Exposure of the population of Lisbon to air pollution in the first period of confinement caused by the pandemic of COVID-19

Marco Roque
marco.roque@tecnico.ulisboa.pt

Instituto Superior Técnico, Lisboa, Portugal

November 2021

Abstract

In air quality studies it is common to use geostatistical methods to interpolate pollutant concentrations in unsampled areas. In this type of studies it is necessary to include a measure of spatial uncertainty, especially when analysing urban areas. These areas possess several sources of pollutant emissions with different intensities, which cause variations of concentrations at short distances.
In this work, the Kriging with External Drift (KED) interpolation method was used to interpolate NO_2 concentrations in unsampled areas. This method was then applied to a geostatistical simulation algorithm (SGS) in order to obtain an approximation of the real events, and quantify the spatial uncertainty.
As a first objective, this work aims to evaluate the temporal evolution of the NO_2 concentrations in the Lisbon Metropolitan Area during the first lockdown caused by the COVID-19 pandemic. The second objective was to quantify the local exposure of the populations living in a set of 86 parishes belonging to the Lisbon Metropolitan Area through geostatistical methods.
The results of the analysis the concentrations of NO_2 for the Lisbon Metropolitan Area revealed a reduction of 41.99% of the average values for the month of April 2020. While the geostatistical analysis of exposure in 2020 revealed that for the month of April, the population of the parish of Sto. António was exposed to an average concentration of NO_2 38.39% lower than the values recorded in 2019.
Keywords: Geostatistics; Kriging with external drift; COVID-19; Air quality; Exposure assessment.

1. Introduction

According to the World Health Organization (WHO), 9 out of 10 people worldwide breathe air with high concentrations of pollutants and about 4.2 million people die annually due to exposure to air pollution [40]. Therefore, air quality monitoring is a critical factor for the protection of public health especially at a time of relevant increases in population density in urban environments, and rapid economic expansion reflected in increased emissions [21]. The emergence of the SARS-CoV-2 virus and the consequent closure of various activities, particularly in the industrial sector, and considerable decrease in traffic, led to a drastic reduction in emissions of various pollutants into the atmosphere from anthropogenic activities [24, 17, 27]. Consequently, this work emerges as a unique opportunity to assess the impact of confinement due to COVID-19 on air quality. Geostatistics aims to characterize the spatial dispersion and the spatio-temporal of magnitudes that define the quantity and quality of natural resources, such as forests, geological, hydrological, ecological resources among other natural phenomena that manifest a structure in space and time [37]. Air quality modelling is positioned as an essential tool in air pollution studies. Monitoring data are indispensable to infer theories or parameters and calibrate or validate computational simulations. However, a good representation of a real phenomenon and its associated dynamics can only be achieved through a well tested and calibrated model. Design and monitoring activities should be integrated with numerical models in order to avoid investments and efforts to collect unnecessary data [30].

2. Data and settings

2.1. Study area

The study region is inserted in the Lisbon Metropolitan Area (LMA) which comprises eighteen municipalities. The river Tagus separates LMA into two zones, LMA North and LMA South, each with nine municipalities.
LMA has a population of 2,871,133 inhabitants in 2021 [8], and is the largest urban area in the coun-

try with a total area of 3,015 km^2. The study area analysed in this thesis represents a sub-area of LMA with 1,307.2 km^2 and composed of 86 parishes .

2.2. Land use data

The relationship between air pollution in urban environments and land use and land cover is well established in the literature [18]. For this work we used the information made available by Direção-Geral do Território, concerning the 2018 land use and land cover cartography in the LMA (COS2018). This information served to improve the quality of the geostatistical modelling of air pollution. The original cartography is composed of homogeneous land cover/land use units, and has a nomenclature structured in a hierarchical system with 83 land cover/land use classes [10].

In order to integrate the COS2018 information in the modelling it was necessary to develop a cartographic generalisation process, maintaining the coherence of the information. This simplification process culminated with the regrouping of the COS2018 classes in three new classes: Transports, Urban, and other uses.

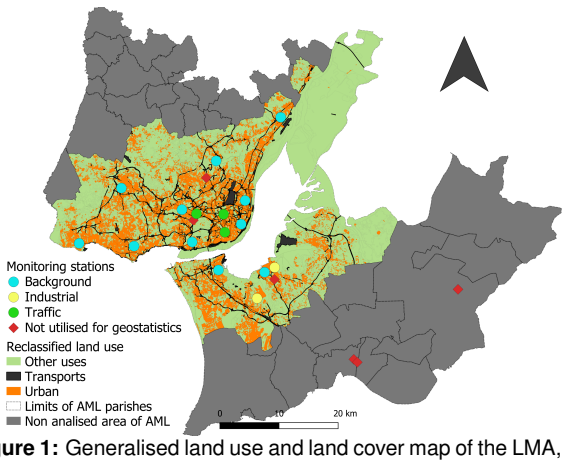


Figure 1: Generalised land use and land cover map of the LMA, and study sub-area.

