Research Report 236

Traffic-Related Air Pollution and Birth Weight: The Roles of Noise, Placental Function, Green Space, Physical Activity, and Socioeconomic Status (FRONTIER)

Payam Dadvand and Jordi Sunyer et al.

Additional Materials 1: FRONTIER / Developing Land Use Regression Models

---

Additional Materials 1 was reviewed by the HEI Review Committee. It has not been fully edited or formatted by HEI.

Correspondence may be addressed to Dr. Payam Dadvand, Barcelona Institute for Global Health (ISGlobal), Doctor Aiguader 88, 08003 Barcelona, Spain; email: *payam.dadvand@isglobal.org*.

Although this document was produced with partial funding by the United States Environmental Protection Agency under Assistance Award CR–83998101 to the Health Effects Institute, it has not been subjected to the Agency's peer and administrative review and may not necessarily reflect the views of the Agency; thus, no official endorsement by it should be inferred. It also has not been reviewed by private-party institutions, including those that support the Health Effects Institute, and may not reflect the views or policies of these parties; thus, no endorsement by them should be inferred.

FRONTIER / Developing Land Use Regression Models

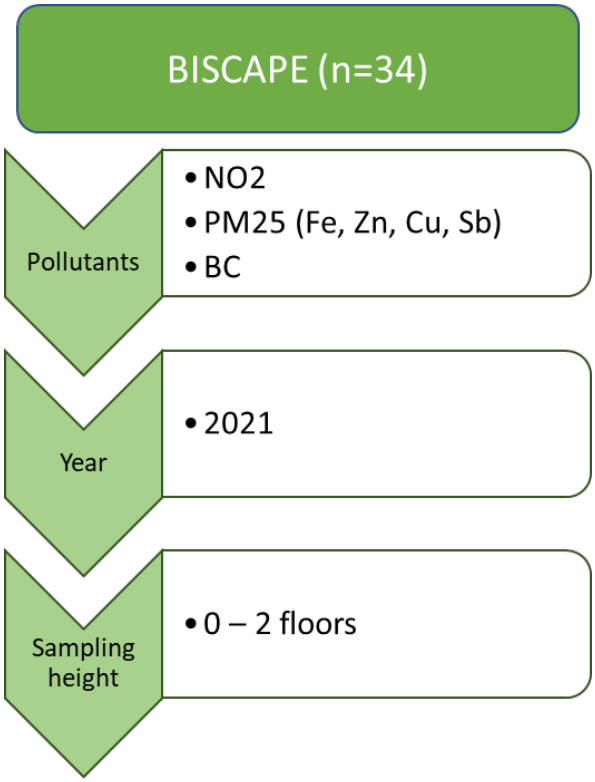
Sampling data

We developed our land use regression (LUR) models using data from two main sources:

1) BiSCAPE campaign at 34 representative locations across the Barcelona Metropolitan Area (including the reference background station of Palau Reial).
2) Measurements of home-outdoor levels at participants' homes.

