Environmental Pollution 289 (2021) 117832

Contents lists available at ScienceDirect

Environmental Pollution

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

Land use regression modelling of NO_2 in São Paulo, Brazil☆

Ornella Luminati [a,b], Bartolomeu Ledebur de Antas de Campos [a,b], Benjamin Flückiger [a,b], Alexandra Brentani [c], Martin Röösli [a,b], Günther Fink [a,b], Kees de Hoogh [a,b,*]

[a] Department of Epidemiology and Public Health, Swiss Tropical and Public Health Institute, Socinstrasse 57, P.O.Box, 4002 Basel, Switzerland
[b] University of Basel, Petersplatz 1, P. O. Box, 4001, Basel, Switzerland
[c] Department of Pediatrics at the Medical School of São Paulo University, São Paulo, Brazil

ARTICLE INFO

Keywords:
nitrogen dioxide
Outdoor air pollution
Population exposure
LUR
São Paulo
Brazil

ABSTRACT

Background: Air pollution is a major global public health problem. The situation is most severe in low- and middle-income countries, where pollution control measures and monitoring systems are largely lacking. Data to quantify the exposure to air pollution in low-income settings are scarce.

Methods: In this study, land use regression models (LUR) were developed to predict the outdoor nitrogen dioxide (NO_2) concentration in the study area of the Western Region Birth Cohort in São Paulo. NO_2 measurements were performed for one week in winter and summer at eighty locations. Additionally, weekly measurements at one regional background location were performed over a full one-year period to create an annual prediction.

Results: Three LUR models were developed (annual, summer, winter) by using a supervised stepwise linear regression method. The winter, summer and annual models explained 52 %, 75 % and 66 % of the variance (R^2) respectively. Cross-holdout validation tests suggest robust models. NO_2 levels ranged from 43.2 μg/m^3 to 93.4 μg/m^3 in the winter and between 28.1 μg/m^3 and 72.8 μg/m^3 in summer. Based on our annual prediction, about 67 % of the population living in the study area is exposed to NO_2 values over the WHO suggested annual guideline of 40 μg/m^3 annual average.

Conclusion: In this study we were able to develop robust models to predict NO_2 residential exposure. We could show that average measures, and therefore the predictions of NO_2, in such a complex urban area are substantially high and that a major variability within the area and especially within the season is present. These findings also suggest that in general a high proportion of the population is exposed to high NO_2 levels.

1. Introduction

Globally, approximately four million deaths are attributable to outdoor air pollution each year, with the highest burden in low and middle-income countries (WHO, 2020). The WHO estimates that 91 % of the world's population breathe polluted air, often over the recommended limits (WHO, 2020). This causes chronic and acute pulmonary as well as ischemic heart disease (WHO, 2013). Air pollution affects multiple organs (Schraufnagel et al., 2019) because of an activation of the immune response, followed by inflammatory problems (Schraufnagel et al., 2019; Babadjouni et al., 2017; Calderón-Garcidueñas et al., 2015; Genc et al., 2012). The central nervous system is one of the affected organs (Babadjouni et al., 2017). Because their rapid development, children are especially vulnerable to high air pollution levels (Schraufnagel et al., 2019; Babadjouni et al., 2017; Calderón-Garcidueñas et al., 2015).

One of the main pollutants related to negative health outcome is nitrogen dioxide (NO_2) (WHO, 2020). The sources are primarily burning fossil fuel and industrial processes (European Environment Agency, 2019). Living in proximity to a major street, resulting in exposure to vehicle pollution, shows a higher risk to develop illnesses (WHO, 2013) such as cardiovascular diseases (Babadjouni et al., 2017; Künzli et al., 2010; Bravo et al., 2016), respiratory diseases and cancer (Schraufnagel

Abbreviations: ESCAPE, European study of cohorts for air pollution effects; CETESB, Environmental Company of the State of São Paulo; GIS, Geographic information system; GPS, Global positioning system; LUR, Land use regression; NDVI, Normalized difference vegetation index; NO_2, Nitrogen Dioxide; OSM, Open Street Map; PM, Particulate matter; RMSE, Root Mean Square Error; SP-ROC, São Paulo Western Region Birth Cohort; Swiss TPH, Swiss Tropical and Public Health Institute; USGS, United States Geological Survey; VIF, Variance Inflation Factor; WHO, World Health Organization.

☆ This paper has been recommended for acceptance by Pavlos Kassomenos.

* Corresponding author. Socinstrasse 57, P.O.Box, 4002, Basel.
E-mail address: c.dehoogh@swisstph.ch (K. de Hoogh).

https://doi.org/10.1016/j.envpol.2021.117832
Received 12 May 2021; Received in revised form 30 June 2021; Accepted 21 July 2021
Available online 24 July 2021

et al., 2019; Ribeiro et al., 2019).

São Paulo city, the capital of the state of São Paulo, is located in south-eastern Brazil, approximately 50 km north-west from the Atlantic Ocean and lies at an elevation of 820 m above sea level (Schneider et al., 2020). With a population of over 20 million, São Paulo is considered a mega city that continues to spread omnidirectionally without major natural constraint. Illegal settlements (favelas), which continue to develop within São Paulo's city boundaries, are characterised by lack of city services and infrastructures, overcrowding, unhealthy living conditions and pollution (Wallenfeldt, 2019).

The city of São Paulo has been improving its air quality during the past thirty years. In this period it has reduced NO_2 emissions by about half (Andrade et al., 2017). However, NO_2 concentrations are mostly high and show large variations within neighbourhoods, depending on the sources close to measuring sites. For the whole São Paulo mega city, at the moment, there are few NO_2 measuring stations, most of them in background urban areas where they measure lower concentrations of NO_2 compared to sites in proximity to roads (The world air quality project, 2021). In 2019, sites near busy roads scarcely met the limits given by the State of São Paulo (annual average of 60 $\mu g/m^3$ NO_2), and widely exceed the WHO guideline of 40 $\mu g/m^3$ annual average (CETESB, 2020). For São Paulo we could not find information on population exposure. Reducing emissions and knowing exposure of the population is extremely important to prevent diseases. A recent study suggests that in European cities, a reduction of NO_2 to the lowest measured concentration in this study (3.5 $\mu g/m^3$) could prevent 79 435 premature deaths per year (Khomenko, 2021).

In epidemiological studies land use regression (LUR) method is widely used to estimate the exposure to air pollution in the absence of dispersion models needing accurate emission inventories (Ryan and LeMasters, 2007; Hoek et al., 2008). LUR is a method to estimate small-scale spatial variation in air pollution levels based on geographical and meteorological predictor variables. This approach has been used for predicting NO_2, NO_x, $PM_{2.5}$ and PM_{10} (Hoek et al., 2008; Araki et al., 2018; Eeftens et al., 2012; Beelen et al., 2013; Saucy et al., 2018). The method is well established in urban settings to explain small-scale variations (Beelen et al., 2013).

We conducted NO_2 measurements at 80 sites during warm and cold season in the Butantã and the Jaguaré districts of São Paulo city and then subsequently developed a LUR model for the study area. This exposure assessment will be used in further studies to estimate NO_2 levels at the residential addresses of children taking part in the São Paulo Western Region Birth Cohort (SP-ROC) (Brentani et al., 2020).

2. Methods

2.1. Study area

The study was conducted in the Western Region of São Paulo city, namely in the Butantã and the Jaguaré districts. The area is about 70 km^2 and is the home of approximately 637′000 inhabitants (census 2010) (São Paulo, 2020).

In this area the street network is complex and dense within different settlements from middle and high class neighbourhoods to informal settlements, also known as favelas. Green areas are extensive and well known for cultural and leisure activities. The area hosts also important structures for the entire city, such as the main campus of the University of São Paulo (São Paulo, 2020).

