross validated model). Both GAM and MLRM were fitted using all predictor variables first, and then stepwise regression was used for model specification aiming to select only those predictors which had a significant effect.

3.1. LUR Model Using NO₂ Data from DT

Both GAM and MLRM were used to regress NO₂ concentrations from DT against the predictor variables. Stepwise regression showed that the following covariates had significant effect: distance to major road (m), distance to minor road (m), residential area (m²), commercial area (m²), distance to bus stops (m), building area (m²), and altitude (m). Among them, building and commercial area had positive coefficients, whereas residential area, distance to major road, distance to minor road, distance to bus stop and altitude had negative coefficients.

Three types of roads were used as covariates in the models: motorway, major roads (A-roads and B-roads), and minor roads. Among these the effect of major and minor roads was significant, the reason is obvious that these roads carry most of the traffic and are spread throughout the study area. In contrast motorway, although very busy has little length inside the study area as shown in Figure 2c and therefore had an insignificant effect. It should be noted that negative coefficients of distance to major roads, minor roads and bus stops show positive effect of these three variables. In other words, as the distance between DT and these variables increase, NO₂ concentrations decrease. Therefore, areas near roads and bus stops have higher concentrations compared to areas away from the roads and bus stops, which is expected. Some researchers have suggested to take inverse or squared-inverse of the distance, e.g., [30] to turn the coefficients positive, however, it does not change the output of the model. Therefore, we stick to the original values. The negative effect of altitude on air quality is well known [31], meaning NO₂ concentration decreases at higher altitude. According to the data used in this study, minimum altitude was 26 m and maximum 551 m (Figure 2a). The effect of altitude was negative on NO₂ concentrations and highly significant. The western side of the city including Peak District National Park had higher altitude and lower NO₂ concentrations as compared to the city centre and eastern side, which had lower altitude and higher level of NO₂. The effect of residential area was negative, probably due to the fact that NO₂ levels were relatively lower at residential areas than at commercial areas and

roadside locations. The effect of industrial area was insignificant probably due to the reasons that not many sensors are installed near industrial area. In addition, due to tall chimneys the emissions from the industry are not read by the local sensors. More recently Munir et al. [21] reported that industrial emissions in Sheffield had significant effect on the level of PM_{10} , but not on the levels of NO_2 .

Coefficients and level of significance of different predictor variables for MLRM are shown in Table 2. Nonlinear model does not provide a single coefficients for each predictor variable, the output of the GAM model showing the association between NO_2 and predictor variables is shown in Figure 3 (only two predictor variables are shown for brevity). Figure 3 shows that NO_2 concentrations decrease drastically as distance from the roadside increases up to approximately 500 m, afterwards the curve is flattened as the distance increase further and there are fewer data points, resulting in wider error bars. In case of altitude, the negative effect is stronger at higher altitudes (altitude >180 m).

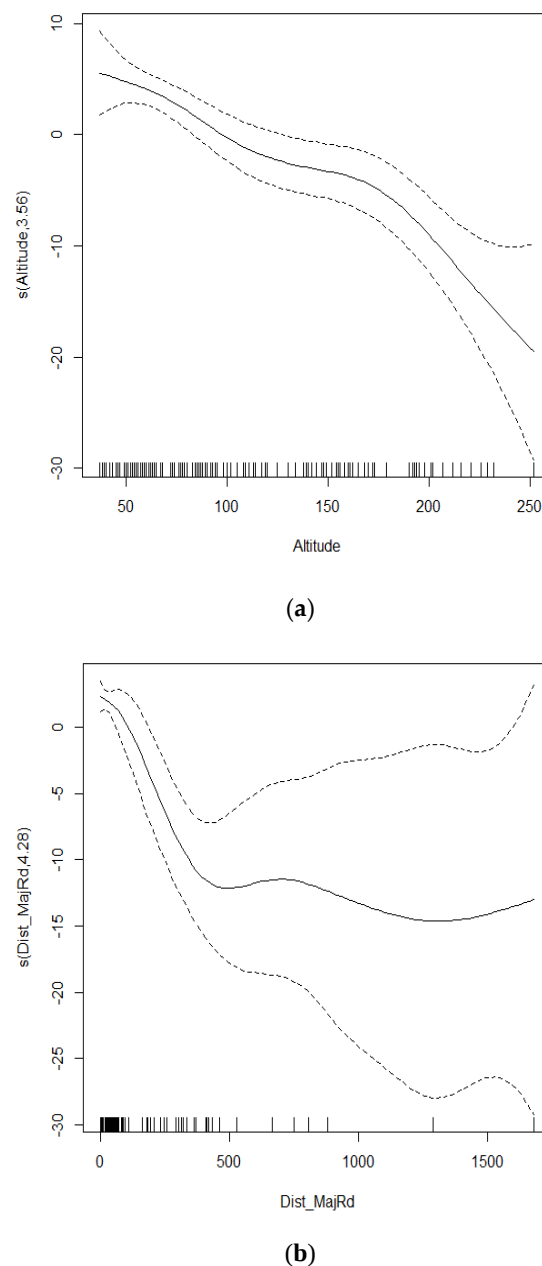


Figure 3. (a) Showing the nonlinear association of NO_2 concentrations ($\mu g/m^3$) with altitude in meters and (b) distance to major roads in meters. The dashed lines are the estimated 95% confidence intervals. The vertical lines on the x-axis show the presence of data.

Table 2. Showing coefficients and significance levels of different predictor variables for MLRM (multiple linear regression models).

Predictor Variable	Coefficient	p-Value
Intercept	43.7025	$<2 \times 10^{-16}$ ***
Building	0.0020	0.075 +
Dist_MajorRd	−0.0114	0.001 **
Dist_MinorRd	−0.0968	0.052 +
Residential	−0.0006	0.002 **
Commercial	0.0005	0.053 +
Altitude	−0.0543	0.000 ***
Dist_Bstop	−0.0308	0.026 *

Note: p-stars relate to how statistically significant the effect is: $p\text{-value} < 0.001 = \text{***}$, $p\text{-value} < 0.01 = \text{**}$, $p\text{-value} < 0.05 = *$ and $p\text{-value} < 0.1 = +$.