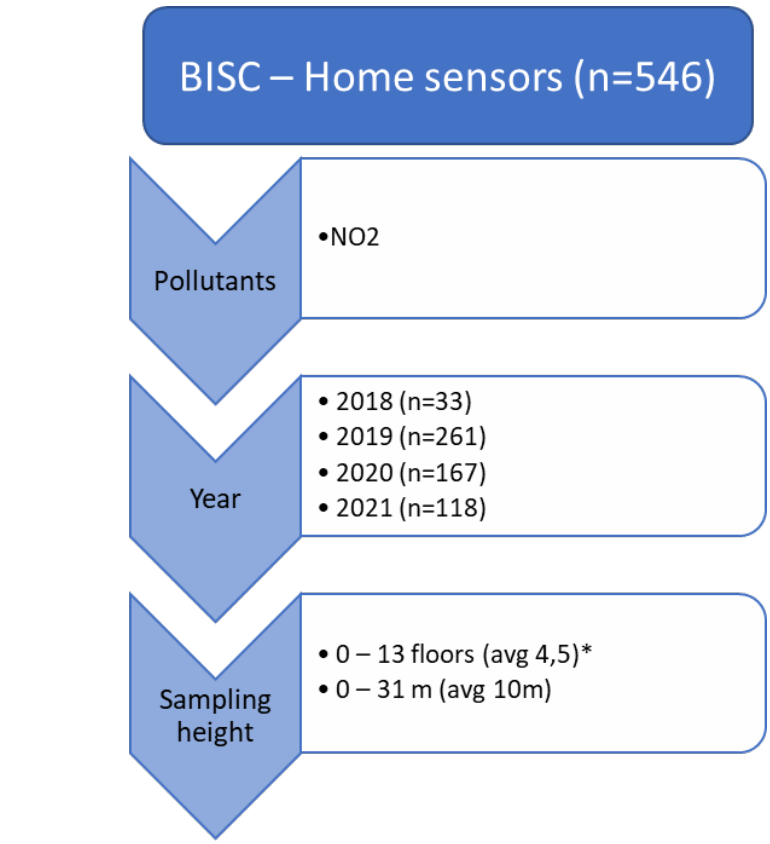
BiSCAPE campaigns were conducted according to the ESCAPE guidelines to generate data for LUR models. It encompassed four campaigns (including 3 campaigns in winter, summer, and autumn of 2021 and one campaign in winter 2022) of measuring NO_2, $PM_{2.5}$, and black carbon (BC) at 34 locations across the Barcelona Metropolitan Area (Barcelona, Esplugues de Llobregat, and L'Hospitalet).
The sites were selected following a strict site selection protocol, i.e sampling height < 4.5m, avoiding intersections, traffic lights and bus stops nearer than 25m, based on ESCAPE protocol (Beelen, et al. 2013). Figure 1 describes the BiSCAPE dataset.

Figure 1. BiSCAPE dataset characteristics.



Home-outdoor NO_2 measurements were also conducted at residential addresses of the BISC participants, one week during the first trimester and one week during the third trimester of pregnancy (2018-2021). For developing LUR models, we only included the locations that had measured home-outdoor NO_2 level data for both sampling campaigns. (i.e., We did not use data from participants who had data available for only one campaign (either first or third trimester). Almost half of the sites were measured before the COVID-19 pandemic started and half of the sites during the pandemic period and its resulting restrictions. Figure 2 describes the dataset of home-outdoor measures used to develop our LUR models.

Figure 2. Characteristics of home-outdoor measures dataset.



* 1 site at floor 20th

Deseasonalizing pollutantion data

For the development of Land-Use Regression (LUR) models data collected at different points of time need to be deseasonalized (process explained below) before being averaged to represent the average value of the baseline period.

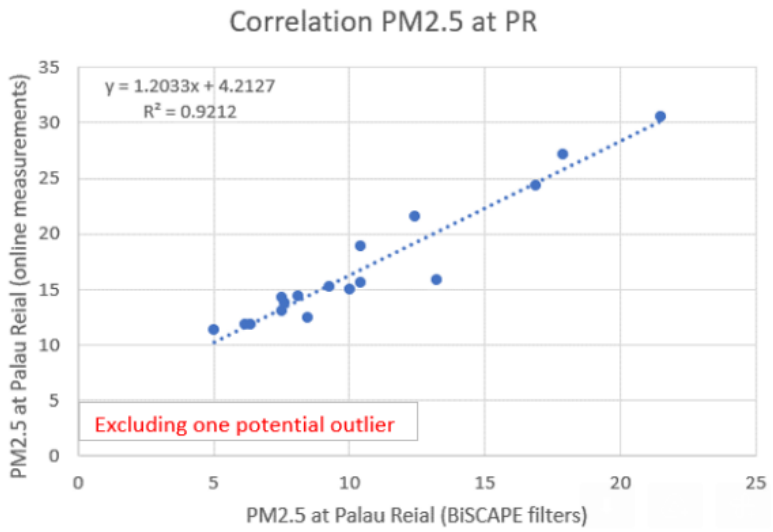
NO_2

For the NO_2 data (BiSCAPE + BiSC) the *difference method* was used for deseasonalisation.