São Paulo enjoys a subtropical climate as classified by Köppen-Geiger (De Souza Rolim et al., 2007) with rainfall occurring throughout the year (IBGE and I.B.d.G.e.E., 2002), peaking in summer (October to March). In winter, July is the coldest month with an average temperature of 17 °C (average maxima and minima of 21 °C and 12 °C). Summer is hot with the highest temperatures in February, averaging 23 °C and with average maxima and minima around 28 °C and 19 °C (Schneider et al., 2020; Canty et al., 2021).

2.2. Measurements

To measure local levels of weekly average NO_2 exposure, 80 locations were selected from the residential addresses of SP-ROC study participants, similarly as done in the ESCAPE Project (Beelen et al., 2013). These sites were purposely chosen to capture the complete range of expected NO_2 concentrations and were divided in different types. The street sites (29), representing the higher concentrations, were located near busy roads. Urban background sites (43) were situated in highly populated areas away from busy roads. Regional background sites (8), representing the lower concentrations, were located in areas with lower population density, higher vegetation density, distant from roads and industrial areas. Fig. 1 shows the study area with the different types of measuring sites.

We implemented two rounds of air pollution measurements, one in the summer and one in winter. The summer campaign took place in February 2019, a hot, humid and rainy month, in which lower air pollution levels were expected. The winter campaign took place in August 2019, which is on average the driest month of the year and where highest air pollution levels were expected. Besides the two seasonal campaigns, one site was selected as a reference location where weekly NO_2 levels were measured continuously throughout one entire year. This reference location, shown in Fig. 1, was chosen in a regional background location, away from major NO_2 emission sources, easily accessible for staff to perform sampler changes, and in a safe location to avoid theft.

Each measurement campaign took place simultaneously at all 80 locations during one week. The rainy summer season measurements were taken between February 12, 2019 and February 19, 2019; the dry winter season measurements were taken between August 7, 2019 and August 14, 2019. The installation and retrieval of the measurement devices was performed by 8 groups of 2 persons within around 5 h. The groups took photos of the locations of the mounted devices as well as the surrounding area and recorded the global positioning system (GPS) coordinates by using a provided mobile phone with GPS and camera function. Each group had printed field forms for each location. These were filled out during the installation and retrieval with information about site characterization, external influences that could be observed in the immediate surroundings, and NO_2 measuring device position. To ensure reliable measurements, explicit instructions were given how and where the sampler had to be installed: within the SP-ROC-Cohort participants property, in a height of 2–3 m and on a post or fence as opposed to a wall to ensure free air circulation. Field workers were told to avoid locations near fuel stations, restaurants or street crossings. Written consent was obtained from all participants.

Week average NO_2 levels were measured using NO_2 passive gas samplers (Yu et al., 2008) from Passam AG, Switzerland. For sampler and measurement's quality control, an extra 10 % blanks and 20 % duplicates were deployed. Before and after the measurement campaign, all samplers were stored in a refrigerator. During the measurement campaign, the samplers were transported using cooling bags and cooling elements. After collection, the samplers were sent to the manufacturer for analysis.

2.3. Predictor variables

For this study, GIS datasets of sufficient resolution were identified to extract predictor variables from which to accurately capture emissions sources and the atmospheric dispersion of NO_2.

Information on road geography was gained directly in QGIS 3.4.1, from Open Street Map (OSM) (OpenStreetMap Contributors). Data on land cover, such as altitude, green space given as Normalized Difference Vegetation Index (NDVI), and built-up environment, were available from remote sensing data (based on Landsat 8 images from the U.S. Geological Survey website) (U.S. Geological Survey, 2021). Areas of informal settlements (favelas) were available from the Centro de Estudos

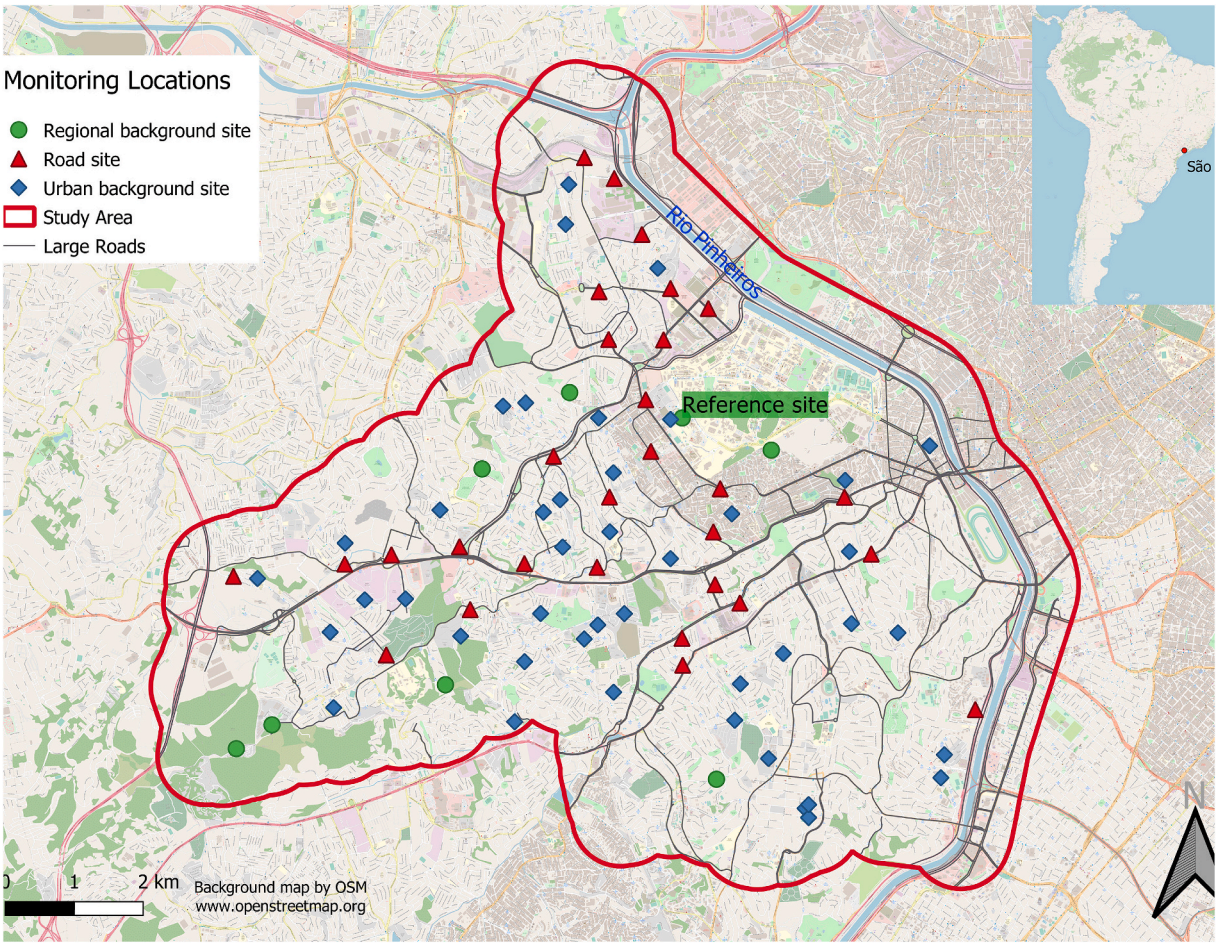


Fig. 1. Study Area with monitoring locations and, reference site.

Table 1
GIS Predictor variables.