Several statistical metrics were calculated for model assessment for both fitted model (FM) and cross validation (CV) (Table 3). Statistical metrics used here were correlation coefficients (r , unitless), fraction of predictions within a factor of two of observations (FAC2, unitless), root mean squared error (RMSE, same units as the quantity being considered, here $\mu\text{g}/\text{m}^3$), mean bias (MB, same units as the quantity being considered, here $\mu\text{g}/\text{m}^3$), mean gross error (MGE, same units as the quantity being considered, here $\mu\text{g}/\text{m}^3$), normalised mean bias (NMB, unitless) and normalised mean gross error (NMGE, unitless). Generally, GAM with r -values 0.73 and 0.70 for training and hold-out dataset, respectively showed better performance than MLRM with r -value 0.67 for both training and hold-out dataset. Other metrics showing error of the model (e.g., NMGE, MGE and RMSE) are slightly greater for MLRM than GAM, indicating better performance of GAM in terms of smaller difference between predicted and observed concentrations. Furthermore, RMSE, NMGE and MGE have positive values, showing slightly over prediction of the model. FAC2 having value of 1 shows acceptable model performance for both MLRM and GAM. Results showed that models performance was not deteriorated considerably when applied to independent testing dataset. Comparison of predicted and measured NO_2 concentrations is made graphically in the form of scatter plots in Figure 4, showing strong correlation between modelled and measured concentrations for both models. Both models have over-predicted at lower levels of NO_2 ($<40 \mu\text{g}/\text{m}^3$) and under-predicted at higher levels ($>40 \mu\text{g}/\text{m}^3$), especially for the testing datasets (Figure 4c,d).

Table 3. Showing various statistical metrics for assessing the performance of MLRM and GAM for both fitted model (FM) using training dataset and cross-validated model (CV) using hold-out dataset based on NO_2 data from DT. MB (mean bias), MGE (mean gross error) and RMSE (root mean square error) have the same units as the quantity being considered (here $\mu\text{g}/\text{m}^3$), whereas FAC2, r , NMGE and NMB are unitless.

Metrics	MLRM		GAM	
	FM	CV	FM	CV
FAC2	1	1	1	1
MB	2×10^{-15}	−0.95	2×10^{-11}	−1.17
MGE	4.98	6.89	4.80	6.55
NMB	2×10^{-17}	−0.03	2×10^{-12}	−0.03
NMGE	0.15	0.19	0.14	0.18
RMSE	6.44	8.98	6.13	8.69
r	0.67	0.67	0.73	0.70