1. Define the baseline period, that is, the total period in which the air pollution measurements were done. For NO_2 the baseline period is from 23rd October 2018 to 16th February 2022.
2. Collect hourly air pollution data from a routine monitoring site covering the baseline period. We used data from Palau Reial monitoring site, as this is the reference site for the BiSCAPE campaigns and the BiSC cohort. NO_2 data at Palau Reial was missing for only 5% of the period. Even though the missing rate is low, we imputed missing data by doing a linear regression with NO_2 levels at the monitoring site of Sants (r=0.84). After recovering those data, the percentage of missing data at Palau Reial decreased to 0.68%.
3. Calculate the average concentration of NO_2 for the routine monitoring site covering the baseline measurement period: $C_{PR}(avg)$
4. Calculate the concentrations of the pollutant at the routine monitoring site for the exact same period (t) as the sample collected at the BiSC and BiSCAPE locations in the different periods: $C_{PR}(t)$
5. Calculate for the routine monitoring site for each sample the difference between the concentration ($C_{PR}(t)$) and the average covering the baseline measurement period: $D\_C_{PR}(t) = C_{PR}(t) - C_{PR}(avg)$.
6. Calculate for each sample at site i ($C_i(t)$) the adjusted concentration ($C_{i,adj}(t)$) by computing: $C_{i,adj}(t) = C_i(t)/ D\_C_{PR}(t)$

$PM_{2.5}$, $PM_{2.5}$ components and BC

For $PM_{2.5}$, $PM_{2.5}$ components and BC the data available was obtained during the BiSCAPE campaigns. The *ratio method* was used for deseasonalisation in these datasets. The method procedure is very similar to the difference method explained before. The rationale for using a different method for these pollutants is that we aim to create LUR models for some $PM_{2.5}$ components. Since we do not have hourly data from the $PM_{2.5}$ components for the baseline period (data is only available as a daily average every fourth day), the deseasonalisation of these components needs to be done with total $PM_{2.5}$ concentrations. As the ratio method is based on substraction, this method cannot be applied to different pollutants than the reference. Also, samples collected within the BiSCAPE campaigns at Palau Reial showed lower concentrations than collocated online measurements at Palau Reial. Using the ratio method in $PM_{2.5}$ concentrations may lead to over-substracting.



1. Define the baseline period, that is, the total period in which the air pollution measurements were done. For PM2.5 and its chemical components and BC the baseline period is from 01/01/2021 to 16/02/2022.
2. Collect hourly air pollution data from a routine monitoring site covering the baseline period. We used data from Palau Reial monitoring site, as this is the reference site for the BiSC cohort.
    a. PM2.5 data at Palau Reial was missing for 26.8% of the period due instrument maintenance. We imputed missing data by doing a linear regression with collocated BC concentrations (r=0.57) as no other station in Barcelona had hourly data for $PM_{2.5}$ during this period. After recovering those data, the percentage of missing data at Palau Reial decreased to 1.0%.
    b. BC data at Palau Reial was missing for only 2.3% of the period.
3. Calculate the average concentration of the corresponding pollutant (j) for the routine monitoring site covering the baseline measurement period: $C_{j,PR}(avg)$. For $PM_{2.5}$ components, in this step *j* corresponds to total $PM_{2.5}$ mass concentrations.
4. Calculate the concentrations of the pollutant *j* at the routine monitoring site for the exact same period (t) as the sample collected at BiSCAPE locations in the different periods: $C_{j,PR}(t)$. For $PM_{2.5}$ components, in this step *j* corresponds to total $PM_{2.5}$ mass concentrations.
5. Calculate for the routine monitoring site for each sample the ratio between the concentration of pollutant *j* ($C_{j,PR}(t)$) and the average covering the baseline measurement period: $R\_C_{j,PR}(t) = C_{PR}(t)/ C_{PR}(avg)$. For $PM_{2.5}$ components, in this step *j* corresponds to total $PM_{2.5}$ mass concentrations.
6. Calculate for each sample at site i ($C_i(t)$) the adjusted concentration of pollutant *j* ($C_{i,adj}(t)$) by computing: $C_{j,i,adj}(t) = C_{j,i}(t)/ R\_C_{j,PR}(t)$.