2.3. Monitoring stations

The assessment of air quality in the country is carried out using Air Quality monitoring networks. The LMA has 22 stations distributed over nine municipalities, these are the responsibility of the Comissão de Coordenação e Desenvolvimento Regional de Lisboa e Vale do Tejo (CCDR-LVT). Data from these stations is measured continuously and transmitted, in semi real time, to regional concentrators and then to the central information system that is based on the QualAr database, which is housed in the Agência Portuguesa do Ambiente (APA). These data are subsequently made available to the public through the QualAr [1] platform. The stations in the air quality measurement network are installed in areas of different types (rural, suburban and urban).

Measurements were collected from 16 monitoring stations (hourly average concentrations in μ/m^3) of NO_2 observed between March and June 2018, 2019 and 2020.

3. Methodology

3.1. Data treatment

The measurements of the air quality stations provided by CCDR-LVT had missing values. Thus, it was necessary to pre-process the monitoring data in order to enable its use for modelling. In a first step, daily averages were calculated for each station and each pollutant for the days in which there were complete measurements (one day was considered has the time interval between 8h and 20h). Then, stations with less than 80% of the observations were removed from this operation. Afterwards, an imputation method was applied using the Predictive Mean Matching (PMM) algorithm. This method calculates the predicted value of the target variable Y according to the specified imputation model. For each missing value, a set of candidate donor values is generated, and this set is formed from complete cases that have a predicted value close to the missing value. A donor is randomly drawn from the candidate set, which in turn is used to replace the missing value [38].

3.2. Descriptive statistics

A descriptive analysis of the data was performed in order to synthesise measures of location and dispersion. Histograms were obtained to evaluate the empirical distribution of observations and the averages for each month in each year were estimated. The analysis of empirical distributions in each month and each year was stratified by station type. Boxplots were used to represent the distribution of pollutants by month year and station type. After intersecting station data with land cover classes, empirical distributions by land cover type were calculated.

3.3. Geostatistics

3.3.1 Linear geostatistical estimator

To model the NO_2 data, the linear estimators of Ordinary Kriging (OK) and the variant of Ordinary Kriging with External Drift (KED) were incorporated into sequential simulation algorithms.

OK is a specific case of weighted average prediction assuming a constant and unknown trend with a homogeneous and known spatial variance. Equation (1) represents the functional form of the linear estimator, where the concentration of the pollutant Z at the unsampled location s_0 is determined (predicted) with the linear estimator $\hat{z}(s_0)$, eq. (1), as a weighted average of the n neighbouring samples

$z(s)$. The coefficient λ_α, is the weight of the neighbouring sample $z(s_\alpha)$ located at the coordinates s_α [33]. The λ_α are determined so as to minimize the estimation variance and satisfy the non-bias condition, i.e., $\sum_{\alpha=1}^{n} \lambda_\alpha = 1$.

$$\hat{z}(s_0) = \sum_{\alpha=1}^{n} \lambda_\alpha z(s_\alpha) \tag{1}$$

The weights λ_α should reflect the structural dependence of the samples. One way to quantify this dependence is by estimating the semivariogram $\gamma(h)$ which measures the spatial continuity between pairs of points separated by a spatial lag distance h, eq. (2) [33].

$$\gamma(h) = \frac{1}{2} E\left\{Z(s_\alpha) - Z(s_{\alpha+h})^2\right\} \tag{2}$$

The experimental variogram (obtained from the sample data), $\hat{\gamma}(h)$ is estimated by maximum likelihood from the observed values, and is function of the distance between the pairs of observations h, eq. (2). If the experimental variogram presents spatial dependence, the values of samples that are closer to another will tend to present similar values. With the increase of the distance between points, the values between pairs of points will tend to differ more until they stop being correlated. The value of the variogram from which it is considered that correlation between the samples ceases to exist is called threshold. The estimate for the value of the sill is obtained by the sample variance.

In order to obtain a semivariogram model for any distance, it is also necessary to fit a mathematical model to an experimental variogram. The most widely used models in the field of earth and environmental sciences are the spherical, exponential or Gaussian model [33, 6]. The model used in this analysis was the spherical model, eq. (3).

$$\gamma(h) = \begin{cases} C_0 + C_e(\frac{2h}{3a} - \frac{h^3}{a^3}) & 0 < h \leqslant a \\ C_0 + C_e & h \geq a \end{cases} \tag{3}$$

In eq. (3), the parameter C_0 represents the nugget effect, C_e partial sill, $C_0 + C_e$ sill, a range and h the lag distance. The parameters C_0, C_e and a can be estimated automatically or manually. For this work a mix of the manual and automatic approach was used.
Nugget effect is assumed to exist when the value of the semivariogram near the origin is non-zero, i.e. $\gamma(h \approx 0) > 0$. This represents the part of the variance without spatial structure, i.e., a randomness or noise [33, 25].
To increase the robustness of the experimental variogram to deal with close and discrepant observations the variogram estimator used is Cressie's [7, 23].
A semivariogram was used to calculate the Kriging weights λ_α, eq. (1), for each time interval.
OK was used concurrently with Kriging With External Drift (KED) eq. (4), which is a more flexible variant of OK. In KED it is possible to combine a linear trend with the stochastic component of the spatial variance (Ordinary Kriging of the model residuals). NO_2 concentration at an unobserved point of coordinates $s(x_0)$, $\hat{z}(s_0)$, is modelled by the sum of a trend, $\hat{m}(s_0)$, and a linear combination of the neighbourhood residuals $e(s_\alpha)$. In this thesis, the functional form chosen for $\hat{m}(s)$ is the regression model with linear predictor given by land cover.