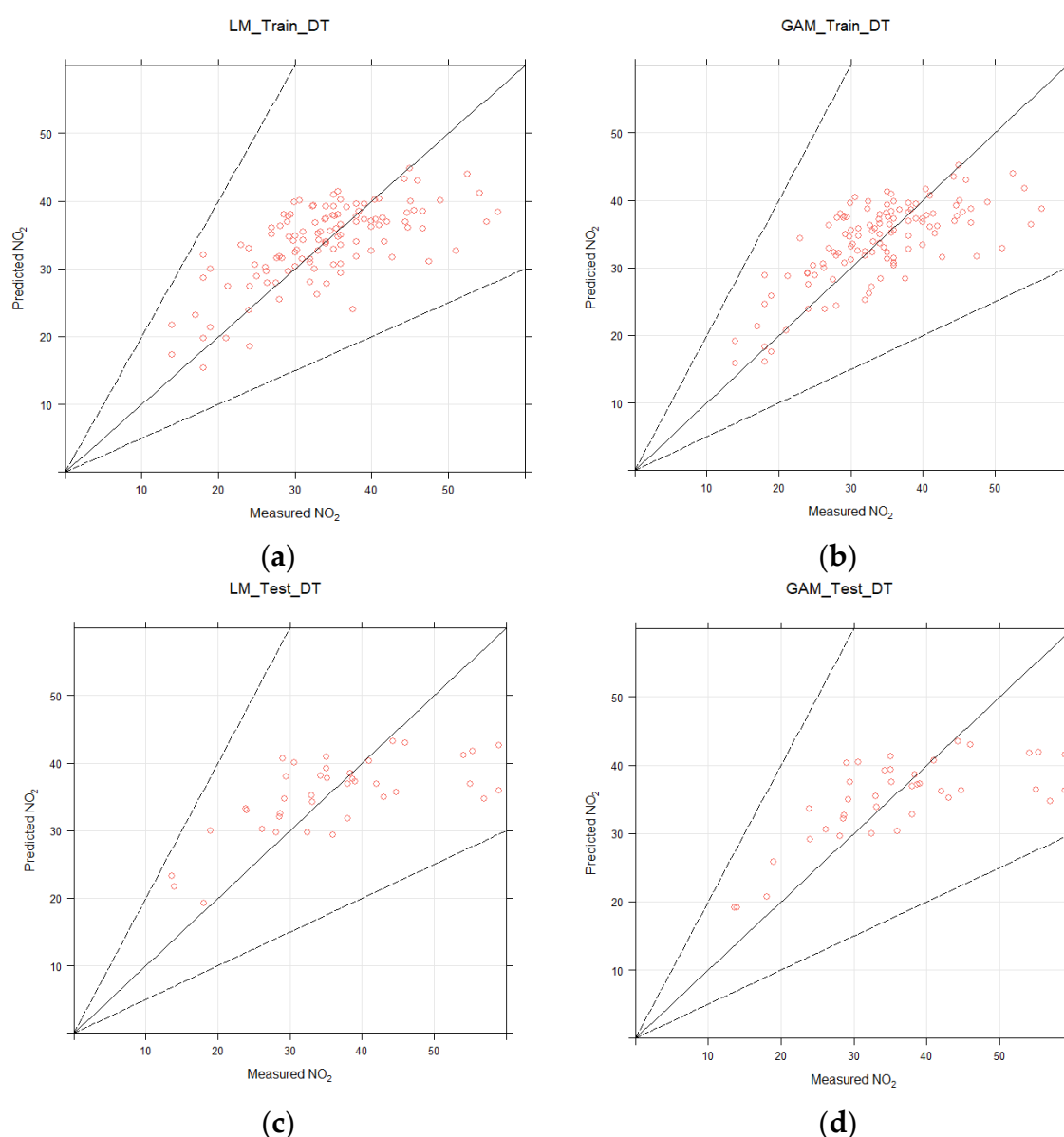
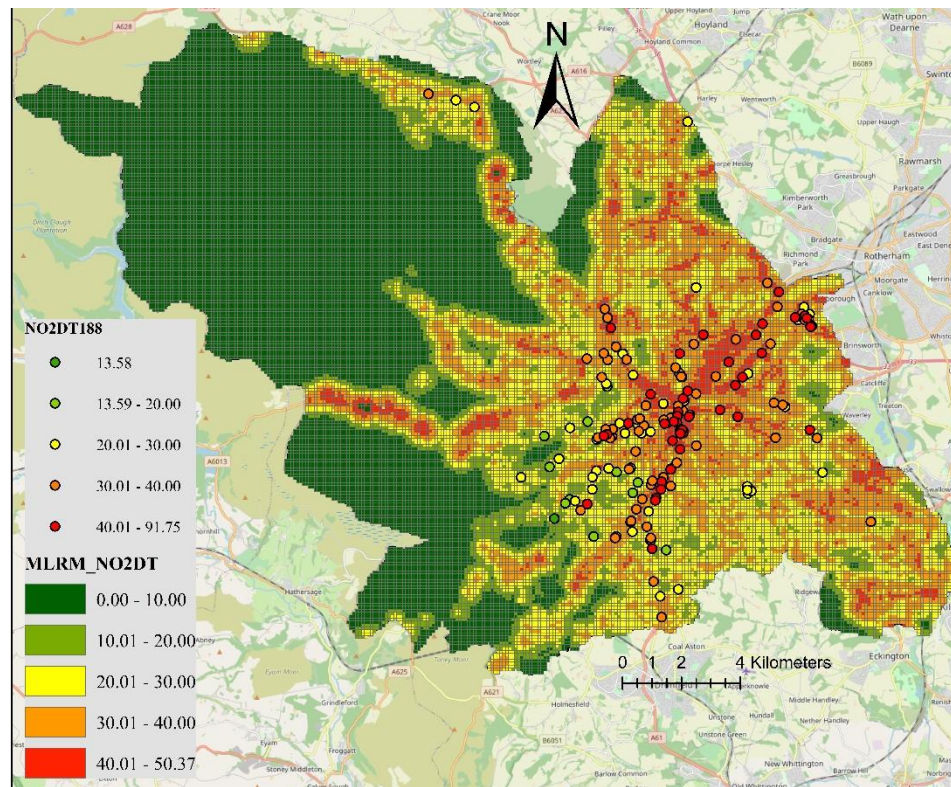
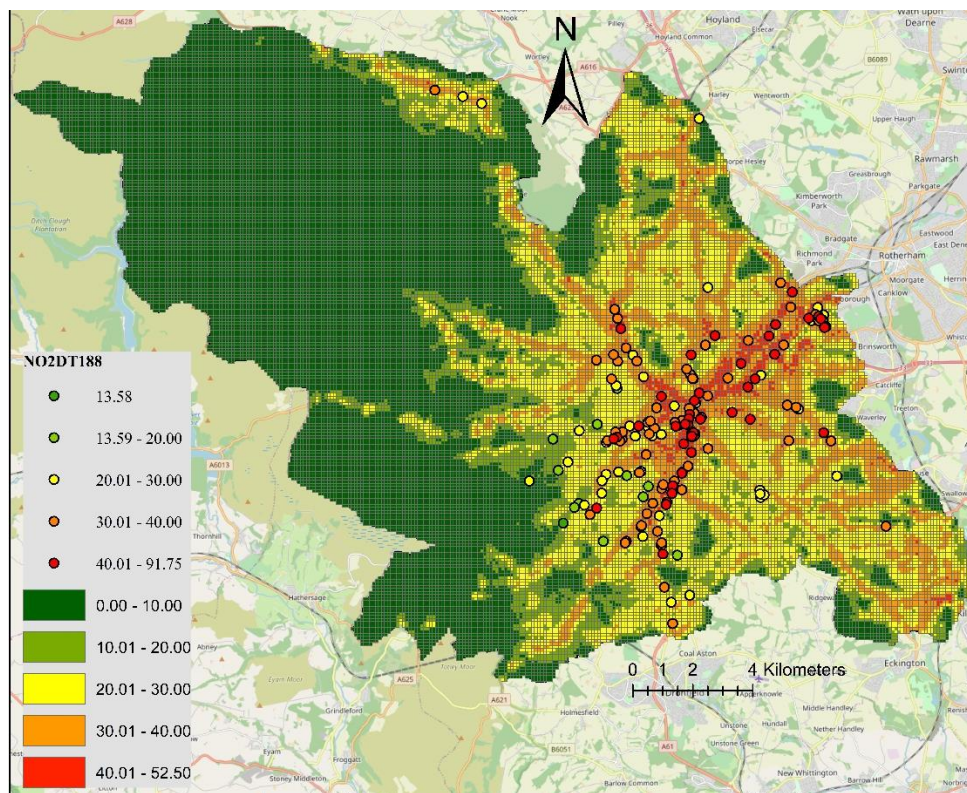


Figure 4. (a,b) Comparing measured and predicted NO₂ concentrations using multiple linear regression model (MLRM) and generalised additive model (GAM) for training (c,d) and testing dataset using NO₂ data from diffusion tubes (DT) only.