Averages for LUR model development

Once the measurements done at different timepoints (t) and sites (i) for the different pollutants (j) were deseasonalised, we proceeded to calculate the average of the adjusted concentrations for each site *i*. Therefore, the average value obtained for each site was considered to be the average for the baseline period at site *i*. These averages were the input values for the development of the LUR model.

Creating predictor variables

To develop a LUR model, we need to calculate those spatial variables that better represent the pollutant concentration distribution across the study area. Following the ESCAPE guidelines, we selected predictor variables from three broad categories: traffic, land use, and urban configuration (see Figure 3). We selected a wide variety of buffer sizes, i.e, 25m, 50m, 100m, 300m, 500m, and 1000m to capture the area-specific predictor variables at every single location. We applied a wide range of local and regional GIS datasets to calculate the predictor variables (see Table 1). A total of 101 predictor variables have been finally created.

Figure 3. Predictor variables categories created for LUR model development.

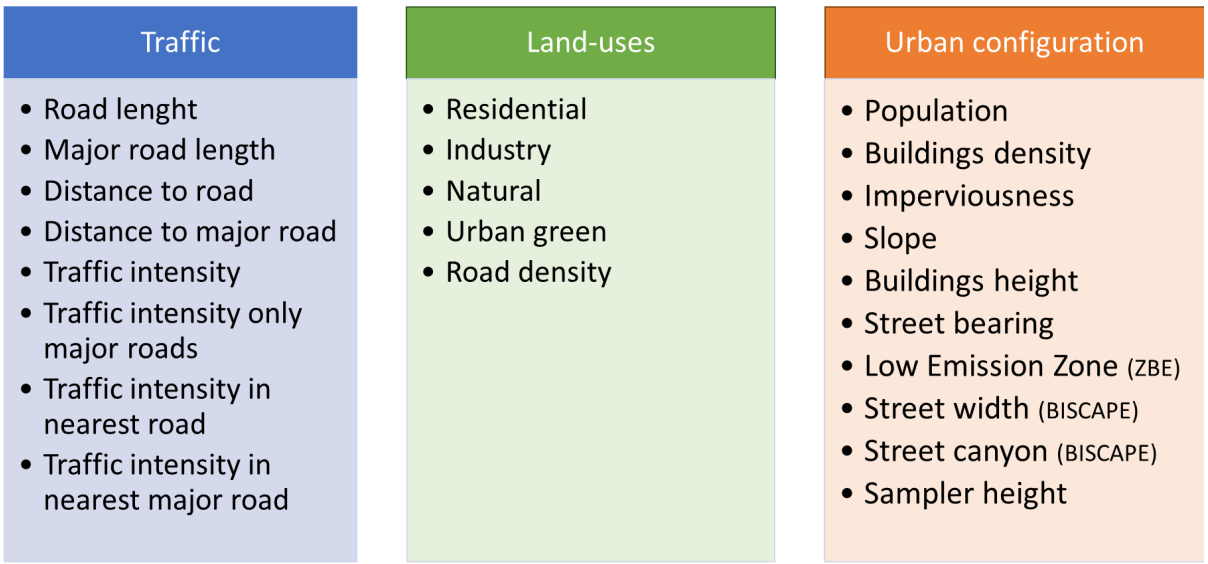


Table 1. Data sources for predictor variables.

Predictor variable	Year	Source
Road network & traffic data	2014	Departament de mobilitat, Ajuntament de Barcelona
Land uses	2018	Urban atlas, Copernicus
Population	2016	Institut d'Estadística de Catalunya.
Buildings height	2009	Modelo Digital de Superficies Edificación - MDSn2,5 (MDSNE)
Imperviousness	2018	Copernicus
Slope	2012	Institut Cartogràfic de Catalunya
LEZ	2020	OSM

* Abbreviations: LEZ: Low Emission Zone, OSM: Open Street Map.