GIS predictor variables						
Category	Source	Year	Unit	Buffer size and Transformation	Expected effect	Type
Land use (green)	OSM/ ESRI	2019	Surface (m^2)	Area in 100, 200, 300, 400, 500, 1000 buffers	–	Shapefile
Land use (industry, residential-campus)	OSM/ ESRI	2019	Surface (m^2)	Area in 100, 200, 300, 400, 500, 1000 buffers	+	Shapefile
Altitude	USGS	2019	meters above sea level	Altitude per point	–	Raster
Roads - All (motorway, trunk, primary, secondary, tertiary, residential) - size M (motorway, trunk, primary, secondary, tertiary) - size L (motorway, trunk, primary, secondary) - size XL (motorway, trunk, primary) - size XXL (motorway, trunk)	OSM	2019	Distance (m) Inverse distance (mˆ-1) Inverse distance squared (mˆ-2) Length (m)	Distance, inverse distance (id) and inverse distance squared (ids) to next road; Road length within 25, 50, 100, 200, 300, 400, 500, 1000 buffers	+	Shapefile
NDVI	USGS	2017	Mean index (−1 to 1)	Index in 100, 200, 300, 400, 500, 1000 buffers	–	Raster
Impervious surface	USGS	2017	Mean index (−1 to 1)	Index in 100, 200, 300, 400, 500, 1000 buffers	+	Raster
Traffic signals	OSM	2019	Distance (m)/Count	Distance, inverse distance and inverse distance squared to next traffic signal; Traffic signal count within 100, 200, 300, 400, 500, 1000 buffers	+	Shapefile
Fuel stations	OSM	2019	Distance (m)/Count	Distance, id and ids to next traffic signal; Traffic signal count within 100, 200, 300, 400, 500, 1000 buffers	+	Shapefile
Bus stations	OSM	2019	Distance (m)/Count	Distance, id and ids to next bus stop; Bus stop count within 100, 200, 300, 400, 500, 1000 buffers	+	Shapefile
Favelas	CEM	2016	Surface (m^2)	Area in 100, 200, 300, 400, 500, 1000 buffers	+	Shapefile

da Metrópole (Centro de Estudos da Metrópole, 2021). As a surrogate for domestic air pollution sources, we used residential land use. Land use categories were generated manually, using aerial photographs from ESRI World Imagery (Environmental Systems Research Institute (Esri), 2021), and data from OSM (OpenStreetMap Contributors). Table 1 summarizes the GIS predictor variables.

Variables such as roads, built-up environment, housing or industry were expected to increase air pollution levels, while open areas and parks or forests were expected to decrease air pollution levels. The expected effect is shown in Table 1.

Measures of distance to nearest attribute (e.g. distance to nearest street or distance to nearest bus station) were calculated directly in QGIS 3.4.4 using the NNjoin plugin and transformed into inverse distance (ID) and inverse distance squared (IDS).

All other variables were summarized using circular buffers of different sizes, representing areas where source emissions could affect the concentrations at each location. For buffer operations we used the Buffer plugin in QGIS 3.4.4. The chosen buffer sizes were 100, 200, 300, 400, 500, 1000 m for all variables. For the roads variables, buffer size of 25 and 50 m were additionally used. In a second step, GIS layers were intersected with the buffered variables using the Intersect plugin. Basic statistics were calculated for each of the buffers. Point data was summarized using the Count points in polygon plugin. Line and area data was summarized in sum length and sum area within the buffer using the attribute table calculator and the Group stats plugin. Mean values for raster data within the buffer were computed using the Zonal statistics plugin.

2.4. Statistical analysis

2.4.1. Temporal adjustment

Annual average NO_2 concentrations were calculated, similar to previous work (Cyrys et al., 2012), by combining the winter and summer measurements with a temporal adjustment using the measurements at the reference site. The annual mean of the reference location, which was measured throughout the year, was calculated and was subtracted from the summer and winter measurements at the same site. The resulting differences were then subtracted from the measurements at the 80 locations. Finally, the mean of the adjusted winter and adjusted summer measurement was calculated for each location and averaged to an adjusted annual average NO_2 concentration. These temporal adjusted annual average NO_2 concentrations were then used to develop the annual model.

2.4.2. Land use regression

A LUR model was developed by applying a multivariable linear regression model with the adjusted summer, winter and annual average NO_2 concentrations as the dependent variable and the predictor data as independent variables. The model was then applied to unmeasured locations to predict air pollution estimate values.

Three LUR models, a summer, a winter and an annual model were developed following protocols developed in the ESCAPE study (Beelen et al., 2013) using the statistical software R-Studio, Version March 1, 1093.

The models were developed by performing a supervised stepwise linear regression. First predictor variables with more than 90 % of null values were removed. A univariate linear regression between the dependent measured NO_2 values and the independent predictor variables was run. The predictor variable that explained most of the variance (highest adj. R^2) was selected first. Sequentially, from all remaining variables, the variable that maximized the observed variance was selected. This was repeated until a variable did not improve the total adjusted R^2 of the model by over 1 %. The accuracy of the model was evaluated by calculating the Root Mean Square Error (RMSE).

Only variables with a coefficient showing the expected direction were entered in the model. The variables showing non-significant p-values (<0.1) were excluded from the model. The final model was tested for correlation between predictor variables (VIF<3) and for potential highly influential sites (Cook's D < 1). Heteroscedasticity, normality and spatial autocorrelation (Moran's I, z-score -1.65 to 1.65, p-value > 0.1) of the residuals were tested to assure independency.

2.5. Validation and mapping

As suggested by Wang et al. (2016) the validation was performed using a cross-holdout validation. By offering all variables, for all N-1 locations (i.e. omitting the measurement from one location), eighty new models were developed as explained above. Each model was used to predict the location left out from the model development. The final validation R^2 was calculated by correlating the measured NO_2 values against the predicted.

Three maps were created at a 25 $\times$ 25 m grid cell level. For each center point of the 25 $\times$ 25 m grid cell the predictor variables were calculated and then used to predict the NO_2 values based on the developed LUR models.

2.6. Population exposed

The percentage of exposed population to the different NO_2 values were calculated by performing area weighting. For this the predicted concentrations surfaces and the census data (São Paulo, 2020) were used.

3. Results

3.1. Measured and temporal adjusted NO_2 values

NO_2 measurements (in $\mu g/m^3$) were sampled in 80 locations for one week in summer and one week in winter. At the reference location, NO_2 was measured continuously through the year (from February 12, 2019 to February 18, 2020). Due to missing data, two locations were left out of the models. Between the first and the second measuring campaign, two locations showed differences in the position, so for the annual model we used these as separate sites. Blanks were all within the norm and duplicates were very similar to each other the mean of the differences between duplicates was 0.9 $\mu g/m^3$ with the lowest being 0 $\mu g/m^3$ and the highest 4.6 $\mu g/m^3$.

As expected, the measurements in August, during the cold season, were higher than in February (summer). Fig. 2 shows box plots for the summer and winter campaigns in yellow and blue respectively. In green, the annual adjusted values are shown and in grey, the NO_2 annual average of the public data by the Environmental Company of the State of São Paulo (CETESB) (CETESB, 2020). For the winter, the NO_2 measurements ranged from 43.2 $\mu g/m^3$ to 93.4 $\mu g/m^3$ with a median of 56 $\mu g/m^3$. The summer measurements were lower, ranging from 28.1 $\mu g/m^3$ to 72.8 $\mu g/m^3$, the median was 36.3 $\mu g/m^3$.

The estimated annual average NO_2 concentrations ranged from 31.7 $\mu g/m^3$ to 79.2 $\mu g/m^3$ with a median of 42.7 $\mu g/m^3$. The Pearson correlation coefficient between estimated annual NO_2 means and unadjusted mean between summer and winter NO_2 measurements for all measurement sites was 0.95.

At the reference location the measured weekly concentration ranged from 17.9 $\mu g/m^3$ to 65.6 $\mu g/m^3$ (median = 36.3 $\mu g/m^3$, mean 37.7 $\mu g/m^3$). These measurements are consistent with the measurements conducted in the city of São Paulo. The most recent annual mean NO_2 concentration reported by CETESB in 2018 at the nearest location to our reference site (Cid.Universitária USP-IPEN) was 31.0 $\mu g/m^3$ (CETESB, 2020).