Furthermore, NO₂ concentration was predicted for 37,605 square grids with 100 m × 100 m resolution to produce maps of NO₂ concentrations for the entire Sheffield City. The resultant maps of predicted NO₂ concentrations for both MLRM and GAM are shown in Figure 5. The model successfully captured spatial variability of NO₂ concentrations in Sheffield, showing higher levels of NO₂ in the city centre and on busy roads around the city. The city centre and the area between the motorway (M1) and the city centre is particularly highlighted with NO₂ levels higher than EU annual limits of 40 µg/m³ (Figure 5). Western part of the city especially Peak District National Park has shown lower level of NO₂ concentrations due to high altitude and limited amount of minor and major roads. Recently, Munir et al. [21] using Airviro dispersion modelling system, have reported similar results in Sheffield. However, the results of this study are much more detailed (100 m × 100 m resolution), successfully capturing micro-level local variations in NO₂ concentrations, intended for quantifying public exposure. The maps show how NO₂ levels vary from street to street. Spatial trends of predicted NO₂ levels closely match with measured NO₂ levels (Figure 5).



(a)



(b)

Figure 5. High resolution maps (100 m × 100 m) of NO₂ concentrations (µg/m³) predicted by (a) MLRM and (b) GAM in Sheffield. The points show values of NO₂ measured by DT.

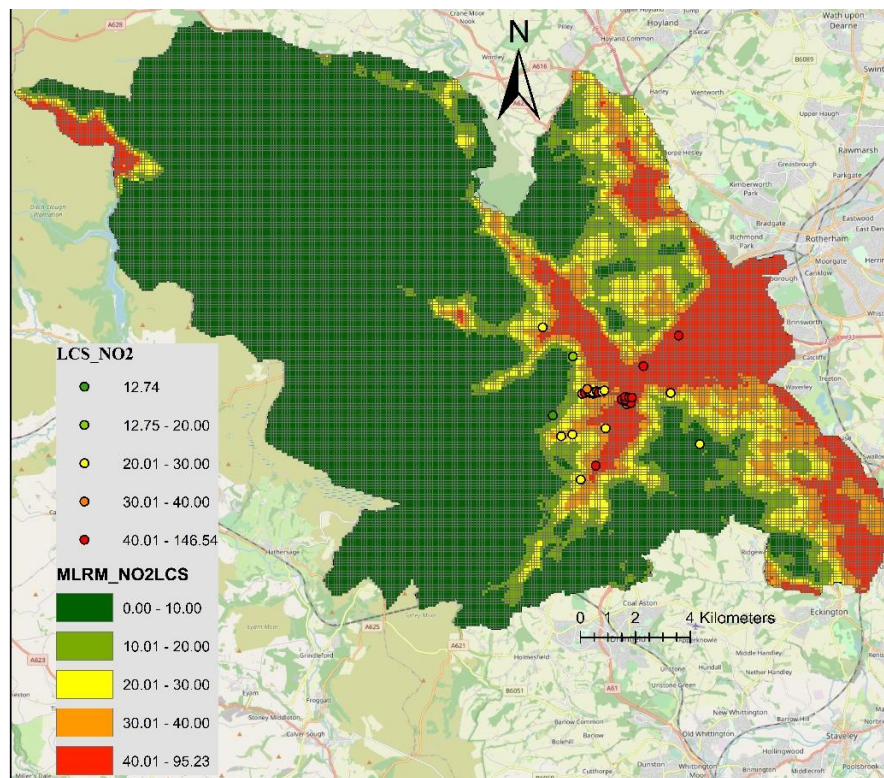
3.2. LUR Model Using NO₂ Data from LCS

Employing MLRM and GAM, NO₂ concentration ($\mu\text{g}/\text{m}^3$) from 40 LCS was modelled using the same predictor variables as in Section 3.1. Using a stepwise regression algorithm 5-covairaites were found to have a significant effect on NO₂ concentrations, which were: distance to major road, distance to minor road, distance to commercial area, distance to residential area, and altitude. Several statistical metrics for both fitted and cross validated models are shown in Table 4. Comparing fitted model, GAM with r -value of 0.89 showed better performance than the MLRM with r -value 0.55. However, when the models were applied to an independent set of data (hold-out dataset), MLRM showed better performance (r -value = 0.78) as compared to GAM (r -value = 0.56). Likewise, metrics (e.g., RMSE, MGE, NMGE, NMB) showing error of the models were lower for training dataset and higher for testing dataset for GAM than MLRM. FAC2 having value of 1 or nearly 1 shows acceptable model performance for both MLRM and GAM. The results of MLRM on hold-out dataset do not seem genuine as r -value is greater and RMSE is less than the fitted model, which probably shows that the model is over-fitted. This is also confirmed by the fact that overall both fitted and cross-validated model showed good performance, however, when the model results were extrapolated to the entire city, both MLRM and GAM models failed to predict the expected spatial variability of NO₂ concentrations in the city (Figure 6), which should have been more like Figure 5. This is probably due to the fact that LCS are mostly installed in the city centre and at the University of Sheffield, and are not representing the wider area of the city. Therefore, extrapolation of the modelled NO₂ outside the calibration range, results in unreasonable values. Secondly, there were only 40 LCS probably not providing enough spatial coverage to represent the whole city. Therefore, the models are over fitted and are not successfully fitting the data from the rest of the city.