LUR development

Following the ESCAPE guidelines (Beelen, et al. 2013), we applied a supervised forward stepwise regression process. Using pollutant concentration as an independent variable we started from a null model and added, one at a time, the predictor variable which improved the model most, following pre-defined conditions and until certain criteria were met. .

Before starting the model development analysis we defined the direction of the coefficients we expected each predictor variable to have in the model. For example, traffic variables were hypothesized to have a positive coefficient if selected in the model because they increase the response, whereas green spaces were hypothesized to reduce the concentration, therefore we expected a negative coefficient if it came into the model. Firstly, univariate regression models were conducted for all predictor variables and the model with the highest adjusted explained variance (adjusted R^2) was selected. We then repeated the process using the previous model and all the remaining predictor variables.

A variable was selected into a model if it followed the next criteria:

- It made increases the model performance (adjusted R^2) by at least 1%.
- The direction of the association (i.e., positive or negative variable coefficient) in the model was the same as what we had hypothesized.
- The p-value of the association for that variable was less than 0.10.
- Addition of that variable did not affect parameters from the previous variables (p-value and direction of the association)

Once we selected the second variable, we repeated the process until no other variable could be included and considered as the intermediate model.

Then, standard model diagnostics were applied including:

- Influential observations: the influence of each observation on the estimates was measured using Cook's Distance. It measures which observations have a large influence on the parameter estimates.
  If there were influential observations, they were removed from the model and the procedure was repeated until getting a new final model.

- Normality of the residuals using tests and plots.

- Heteroscedasticity of the residuals: a plot was made with the monitored concentrations and the residuals to evaluate whether there was heteroscedasticity.

- Spatial autocorrelation: Moran's I test was conducted to assess spatial autocorrelation.

Once the checks for the model diagnostics were accomplished, we considered the model as final.

Final LUR models

Table 2 describes the final models obtained for black carbon, $PM_{2.5}$, and its constituents (Fe, Zn, and Cu), including the number of sites finally used to develop the models, the predictor variables, the model adjusted R^2 and the LOOCV R^2.

Table 2. Final LUR models for BC, $PM_{2.5}$, Cu, Fe, and Zn.

Pollutant	Year	N	*Adj-R2*	*CV - R2*	*RSE*	Predictor variables (1)
$PM_{2.5}$	2021	34	0.47	0.45	1.47	hdres500 + trafnear + LEZ
BC	2021	30	0.85	0.83	0.18	hdres50 + linesnear + pop300 + trafload500 + roads500
$PM_{2.5}$ Cu	2021	31	0.90	0.87	0.72	trafnear + roads1000 + ind1000 + pop25
$PM_{2.5}$ Fe	2021	34	0.91	0.89	0.03	trafnear - lat + pop50 - LEZ
$PM_{2.5}$ Zn	2021	31	0.89	0.85	6.99	- LEZ + roads50 + ind1000 + build25

(1) Predictor variables (roadlength: total road length (m), majorroadlength: total major road length (m), trafload: total traffic intensity (veh/day), trafnear: traffic intensity at the nearest road (veh/day), linesnear: number traffic lines on nearest street, LEZ: Low Emissions Zone (Yes/No, ref value=No), hdres: high-density residential area (m2), roads: roads surface area (m2), ind: industry area (m2), pop: population density (inhabitants), build: building area (m2), lat: latitude (m), sqralt: squared root altitude (m1/2)).

Table 3 describes the potential NO_2 models that were explored. Models number 1 and 2 were developed using measurements done during the same year (i.e., We only included measured home-outdoor level data from those participants who had both of their measurements (1st and 3rd trimesters) in the same calendar year and excluded participants whose two measurements were in two different calendar years or those who had only data on one measurement). For these models, the deseasonalization was done separately for each year (2019, 2020, and 2021). Models 3 and 4 were developed including data from all participants with two measurements regardless of the sampling date (i.e., We only excluded participants with one measurement available). For these models, the deseasonalization was done for the entire sampling period (2018-2021).

Models 1 and 3 were built using all sites regardless of the sampling height whereas models 2 and 4 restrict sites from the ground floor to the 4th floor (included). This was informed by the findings of a study in Barcelona showing that there was only a 10% reduction in levels of air pollutants within 3 and 15 m of vertical elevation, which approximately correspond to a 4th-5th floor (Amato, F. et al, 2019).