As expected, measuring locations near large roads reported higher NO_2 concentrations than urban background sites and regional background sites, as reported in Fig. 3, with one location very close to a busy road measuring over 70, respectively 90 $\mu g/m^3$ in the summer and

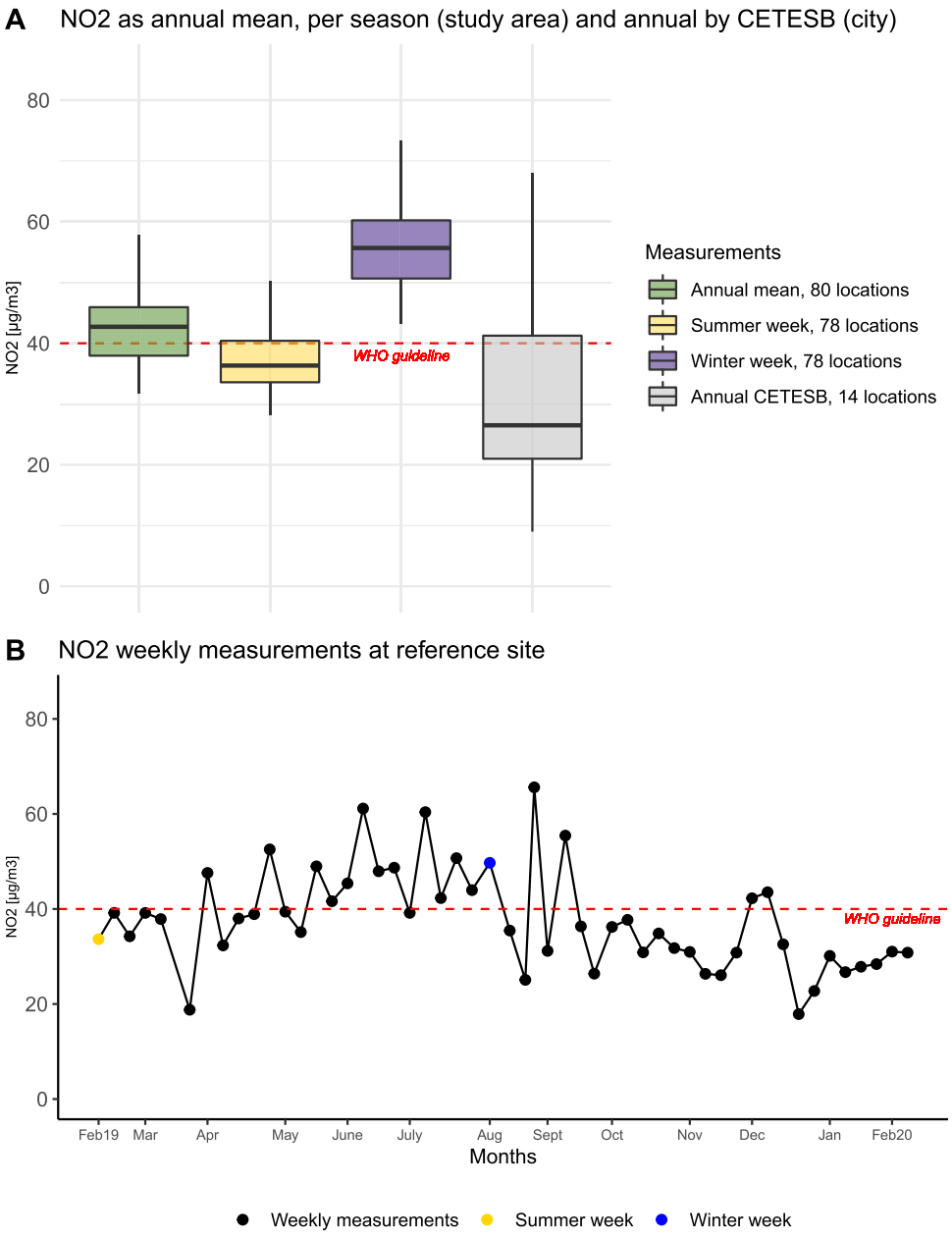


Fig. 2. A: Distribution of NO_2 measurements (annual mean, summer week, winter week and annual public data by CETESB), B: Weekly measurement of NO_2 at reference site.

winter campaigns.

3.2. LUR models

Three LUR models were developed, one for summer, one for winter, and one for the annual adjusted mean. Prior to the model development, we log transformed the NO_2 concentrations to address the skewness in the distribution. Afterwards the predicted values were back transformed.

The final models are presented in Table 2 together with the performance statistics R-Squared (R^2), root mean squared error (RMSE), the number of observations in the model (N), the mean variance inflation factor (VIF) of the model, and the Cook's D (CD). The annual NO_2 LUR model explained 66 % of the spatial variability in the NO_2 adjusted concentrations. For the annual model, 80 locations were used to build the model. The summer and winter models explained 75 % and 52 % of spatial variability in NO_2 concentrations and were built with 78 and 77 measuring sites respectively. For the winter model we excluded location 68 from the analysis, because this was an outlier (Cook's D $>$ 10). As expected, most of the predictor variables in the models were road variables. The road variables were divided by type of roads and were expressed as meters road in a buffer or as distance or inverse distance to a road. The variable bus_100, which contained the number of bus stops in a 100 m buffer around the measuring locations, was present in all three models. The annual and summer model shared five variables (Road_XXL_500, Road_XL_ids, Road_M_d, Imp_surf_300 and bus_100). The annual and winter model shared two variables (Road_M_25 and bus_100), and the summer and winter models shared one variable (bus_100).

The variables entering the models were all significant, for more information see appendix A (Fig. A.1.1, Fig. A.2.1, Fig. A.3.1). Accurate model's predictions are shown by the RMSE ranging from 3 µg/m^3 to

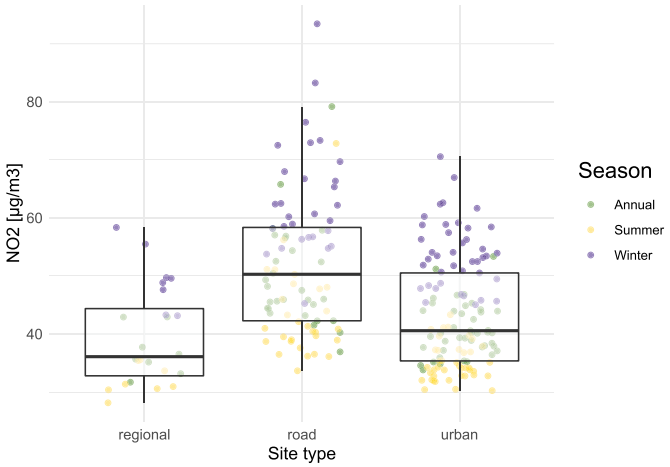


Fig. 3. Box plots of nitrogen dioxide values divided by site type.

5.4 µg/m^3 for the three models. There was no collinearity, the VIF range from 1.18 to 1.23. The sites had Cook's D values below 0.25, which means they were all influencing the model by a similar amount. The models were checked for spatial autocorrelation with the Moran's I statistic. All models showed a random distribution with Moran's I Index ranging from −0.028 to 0.02. In appendix A (Fig. A.1.2, Fig. A.2.2, Fig. A.3.2), graphs illustrating homoskedasticity are reported.

Fig. 4 presents the predicted NO_2 concentrations for annual, summer and winter. As expected, areas in proximity to roads were clearly showing higher predicted NO_2 values. The NO_2 annual average predictions ranged from 31.1 µg/m^3 to 128.8 µg/m^3. For summer and winter the predictions ranged from: 27.5 µg/m^3 to 117 µg/m^3 and 45.3 µg/m^3 to 117.1 µg/m^3 respectively.