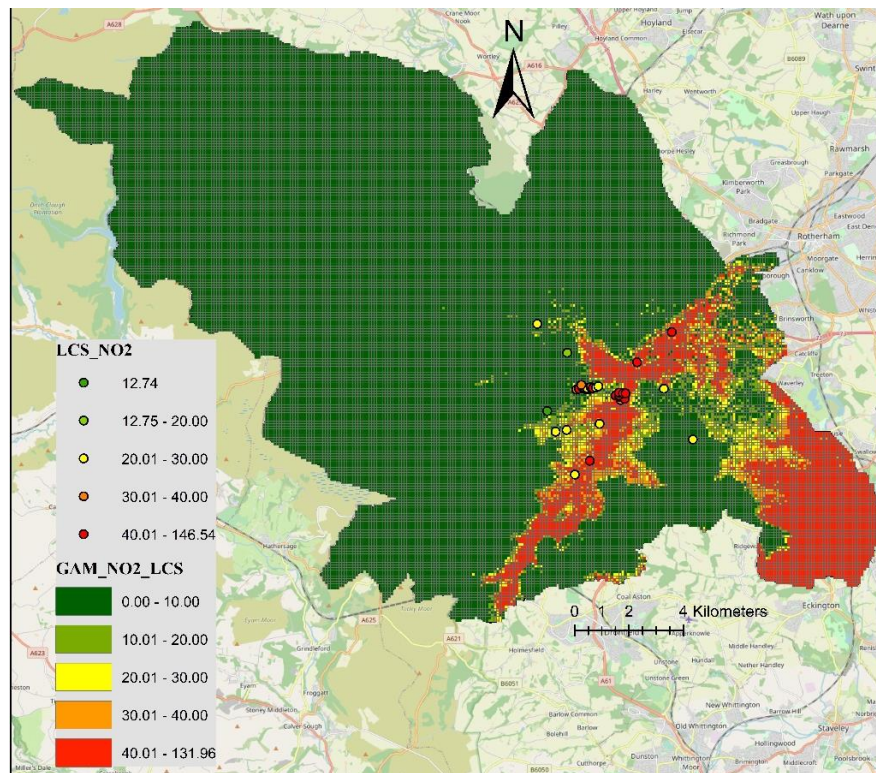
Table 4. Showing various statistical metrics for assessing the performance of MLRM and GAM for both fitted model (FM) using training dataset and cross-validation (CV) using hold-out dataset based on NO₂ data from LCS.

Metrics	MLRM		GAM	
	FM	CV	FM	CV
FAC2	0.97	1	1	0.90
MB	2×10^{-15}	−0.48	2×10^{-10}	−3.76
MGE	11.07	8.82	7.60	16.35
NMB	2×10^{-16}	−0.01	2×10^{-12}	−0.10
NMGE	0.29	0.24	0.20	0.44
RMSE	19.31	12.56	10.40	22.21
r	0.55	0.78	0.89	0.56

Gillespie et al. [18] developed an LUR model to estimate exposure to NO₂ in Glasgow, Scotland. They used 135 NO₂ passive diffusion tubes, which were divided to four groups (32–35 sites per group) and models were developed using a combination of 1 to 3 groups as training sites to assess how the number of training sites affected the model performance. The explanatory variables used in the models were major road length, minor road length, all urban areas, building volume, distance to nearest major or minor road, green rural area, minor road length, and street configuration. The models were able to explain moderate to high variance in the data, where R^2 ranged from 0.62 to 0.89 for training dataset and 0.44 to 0.85 for hold-out dataset. Precision of estimated exposure was increased with increasing number of training sites. Gillespie et al. [16] concluded that the use of more than 60 training sites in LUR model has quantifiable benefits in epidemiological application. Therefore, this might suggest that probably in present study the number of sites (40 LCS) were not enough and resulted in over fitting of the model.



(a)



(b)

Figure 6. High resolution (100 m × 100 m) maps of NO₂ concentrations ($\mu\text{g}/\text{m}^3$) in Sheffield, predicted by (a) MLRM and (b) GAM. The points show values of NO₂ measured by LCS.

3.3. LUR Model Using NO₂ Data from DT and LCS

In this section NO₂ data from 188 DT and 40 LCS were combined to increase the number of monitoring sites and see how it affects the model outputs. At the model selection stage six predictors showed significant effect, which were distance to major road, minor roads length, commercial area, population, distance to commercial area and altitude. Performance of the model was assessed by calculating several statistical metrics as shown in Table 5. In case of MLRM, *r*-values were 0.60 and 0.52 for training and hold-out dataset, whereas in case of GAM *r*-values were 0.69 and 0.53, respectively. Metrics (e.g., RMSE, MGE, NMGE, MB and NMB) expressing error of the models have smaller values for GAM than MLRM, showing GAM outperforms MLRM. FAC2 also shows that prediction of GAM is closer to the measured values. Furthermore, comparing Table 5 with Table 3, it can be clearly observed that combining DT and LCS did not improve the models performance compared to the models using DT data only. This is probably due to the fact that DT and LCS use different techniques for measuring NO₂ concentrations and combining their data might not be the right thing to do. Second, DT provide long-term NO₂ concentrations (e.g., annual mean), whereas LCS provide short-term average (e.g., 5 min to 1 h mean), which is converted to annual mean. It probably indicates that ideally all sensors should be of the same type and mixing of sensors of different grades might cause conflict and affect the model outputs. Finally, 188 DT are enough to build an LUR model, further increasing the number does not improve the model outputs.

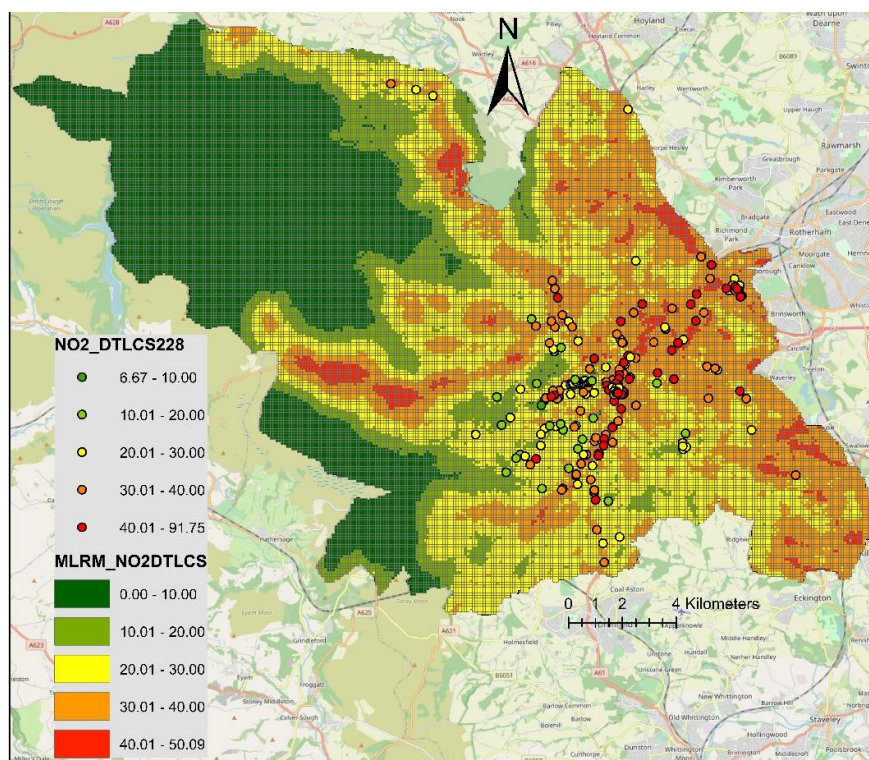
Table 5. Showing various statistical metrics for assessing the performance of the MLRM and GAM for both fitted model (FM) using training dataset and cross-validation (CV) using hold-out dataset based on NO₂ data obtained from DT and LCS.