The final model selected was number 3 based on its better validation performance and because it was built using a wider range of the sample that could be more representative of the study population. The description of this model is presented in Figure 4.

Table 3. Scenarios for the potential NO_2 LUR models.

		MODEL DEVELOPMENT							VALIDATION	
# Model	**Dataset**	**N**	**Floors**	**Year**	**N model**	**Adj R2**	**k-fold CV**	**Predictor variables**	**N**	**Adj R2**
1	**BISC - homes + BISCAPE**	384	ALL	2019-2021	315	**0.74**	**0.74**	total traffic intensity within 25m buffer + period + squared root altitude + total major road length within 50m buffer + total road length within 25m buffer + total traffic intensity within 300m buffer	602	0.51
2	**BISC - homes + BISCAPE**	250	0 to 4th	2019-2021	205	**0.70**	**0.70**	total traffic intensity within 25m buffer + period + squared root altitude + total road length within 25m buffer + otal major road length within 100m buffer + averaged buildings height within 25m buffer	602	0.49
3	**BISC - homes + BISCAPE (all period)**	579	ALL	2018-2021	489	**0.62**	**0.62**	total traffic intensity within 25m buffer + squared root altitude + total major road length within 50m buffer + total road length within 25m buffer + total major road length within 300m buffer	602	0.56
4	**BISC - homes + BISCAPE (all period)**	360	0 to 4th	2018-2021	292	**0.55**	**0.56**	total traffic intensity within 25m buffer + squared root altitude + total major road length within 100m buffer + total road length within 25m buffer	602	0.56

Figure 4. NO_2 model selected.

```
Call:
lm(formula = no2 ~ trafload25 + sqralt + majorroadlength50 +
    roadlength25 + majorroadlength300, data = dat)

Residuals:
     Min       1Q   Median       3Q      Max
-11.3101  -2.7445  -0.1588   2.5849  11.4221

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)         3.975e+01  7.812e-01  50.875  < 2e-16 ***
trafload25          8.252e-06  1.008e-06   8.188 2.40e-15 ***
sqralt             -1.259e+00  8.289e-02 -15.185  < 2e-16 ***
majorroadlength50   3.162e-02  4.878e-03   6.482 2.24e-10 ***
roadlength25        4.718e-02  6.689e-03   7.054 6.06e-12 ***
majorroadlength300  1.095e-03  2.213e-04   4.948 1.04e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.028 on 483 degrees of freedom
Multiple R-squared:  0.6254,     Adjusted R-squared:  0.6215
F-statistic: 161.2 on 5 and 483 DF,  p-value: < 2.2e-16
```

Validation process

For our LUR models for black carbon, $PM_{2.5}$, and its constituents (Cu, Fe, and Zn) we conducted internal validation using leave-one-out cross-validation (LOOCV) process. We iterated model construction over the dataset n times (where n is the total number of observations in the dataset) and each time we built the model using *n-1* observations and recorded the resulting adjusted R^2. Eventually, we averaged the metrics and provided a CV-adjusted R^2. For these models, we were not able to conduct external validation, because of the lack of external data.

For our LUR models for NO_2 we carried out both internal and external validations. For internal validation, we applied K-fold cross-validation analysis. We selected 10-fold cross-validation, therefore the sample data was divided into ten sub-samples and the model was run over each of the samples. The model was trained using the *k-1* of the folds as a training dataset and the resulting model was validated against the remaining fold. The errors and adjusted R^2 were recorded and the process was repeated *k* times. The averaged measures were reported as the indicators of the model validity.

For the external validation of NO_2 models, we applied data on measured home-outdoor NO_2 levels at participants' homes for those participants who had data available for only one measurement campaign (either the first or the third trimester measurement) and therefore their data were not used for developing LUR models.

For each candidate model, we temporally adjusted spatial estimates to abstract the predicted levels for the same home-outdoor NO_2 sampling period for each external validation point (i.e., participant's home). The results of this external validation results are shown in table 3 for each candidate NO_2 model.