Approximately 637′000 people live in the study area (census 2010) (São Paulo, 2020). Table 3 show the percentage of people living in the study area exposed to different ranges of NO_2 concentrations annually, in summer and winter. According to our prediction, about 67.6% of these people (i.e. 430′500 individuals) are exposed to NO_2 values exceeding the WHO guideline of 40 µg/m3 annual average. . In summer the proportion of people exposed to higher than recommended NO_2 values is around 15 %. For winter the corresponding proportion reaches 94 %.

3.3. Validation

Cross-holdout validation yielded R^2's of 0.59, 0.71 and 0.4 respectively for the annual, summer and winter model. The difference between the model R^2 and the validation R^2 were 7 % for annual, 4 % for summer and 12 % for winter, depicting robust stability of the models. The measured NO_2 values vs. the predicted NO_2 values from the cross-holdout annual, summer and winter models are plotted in Fig. 5.

4. Discussion

Land use regression models have been widely used in epidemiological studies in Europe, America and Asia (Ryan and LeMasters, 2007; Hoek et al., 2008). Only few studies applied LUR modelling for air pollution in São Paulo (Habermann and Gouveia, 2012), others used NO_2 exposures from a global LUR model (Ribeiro et al., 2019; Ribeiro et al., 2020).

To our knowledge, this is the first time a LUR model has been applied for predicting NO_2 in a city district in São Paulo. This district is particularly suited for such a model as it contains such a variety of different living, industrial and leisure areas.

Three models were developed (annual, summer, winter) of which the summer model was the best performing in terms of explained variance (0.75). The summer and annual models were the most similar and

Table 2
LUR models.

Period	LUR Model*	Model R^2	RMSE (µg/m^3)	N	VIF	CD
Annual	3.596699 +0.0016636*Road_M_25 +15.2392*Road_XL_ids - 0.0002183*Road_M_d +0.0019168*Imp_surf_300 +0.0415239*bus_100 +0.0000367*Road_XXL_500	0.66	4.04	80	1.35	<0.15
Summer	3.464911 +0.0000283*Road_XXL_500 +21.43656*Road_XL_ids - 0.0002057*Road_M_d +0.0014305*Imp_surf_300 +0.0000000689*landuse_ind_1000 +0.0548767*bus_100 +0.4536094*road_L_id	0.75	3.04	78	1.26	<0.2
Winter	4.023391 +0.0020733*Road_M_25 +0.00000825*Road_XL_1000 +0.0474643*bus_100 - 0.2994813*NDVI_300	0.52	5.41	77	1.18	<0.25

*Variables explanation.
Road_M_25 Length of roads (M) in a 25m buffer, unit: meters
Road_XL_ids Inverse distance squared to roads (XL), unit: meters^{-2}
Road_M_d Distance to roads (M), unit: meters
Imp_surf_300 Impervious surfaces given as mean index, unit free
bus_100 Number of bus stops in a 100m buffer, unit free
Road_XXL_500 Length of roads (XXL) in a 500m buffer, unit: meters
landuse_ind_1000 Surface of industry ground in a 1000m buffer, unit meters^2
Road_L_id Inverse distance to roads (L), unit: m^{-1}
Road_XL_1000 Length of roads (XL) in a 1000m buffer, unit: m
NDVI_300 Mean NDVI index in a 300m buffer, unit free

Fig. 4. Annual, summer and winter NO_2 predictive maps.

shared the highest number of predictor variables. The number of bus stations within 100 m (Bus_100) was present in all models. The cross hold out validation showed stable models.

Our measurements are consistent with the annual monitoring measurements in proximity to roads in the São Paulo state (CETESB, 2020) as with the differences between measuring location at background, residential or road sites (JúlioBarbozaChiquetto et al., 2020). The seasonality (winter higher values than summer) is due to the greater air instability and higher precipitation in summer (JúlioBarbozaChiquetto et al., 2020). One interesting and sobering finding is that, accordingly to our annual predictions, 67 % of the people living in the study area are exposed to NO_2 values exceeding the WHO guideline.

Table 3
Percentage of people living in the study area exposed to NO_2.

Percentage (%) of people living in the study area exposed to NO_2 concentrations			
	Models		
NO_2 ($\mu g/m^3$) Concentration in Categories	Annual	Summer	Winter
0–35	6.4	43.5	5.2
36–40	26.1	41.7	0.6
41–45	51	10.8	0.8
46–50	11.5	3.2	3.3
51–55	4.1	0.5	50.7
56–60	0.8	0.2	32.6
61–100	0.2	0.04	6.8

LUR models are suitable for assessing people's exposure to air pollution as these models capture the within-city variations (Crouse et al., 2015). In Australia a standard LUR model was compared with a national LUR using satellite estimates and a Bayesian blended model. They found that all three methods performed similarly and could reaffirm the standard LUR approach for predicting NO_2 in small study areas of cohort studies (Cowie et al., 2019).

The best predictor variables in our LUR models were, as expected, road or traffic-related variables in line with previous NO_2 modelling studies in urban areas where traffic related variables, like traffic count and road type, are often found to be the best predictors (Beelen et al., 2013; Lee et al., 2014), followed by land use variable and altitude (Ryan and LeMasters, 2007). Altitude is not present in any model, probably because of the low variability in altitude (less than 100 m) in our small study area.

Our models perform similar as other models in Europe (R^2 ranged from 0.55 to 0.92, median $R^2 = 0.82$, over 36 study areas) (Beelen et al., 2013), in Canada (R^2 ranged from 0.61 to 0.84, median $R^2 = 0.78$, over 10 cities) (Crouse et al., 2015), in Australia ($R^2 = 0.84$) (Cowie et al., 2019), in China (R^2 ranging from 0.42 to 0.87) (He et al., 2018) and in Japan ($R^2 = 0.68$, for all locations and $R^2 = 0.76$ for background sites only) (Kashima et al., 2018). Similar to us, Saucy et al. (2018) also developed NO_2 LUR models for annual average, cold and warm season in informal settlements in South Africa. The measured NO_2 values were lower than in our study area, showing an annual mean ranging between 9.9 $\mu g/m^3$ and 39.1 $\mu g/m^3$. Their models explained a slightly higher similar spatial variability than our models of 76 %, 62 % and 77 % in the NO_2 concentrations for the annual, warm and cold season respectively. As advised by Wang et al. (2016), to avoid overfitting problems in the validation, we did a cross hold out validation, instead of a leave-one-out cross validation. A hold out validation is not suitable in this case, because of the few locations remaining for building a model. With the performed cross hold out validation, we can be confident that our models are robust with validations R-Squared ranging from 0.40 to 0.71.

One limitation of the LUR method is the almost endless potential combinations of different variables entered as linear, quadratic or higher order terms that practically cannot be manually tested. To overcome this, machine learning systems have been used for model parameter selection instead of supervised stepwise linear regression (Araki et al., 2018; Cowie et al., 2019; Chen et al., 2020). We tested the machine learning method Random Forest on our dataset using the function Ranger in R-Studio March 1, 1093. The Random Forest model predicting annual NO_2 ($R^2 = 0.45$) did not perform better than our LUR annual model ($R^2 = 0.66$). This is probably due to the small amount of locations to be able to predict by using a machine learning algorithm.