Metrics	MLRM		GAM	
	FM	CV	FM	CV
FAC2	0.93	0.89	0.99	0.91
MB	2×10^{-15}	−2.28	−4.01	−2.03
MGE	7.31	9.32	6.49	9.21
NMB	2×10^{-16}	−0.09	−1.28	−0.07
NMGE	0.23	0.31	0.21	0.30
RMSE	9.47	12.33	8.62	12.22
r	0.60	0.52	0.69	0.53

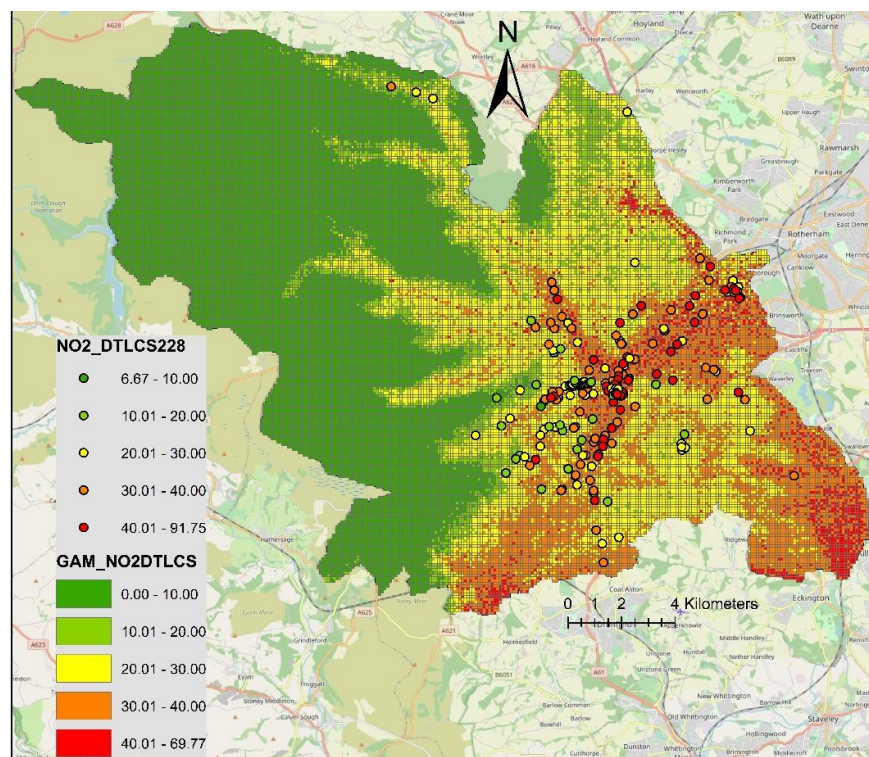
Figure 7 shows the maps of predicted NO₂ concentrations, where NO₂ concentrations predicted by MLRM range from 0 to 50 µg/m³, whereas NO₂ concentrations predicted by GAM range from 0 to 70 µg/m³. City centre and the eastern side of the city towards the motorway are highlighted as having high levels of NO₂ due to lower altitude and greater length of major and minor roads and commercial areas.

Major roads, minor roads, commercial areas and altitude had significant effects in all three models (described in Sections 3.1–3.3). Residential areas had significant effect in two models, whereas bus stops and population had significant effect only in one model. This shows that in addition to emission sources, topography especially altitude plays an important role in controlling the levels of air pollution. However, the effect of predictor variables is not the same everywhere and may change from region to region. To build an LUR model, researchers have used different sets of land use and traffic related features. The most common predictor variables used are topography, land use, traffic, population density, altitude and meteorological parameters. The use of these variables in the regression model are known to capture small scale variability [17]. The effect of predictor variables may vary from one area to another, depending on the nature of geographical conditions, size of urban area, the type of environment of the monitoring site (e.g., roadside, urban background or rural) and type of the pollutant modelled (e.g., NO₂, O₃, VOCs, PM₁₀, PM_{2.5} etc.). The main constraint for building an LUR model is the unavailability of reasonable size monitoring network to provide measured data of pollutant for fitting and validating the model. This is particularly a problem in poor and low income countries having no air quality monitoring networks. To overcome this problem, Molter et al. [32] have suggested to use

existing data from a dispersion air quality model. However, this is a problem in itself as dispersion model requires emission, geographical and meteorological data.



(a)



(b)

Figure 7. High resolution (100 m $\times$ 100 m) maps of predicted NO₂ concentrations ($\mu\text{g}/\text{m}^3$) in Sheffield, estimated by (a) MLRM, and (b) GAM. The points show values of NO₂ measured by both DT and LCS.

To the best of our knowledge, almost all previous investigations have used linear regression approach for building LUR model. Therefore, linear regression has become a default methodology for developing LUR models. This is due to the fact that in contrast to linear regressions, nonlinear regressions do not produce a single coefficient for each explanatory variable and therefore need to be applied in a different way to the traditional linear methods. With advances in IT and data analysis software (e.g., R programming language and Python), nonlinear LUR models can be developed and predicted in these software. These software have several special packages for spatial analysis and producing maps of measured or predicted pollutants. Therefore, maps of the predicted pollutants can be either produced in these software or alternatively the predicted pollutants can be exported to GIS software (e.g., ArcGIS) for further analysis. In this paper, the nonlinear model was developed using both R programming language and ArcGIS. The proposed method provides a basis for further application of advanced nonlinear modelling approaches to constructing LUR models in urban areas which enable quantifying small scale variability in pollution levels.