Figures 6 show the distribution of the data from the different NO_2 datasets used in modelling and validation:

- Modelled data: NO_2 used to build the final model in each scenario.
- Measured data: all NO_2 outdoor measurements at participants' homes.
- Predicted data: NO_2 predicted by the models for the external validation dataset.

Figure 7 shows the association between the predicted and measured NO2 data on the external validation dataset. On the left-top side is shown the coefficient of determination.

Figure 6. Distribution of NO_2 data for the datasets used in modelling and validation.

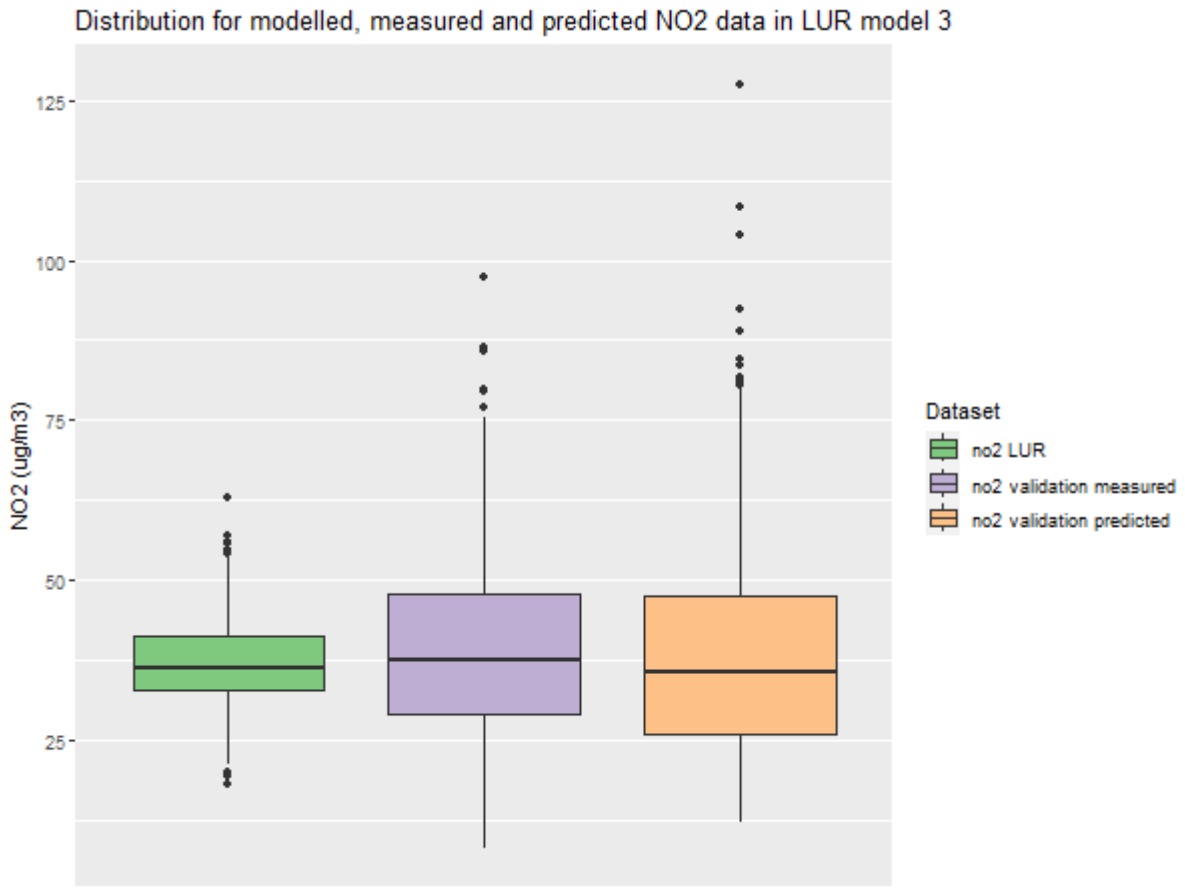


Figure 7. External validation using model predictions over the real measured data for model 3.