The models were developed by using measurements at approximately 80 monitoring sites. This number of sites is at the lower end of the recommended number of monitoring sites to develop a LUR model in a complex urban setting (Basagaña et al., 2012). Another limitation of this study is the use of the predictor variables roads as a proxy for traffic, instead to use traffic counts, which have been shown to improve substantially the model development (Beelen et al., 2013). Such data,

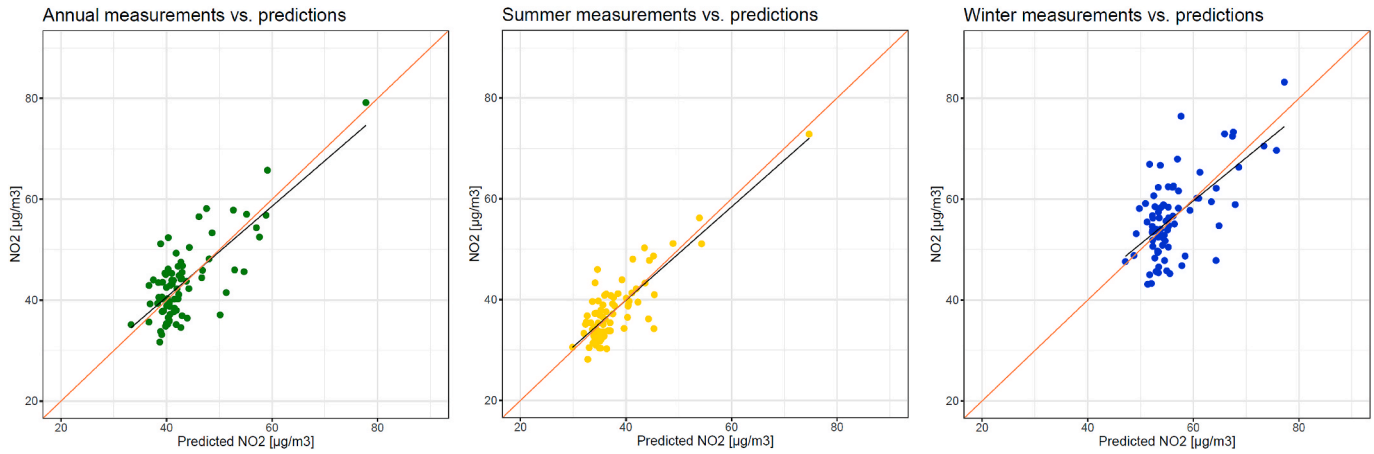


Fig. 5. Measured vs. predicted NO_2 concentrations ($\mu g/m^3$) for adjusted annual, summer and winter. The black line correspond to the regression line, the pink line correspond to the 1:1 line. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

however, were not available for the study area. A further limitation of our study is that we did not include local cooking sources, which can vary a lot between residential area and favelas and showed an impact in previous particulate matter models (Saucy et al., 2018), but are unlikely to play a major role for ambient NO_2 concentrations. Also we did not include the height of the buildings, which vary considerably between different settlements and have been shown to have an impact on the LUR development (Jin et al., 2019; Rotko et al., 2002). It is unclear how well the model performs outside the initial study area, so we recommend not to transfer these models to other parts of the city without external validation.

We developed the annual model to predict the annual NO_2 exposure at location of the residential address of the children taking part in the SPROC health study (Brentani et al., 2020).

5. Conclusion

In this study we were able to develop robust models to predict NO_2 exposure. We showed that NO_2 concentrations in our study area are generally high with a large within-study area and between-season variability. Our LUR model estimates that around 67 % of the population in the study area was exposed to NO_2 values exceeding the suggested guideline values by the WHO. The resulting NO_2 concentrations maps will facilitate epidemiological studies in São Paulo investigating the effects of air pollution on human health.

Ethical approval

Approval was obtained from the Swiss ethics committee (AO_2020–00024) and the Hospital das Clínicas at the University of São Paulo ethics committee (CAPPESQ HC_FMUSP). This article does not contain any studies involving human participants performed by any of the authors.

Funding

The Eckenstein-Geigy Professorship supported the data collection.

Author statement

Ornella Luminati, Formal analysis, Visualization, Writing – original draft. Bartolomeu Ledebur de Antas de Campos, Conceptualization, Formal analysis, Investigation, Writing – review & editing. Benjamin Flückiger, Software, Formal analysis, Investigation, Resources, Writing – review & editing. Alexandra Brentani, Investigation, Resources, Writing – review & editing. Martin Röösli, Conceptualization, Writing – review & editing. Günther Fink, Conceptualization, Writing – review & editing. Kees de Hoogh, Conceptualization, Formal analysis, Writing – review & editing, Supervision.

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.

acknowledgement

Data collection was supported by the Eckenstein-Geigy Professorship.

APPENDIX A

A. LUR models

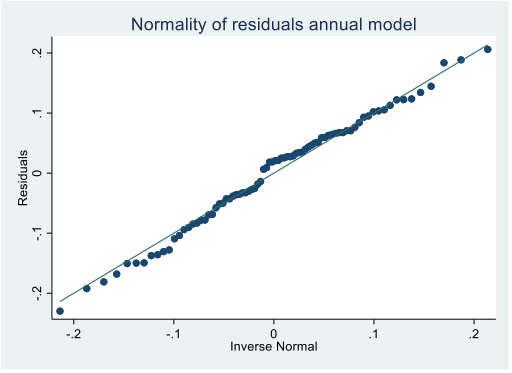
A.1. Annual model

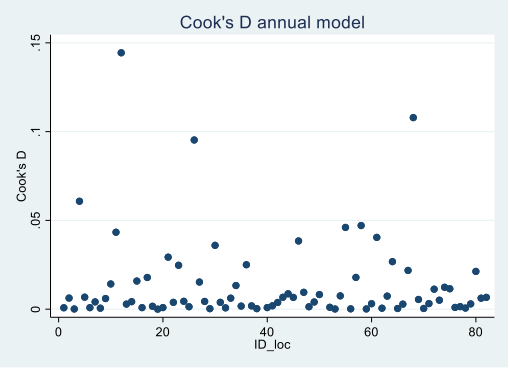
```
. regress logNO2 Road_M_25 Road_XL_ids Road_M_d Imp_surf_300 bus_100 Road_XXL_500
```

Source	SS	df	MS		
				Number of obs	= 80
				F(6, 73)	= 24.10
Model	1.41690967	6	.236151611	Prob > F	= 0.0000
Residual	.71531744	73	.009798869	R-squared	= 0.6645
				Adj R-squared	= 0.6369
Total	2.13222711	79	.026990217	Root MSE	= .09899

logNO2	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Road_M_25	.0016636	.0005065	3.28	0.002	.0006541	.0026731
Road_XL_ids	15.2392	4.907325	3.11	0.003	5.458916	25.01949
Road_M_d	-.0002183	.0001165	-1.87	0.065	-.0004505	.0000139
Imp_surf_300	.0019168	.0007682	2.50	0.015	.0003857	.0034479
bus_100	.0415239	.0162293	2.56	0.013	.0091789	.073869
Road_XXL_500	.0000367	.0000111	3.29	0.002	.0000145	.0000589
_cons	3.596699	.0527828	68.14	0.000	3.491503	3.701895

Fig. A.1.1. Regression.





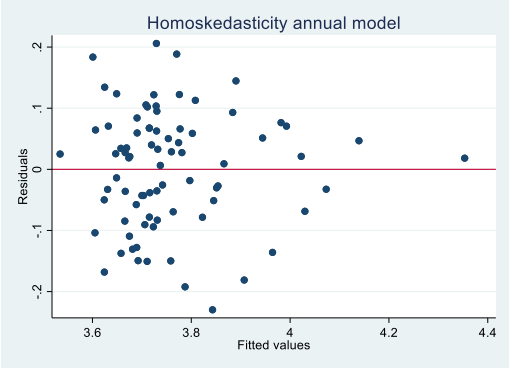


Fig. A.1.2. Normality of the residuals, Cook's D and Homoscedasticity.

A.2. Summer model

```
. regress logNO2 Road_XXL_500 Road_XL_ids Road_M_d Imp_surf_300 landuse_ind_1000 bus_100 road_L_id
```

Source	SS	df	MS		
Model	1.51637539	7	.216625055	Number of obs	= 78
Residual	.497235192	70	.00710336	F(7, 70)	= 30.50
				Prob > F	= 0.0000
				R-squared	= 0.7531
				Adj R-squared	= 0.7284
Total	2.01361058	77	.026150787	Root MSE	= .08428

logNO2	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Road_XXL_500	.0000283	.0000103	2.74	0.008	7.71e-06	.0000489
Road_XL_ids	21.43656	3.925728	5.46	0.000	13.60695	29.26618
Road_M_d	-.0002057	.0000944	-2.18	0.033	-.000394	-.0000174
Imp_surf_300	.0014305	.0006632	2.16	0.034	.0001077	.0027533
landuse_ind_1000	6.89e-08	2.29e-08	3.01	0.004	2.33e-08	1.14e-07
bus_100	.0548767	.0141814	3.87	0.000	.0265927	.0831606
road_L_id	.4536094	.1276973	3.55	0.001	.1989252	.7082935
_cons	3.464911	.0442464	78.31	0.000	3.376664	3.553158

Fig. A.2.1. Regression.