4. Conclusions

It has been a common practice to use the linear regression for developing LUR models, however, the association between air pollutant levels and spatial features is not always linear, therefore, ideally nonlinear modelling approaches should be used for developing LUR models, which can help in understanding small scale spatial variability of different air pollutants in urban areas. In this paper, the GAM was fitted and predicted in R programming language and the predicted concentrations was then transferred to ArcGIS for producing maps of NO₂ in Sheffield. Alternatively, R has several special packages that can be used for mapping and spatial analysis of the predicted concentrations.

In this paper spatial variability of NO₂ concentration is modelled in the city of Sheffield for year 2019. MLRM is a traditional and most commonly used approach for developing LUR models, here in addition to the linear approach, a nonlinear GAM model is employed and its benefits and the way it is applied are discussed. Three datasets of NO₂ measurements were used: (a) NO₂ data from 188 DT, (b) NO₂ from 40 LCS, and (c) NO₂ data from both DT and LCS. The first group performed better than the other two groups. Among predictor variables altitude (negative effect), major roads (positive effect), minor roads (positive effect), and commercial area (positive effect) had significant effect in all three groups. The model successfully captured the spatial variability of NO₂ in Sheffield, estimating high levels of NO₂ in the city centre and on major roads around the city. The eastern area between the city centre and motorway (M1) showed particularly high levels, whereas the western area (Peak District National Park) demonstrated lower levels of NO₂ concentrations.

The main contributions of this work are summarised as follows: (a) An advanced nonlinear GAM is proposed for developing an LUR model, which outperforms the linear counterpart. (b) High resolution maps (100 m × 100 m) of NO₂ are developed in Sheffield using a nonlinear LUR model for quantifying public exposure to NO₂ and determining how the exposure varies at small scales in the city. (c) NO₂ data measured by a network of DT and LCS are integrated to developed high resolution maps. (d) It is confirmed that Sheffield City Centre and its eastern sides experience relatively higher levels of NO₂ pollution.

Future work could include developing a spatiotemporal LUR model using meteorology, traffic counts and fleet composition data. In this study, the effect of major and minor roads is analysed on NO₂ levels, however, road traffic and composition may vary from time to time on a given road. It is, therefore, important to capture temporal and spatial variability in meteorology and traffic data and feed it to the nonlinear LUR models. This will probably further improve the model performance, depending on the quality and temporal resolution of the data.

Author Contributions: S.M., M.M. and D.C. planned and designed the original idea. S.M. obtained the data and did the modelling. M.M. and D.C. won the funding for this research project. S.M. wrote the first draft of the manuscript under the supervision of M.M. and D.C. L.S.M., M.M. and D.C. reviewed and edited the manuscript. All the authors have read and approved the content of the manuscript. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Engineering and Physical Sciences Research Council (EPSRC) (grant number - EP/R512175/1) and Siemens plc. The APC was paid by UKRI/EPSRC.

Acknowledgments: We would like to express our thanks to the Engineering and Physical Sciences Research Council (EPSRC), grant number - EP/R512175/1 and Siemens plc for funding this project. We are also thankful to Sheffield City Council for providing NO₂ diffusion tubes data.

Conflicts of Interest: All the authors declare no conflict of interest.

References

- Landrigan, P.J. Air pollution and health. *Lancet Public Health* **2016**, *2*, E4–E5. [CrossRef]
- WHO. Review of Evidence on Health Aspects of Air Pollution-REVIHAAP Project: Final Technical Report. World Health Organization Regional Office for Europe, 2013. Available online: <http://www.euro.who.int/en/health-topics/environment-and-health/airquality/publications/2013/review-of-evidence-on-health-aspects-of-air-pollutionrevihaap-project-final-technical-report> (accessed on 12 February 2020).
- DEFRA. Improving Air Quality in the UK Tackling Nitrogen Dioxide in Our Towns and Cities, UK Overview Document. December 2015. Available online: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/486636/aq-plan-2015-overview-document.pdf (accessed on 9 April 2020).
- Hsu, S.; Mavrogianni, A.; Hamilton, I. Comparing spatial interpolation techniques of local urban temperature for heat-related health risk estimation in a subtropical city. *Procedia Eng.* **2017**, *198*, 354–365. [CrossRef]
- Briggs, D.J. The Role of Gis: Coping with Space (And Time) in Air Pollution Exposure Assessment. *J. Toxicol. Environ. Health Part A* **2005**, *68*, 1243–1261. [CrossRef] [PubMed]
- Schneider, P.; Castell, N.; Vogt, M.; Dauge, F.R.; Lahoz, W.A.; Bartonova, A. Mapping urban air quality in near real-time using observations from lowcost sensors and model information. *Environ. Int.* **2017**, *106*, 234–247. [CrossRef] [PubMed]
- Briggs, D.J.; De Hough, C.; Gulliver, J.; Wills, J.; Elliott, P.; Kingham, S.; Smallbone, K. A regression-based method for mapping traffic-related air pollution: Application and testing in four contrasting urban environments. *Sci. Total Environ.* **2000**, *253*, 151–167. [CrossRef]
- Hoek, G.; Beelen, R.; De Hoogh, K.; 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**, *42*, 7561–7578. [CrossRef]
- Briggs, D.J.; Collins, S.; Elliott, P.; Fischer, P.; Kingham, S.; Erik Lebrecht, K.P.; Van Reeuwijk, H.; Smallbone, K.; Van der Veen, A. Mapping urban air pollution using GIS: A regression-based approach. *Int. J. Geogr. Inf. Sci.* **1997**, *11*, 699–718. [CrossRef]
- Eeftens, M.; Beelen, R.; De Hoogh, K.; Bellander, T.; Cesaroni, G.; Cirach, M.; Declercq, C.; Dedelè, A.; Dons, E.; De Nazelle, A.; et al. Development of land use regression models for PM_{2.5}, PM_{2.5} absorbance, PM₁₀ and PM coarse in 20 European Study areas; results of the ESCAPE project. *Environ. Sci. Technol.* **2012**, *46*, 11195–11205. [CrossRef]
- Lee, J.H.; Wu, C.F.; Hoek, G.; Hoogh, K.; Beelen, R.; Brunekreef, B.; Chan, C.-C. Land use regression models for estimating individual NO_x and NO₂ exposures in a metropolis with a high density of traffic roads and population. *Sci. Total Environ.* **2013**, *472*, 1163–1171. [CrossRef] [PubMed]
- Muttoo, S.; Ramsay, L.; Brunekreef, B.; Beelen, R.; Meliefste, K.; Naidoo, R.N. Land use regression modelling estimating nitrogen oxides exposure in industrial south Durban, South Africa. *Sci. Total Environ.* **2017**, *610*–*611*, 1439–1447. [CrossRef]
- Rahman, M.M.; Yeganeh, B.; Clifford, S.; Knibbs, L.D.; Morawska, L. Development of a land use regression model for daily NO₂ and NO_x concentrations in the Brisbane metropolitan area, Australia. *Environ. Model. Softw.* **2017**, *95*, 168–179. [CrossRef]
- Stedman, J.; Vincent, K.; Campbell, G.; Goodwin, J.; Downing, C. New high resolution maps of estimated background ambient NO_x and NO₂ concentrations in the U.K. *Atmos. Environ.* **1997**, *31*, 3591–3602. [CrossRef]
- Beelen, R.; Hoek, G.; Fischer, P.; Van den Brandt, P.A.; Brunekreef, B. Estimated long-term outdoor air pollution concentrations in a cohort study. *Atmos. Environ.* **2007**, *41*, 1343–1358. [CrossRef]
- Ryan, P.H.; LeMasters, G.K.; Biswas, P.; Levin, L.; Hu, S.; Lindsey, M.; Bernstein, D.I.; Lockey, J.; Villareal, M.; Hershey, G.K.K.; et al. A comparison of proximity and land use regression traffic exposure models and wheezing in infants. *Environ. Health Perspect.* **2007**, *115*, 278–284. [CrossRef]