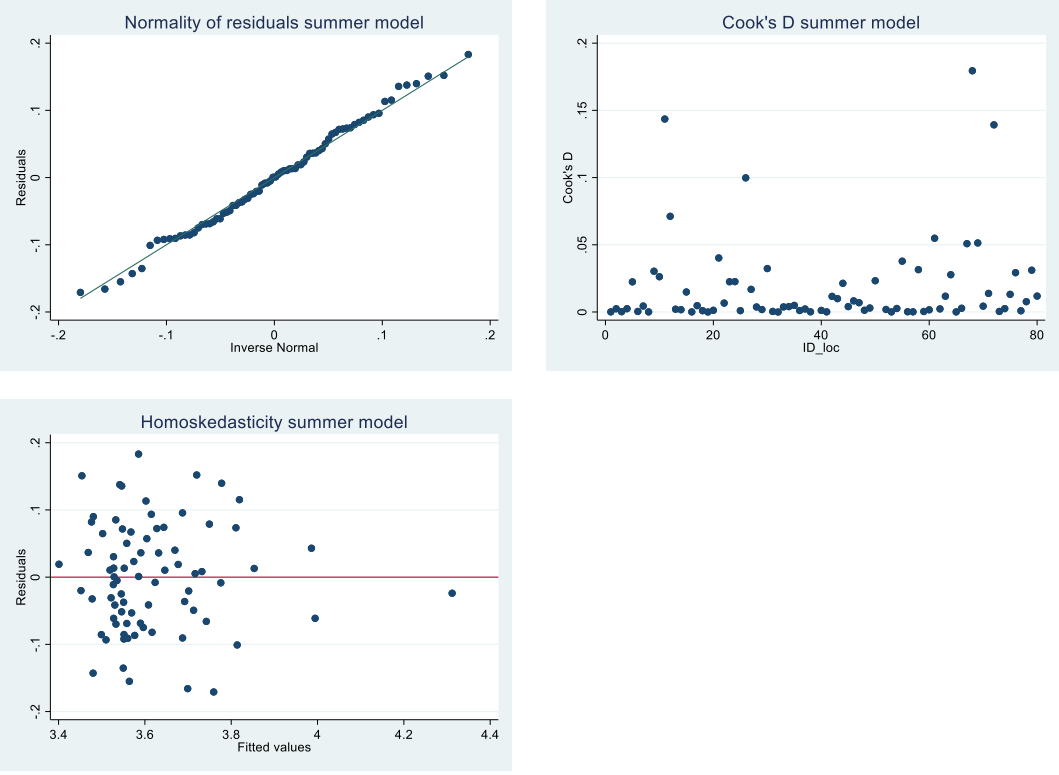


Fig. A.2.2. Normality of the residuals, Cook's D and Homoscedasticity.

A.3. Winter model

```
.
. regress logNO2 Road_M_25 Road_XL_1000 bus_100 NDVI_300
```

Source	SS	df	MS		
				Number of obs	= 77
				F(4, 72)	= 19.77
Model	.782207014	4	.195551753	Prob > F	= 0.0000
Residual	.712211229	72	.009891823	R-squared	= 0.5234
				Adj R-squared	= 0.4969
Total	1.49441824	76	.019663398	Root MSE	= .09946

logNO2	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
Road_M_25	.0020733	.0004306	4.81	0.000	.0012149	.0029317
Road_XL_1000	8.25e-06	3.69e-06	2.24	0.029	8.92e-07	.0000156
bus_100	.0474643	.0158901	2.99	0.004	.0157879	.0791407
NDVI_300	-.2994813	.1340462	-2.23	0.029	-.5666975	-.0322652
_cons	4.023391	.0525558	76.55	0.000	3.918623	4.128159

Fig. A.3.1. Regression.

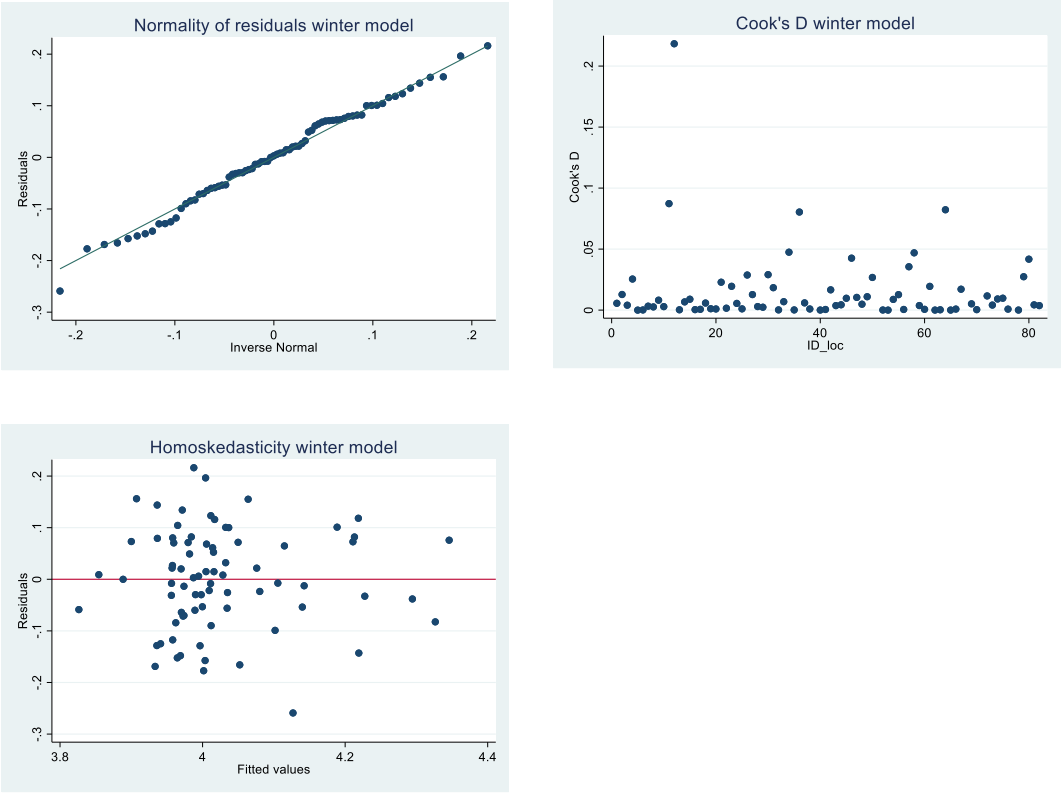


Fig. A.3.2. Normality of the residuals, Cook's D and Homoscedasticity.

References

Andrade, M.D.F., et al., 2017. Air quality in the megacity of São Paulo: evolution over the last 30 years and future perspectives. Atmos. Environ. 159, 66–82.

Araki, S., Shima, M., Yamamoto, K., 2018. Spatiotemporal land use random forest model for estimating metropolitan NO2 exposure in Japan. Sci. Total Environ. 634, 1269–1277.

Babadjouni, R.M., et al., 2017. Clinical effects of air pollution on the central nervous system; a review. J. Clin. Neurosci. 43, 16–24.

Basagaña, X., et al., 2012. Effect of the number of measurement sites on land use regression models in estimating local air pollution. Atmos. Environ. 54, 634–642.

Beelen, R., et al., 2013. Development of NO2 and NOx land use regression models for estimating air pollution exposure in 36 study areas in Europe – the ESCAPE project. Atmos. Environ. 72, 10–23.

Bravo, M.A., et al., 2016. Air pollution and mortality in São Paulo, Brazil: effects of multiple pollutants and analysis of susceptible populations. J. Expo. Sci. Environ. Epidemiol. 26 (2), 150–161.

Brentani, A., et al., 2020. Cohort Profile: São Paulo Western Region Birth Cohort. ROC.

Calderón-Garcidueñas, L., et al., 2015. Air pollution and your brain: what do you need to know right now. Prim. Health Care Res. Dev. 16 (4), 329–345.

Canty, J.L., Frischling, B., Frischling, D., 2021. Wheatherbase. São Paulo. Available from: http://www.weatherbase.com/weather/weather.php3?s=8738. (Accessed 24 June 2021).

Centro de Estudos da Metrópole. *Relatórios Favelas e Loteamentos - Estudo do CEM para Sehab/PMSP*. 21.01.2021]; Available from:: http://centrodametropole.fflch.usp.br/pt-br/downloads-de-dados/relatorios-favelas-e-loteamentos-estudo-do-cem-para-sehabpmsp.

CETESB, 2020. Qualidade do ar no estado de São Paulo 2019 [recurso electronico], p. 228.

Chen, J., et al., 2020. Development of europe-wide models for particle elemental composition using supervised linear regression and random forest. Environ. Sci. Technol. 54 (24), 15698–15709.
Cowie, C.T., et al., 2019. Comparison of model estimates from an intra-city land use regression model with a national satellite-LUR and a regional Bayesian Maximum Entropy model, in estimating NO(2) for a birth cohort in Sydney, Australia. Environ. Res. 174, 24–34.
Crouse, D.L., et al., 2015. Within- and between-city contrasts in nitrogen dioxide and mortality in 10 Canadian cities; a subset of the Canadian Census Health and Environment Cohort (CanCHEC). J. Expo. Sci. Environ. Epidemiol. 25 (5), 482–489.
Cyrys, J., et al., 2012. Variation of NO2 and NOx concentrations between and within 36 European study areas: results from the ESCAPE study. Atmos. Environ. 62, 374–390.
De Souza Rolim, G., et al., 2007. Classificação climática de Köppen e de Thornthwaite e sua aplicabilidade na determinação de zonas agroclimáticas para o estado de São Paulo. Bragantia 66, 711–720.
Eeftens, M., et al., 2012. Development of Land Use Regression models for PM(2.5), PM (2.5) absorbance, PM(10) and PM(coarse) in 20 European study areas; results of the ESCAPE project. Environ. Sci. Technol. 46 (20), 11195–11205.
Environmental Systems Research Institute (Esri). 21.01.2021]; Available from:: https://www.esri.com.
European Environment Agency, 2019. Air Pollution Sources. Available from: https://www.eea.europa.eu/themes/air/air-pollution-sources-1. (Accessed 10 July 2020).
Genc, S., et al., 2012. The adverse effects of air pollution on the nervous system. J. Toxicol. 2012, 782462.
Habermann, M., Gouveia, N., 2012. Application of land use regression to predict the concentration of inhalable particular matter in São Paulo city, Brazil. Eng. Sanitária Ambient. 17, 155–162.
He, B., Heal, M.R., Reis, S., 2018. Land-use regression modelling of intra-urban air pollution variation in China: current status and future needs, 9 (4), 134.
Hoek, G., et al., 2008. A review of land-use regression models to assess spatial variation of outdoor air pollution. Atmos. Environ. 42 (33), 7561–7578.
IBGE, I.B.d.G.e.E., 2002. Mapa de clima do Brasil.
Jin, L., et al., 2019. A land use regression model of nitrogen dioxide and fine particulate matter in a complex urban core in Lanzhou, China. Environ. Res. 177, 108597.
Júlio Barboza Chiquetto, M.E.S.S., Ynoue, Rita Yuri, Dutra Ribieiro, Flávia Noronha, Alvim, Débora Souza, Rozante, José Roberto, Cabral-Miranda, William, Swap, Robert John, 2020. The impact of different urban land use types on air pollution in the megacity of São Paulo. Revista Rresença Geografica (RPGeo) 7.
Kashima, S., et al., 2018. Comparison of land use regression models for NO(2) based on routine and campaign monitoring data from an urban area of Japan. Sci. Total Environ. 631–632, 1029–1037.
Khomenko, S., et al., 2021. Premature mortality due to air pollution in European cities: a health impact assessment. Lancet Planetary Health 5 (3), e121–e134. https://doi.org/10.1016/S2542-5196(20)30272-2.
Künzli, N., et al., 2010. Ambient air pollution and the progression of atherosclerosis in adults. PLoS One 5 (2), e9096.
Lee, J.H., et al., 2014. Land use regression models for estimating individual NOx and NO_2 exposures in a metropolis with a high density of traffic roads and population. Sci. Total Environ. 472, 1163–1171.
OpenStreetMap Contributors. OpenStreetMap. Available from:: https://www.openstreetmap.org.
Ribeiro, A.G., et al., 2019. Incidence and mortality for respiratory cancer and traffic-related air pollution in São Paulo, Brazil. Environ. Res. 170, 243–251.
Ribeiro, A.G., et al., 2020. Residential traffic exposure and lymphohematopoietic malignancies among children in the city of São Paulo, Brazil: an ecological study. Canc. Epidemiol. 70, 101859.
Rotko, T., et al., 2002. Determinants of perceived air pollution annoyance and association between annoyance scores and air pollution (PM2.5, NO2) concentrations in the European EXPOLIS study. Atmos. Environ. 36 (29), 4593–4602.
Ryan, P.H., LeMasters, G.K., 2007. A review of land-use regression models for characterizing intraurban air pollution exposure. Inhal. Toxicol. 19 (Suppl. 1), 127–133.
São Paulo, P., 2020. Mapa Digital da Cidade de São Paulo. Available from: http://geosampa.prefeitura.sp.gov.br/PaginasPublicas/_SBC.aspx#. (Accessed 7 July 2020).
Saucy, A., et al., 2018. Land use regression modelling of outdoor NO_2 and PM(2.5) concentrations in three low income areas in the western cape province, South Africa. Int. J. Environ. Res. Publ. Health 15 (7).
Schneider, R.M., Minkel, C.W., Leite, A., 2020. São Paulo. Available from: https://global.britannica.com/place/Sao-Paulo-Brazil. (Accessed 9 July 2020).
Schraufnagel, D.E., et al., 2019. Air pollution and noncommunicable diseases A review by the forum of international respiratory societies' environmental committee, Part 2: air pollution and organ systems. Chest 155 (2), 417–426.
The world air quality project, 2021. Air Pollution in Sao Paulo: Real-Time Air Quality Index Visual Map. Available from: https://aqicn.org/map/saopaulo/. (Accessed 5 June 2021).
U.S. Geological Survey. *Landsat 8 Mission* 21.01.2021]; Available from:: https://www.usgs.gov.
Wallenfeldt, J., 2019. Favela. Available from: https://www.britannica.com/topic/favela. (Accessed 21 June 2021).
Wang, M., et al., 2016. A new technique for evaluating land-use regression models and their impact on health effect estimates. Epidemiology 27 (1), 51–56.
W.H.O., 2013. Review of Evidence on Health Aspects of Air Pollution - REVIHAAP Project.
W.H.O, 2020. Health Topics - Air Pollution. Available from: https://www.who.int/health-topics/air-pollution#tab=tab_1. (Accessed 7 October 2020).
Yu, C.H., Morandi, M.T., Weisel, C.P., 2008. Passive dosimeters for nitrogen dioxide in personal/indoor air sampling: a review. J. Expo. Sci. Environ. Epidemiol. 18 (5), 441–451.