17. Ryan, P.H.; LeMasters, G.K. A review of land-use regression models for characterizing intraurban air pollution exposure. *Inhal. Toxicol.* **2007**, *19* (Suppl. 1), 127–133. [\[CrossRef\]](#)
18. Gillespie, J.; Beverland, I.J.; Hamilton, S.; Padmanabhan, S. Development, Evaluation, and Comparison of Land Use Regression Modeling Methods to Estimate Residential Exposure to Nitrogen Dioxide in a Cohort Study. *Environ. Sci. Technol.* **2016**, *50*, 11085–11093. [\[CrossRef\]](#) [\[PubMed\]](#)
19. Beelen, R.; Hoek, G.; Vienneau, D.; Eeftens, M.; Dimakopoulou, K.; Pedeli, X.; Tsai, M.Y. Development of NO₂ and NO_x land use regression models for estimating air pollution exposure in 36 study areas in Europe e-The ESCAPE project. *Atmos. Environ.* **2013**, *72*, 10–23. [\[CrossRef\]](#)
20. Vienneau, D.; De Hoogh, K.; Beelen, R.; Fischer, P.; Hoek, G.; Briggs, D. Comparison of land-use regression models between Great Britain and the Netherlands. *Atmos. Environ.* **2010**, *44*, 688–696. [\[CrossRef\]](#)
21. Munir, S.; Mayfield, M.; Coca, D.; Mihaylova, L.S.; Osammor, O. Analysis of air pollution in urban areas with Airviro dispersion—A Case Study in the City of Sheffield, United Kingdom. *Atmosphere* **2020**, *11*, 285. [\[CrossRef\]](#)
22. Munir, S.; Mayfield, M.; Coca, D.; Jubb, S.A. Structuring an Integrated Air Quality Monitoring Network in Large Urban Areas—Discussing the Purpose, Criteria and Deployment Strategy. *Atmos. Environ. X* **2019**, *2*, 100027.
23. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2019; Available online: <https://www.R-project.org/>.
24. Wood, S.N.; Augustin, N.H. GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. *Ecol. Model.* **2002**, *157*, 157–177. [\[CrossRef\]](#)
25. Wood, S.N. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. B* **2011**, *73*, 3–36. [\[CrossRef\]](#)
26. Wood, S.N. *Generalized Additive Models: An Introduction with R* 2006; Chapman and Hall: London, UK; CRC Press: Boca Raton, FL, USA, 2006.
27. Hastie, T.J.; Tibshirani, R.J. *Generalised Additive Models* 1990; Chapman and Hall: London, UK, 1990.
28. Venables, W.N.; Ripley, B.D. *Modern Applied Statistics with S*, 4th ed.; Springer: New York, NY, USA, 2002; ISBN 0-387-95457-0.
29. Carslaw, D.C.; Ropkins, K. Openair—An R package for air quality data analysis. *Environ. Model. Softw.* **2012**, *27–28*, 52–61. [\[CrossRef\]](#)
30. Korek, M.; Johansson, C.; Svensson, N.; Lind, T.; Beelen, R.; Hoek, G.; Pershagen, G.; Bellander, T. Can dispersion modeling of air pollution be improved by land-use regression? An example from Stockholm, Sweden. *J. Expo. Sci. Environ. Epidemiol.* **2017**, *27*, 575–581. [\[CrossRef\]](#)
31. Ji, H.; Chen, S.; Zhang, Y.; Chen, H.; Guo, P.; Zhao, P. Comparison of air quality at different altitudes from multi-platform measurements in Beijing. *Atmos. Chem. Phys.* **2018**, *18*, 10645–10653. [\[CrossRef\]](#)
32. Molter, A.; Lindley, S.; De Vocht, F.; Simpson, A.; Agius, R. Modelling air pollution for epidemiologic research—Part I: A novel approach combining land use regression and air dispersion. *Sci. Total Environ.* **2010**, *408*, 5862–5869. [\[CrossRef\]](#)



© 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).