$$\hat{z}(s_0) = \hat{m}(s_0) + \sum_{\alpha=1}^{n} \lambda_\alpha e(s_\alpha) \tag{4}$$

In these type of linear models, a common problem is the possibility of returning physically impossible values (e.g., negative values). In these cases, the solution found in this thesis was to manually replace those values by admissible values. In the cases where the function returned negative value, they were replaced by 0 as proposed by Goovaerts in [14].

3.3.2 Trend

The linear regression model was chosen to predict the expected value of NO_2, with the linear predictor of land use represented by the following formula:

$$m_{ij}(s_0) = \beta_0 + \beta_1 U(s_0) + \beta_2 T(s_0) \tag{5}$$

where $m_{ij}(s_0)$ is the expected value of NO_2 for month i of year j at coordinate point s_0, $U(s_0)$ represents the value of the land cover "Urban" and $T(s_0)$ the value in the class "Transports" at point s_0. The parameters β_1, β_2 are the coefficients associated with $U(s_0)$ and $T(s_0)$ respectively, β_0 is the ordinate at the origin, and represents the expected value of NO_2 when the class at point s_0 is "Other Uses" (called the reference class). Thus, GIS was used to assign to each air monitoring station, the occupation classes predominant in its area of influence. To consider that a land use class exists in the area of influence of a station, the proportion of the area of each land cover class was calculated for a radius of influence of 1 km. The following criteria were applied to each station:

$$U(s_0) = \begin{cases} 0 & \text{, \% of coverage "Urban"} < \text{P}_{63\%} \\ 1 & \text{, \% of coverage "Urban"} \geq \text{P}_{63\%} \end{cases} \tag{6}$$

$$T(s_0) = \begin{cases} 0 & \text{, \% of coverage "Transports"} < \text{P}_{25\%} \\ 1 & \text{, \% of coverage "Transports"} \geq \text{P}_{25\%} \end{cases} \quad (7)$$

The reference class ("Other uses") takes the value 1 whenever $U(s_0) = 0 \wedge T(s_0) = 0$ or 0 otherwise, i.e., $U(s_0) = 1 \vee T(s_0) = 1$.

3.3.3 Geostatistical Simulation

Spatial uncertainty is an important aspect when modelling air pollution. Interpolation by OK and KED provides a simplification of reality, but not a measure of spatial uncertainty. Thus, one way to measure the uncertainty associated with predictions is to generate simulations with the aid of geostatistical interpolators, to generate new simulated maps that reproduce the statistical properties of the observed data [33]. The set of simulated maps allows the estimation of the expected values at each point (also referred to as *E-type* values) and an associated spatial uncertainty measure, given by the interquartile range of the simulated values.

The simulation method used was the Sequential Gaussian Simulation (SGS), where the whole simulation process is developed under an environment where it is admitted that the variable of interest has a Gaussian distribution at any point s of the spatial domain. The first step consists in transforming the original values into Gaussian values $Y(s) = \varphi[Z(s)]$, being then applied the following methodology [37]:

- Estimation at a point s_i, randomly located in area A, where the mean and variance are to be simulated. Then, a p-value is generated from a uniform distribution between 0 and 1. The simulated Gaussian value $Y(s_i)$ is then obtained from the inverse function of the local cumulative Gaussian distribution, integrating the conditional set;
- The previous step is repeated for other points, until the last $Y(s_N)$ value of A is simulated based on the conditional values;
- The simulated Gaussian map $Y(s)$ is subsequently transformed into the values $Z(s)$ by the inverse transform:

$$Z(s) = \varphi^{-1}[Y(s)] \quad (8)$$

In the simulation algorithm, the expected value of $Y(s_i)$ in the simulated point of coordinates s_i is obtained by the simple Kriging estimator. The use of simulations allows quantifying spatial uncertainty and obtaining various representations of reality (from observed data). With SGS, 300 maps were generated with statistical properties similar to those observed in the concentrations of NO_2. Finally, the variance of the set of simulations and the interquartile range were calculated for each grid point. Both these metrics allow the evaluation the dispersion of the simulations for each location.

3.4. Exposure analysis

To quantify the levels of NO_2 that the population was exposed in 2020, the areas of the parishes where the urban centres are located were considered and these areas were crossed with the simulated concentrations. Thus the average exposure was calculated only for the areas where the population lived (thus excluding from the exposure calculations the values simulated in agricultural or forested areas).

Thus, a new polygon vector layer was created with land use (vector layers represent geographic objects, with an associated geographic database [19]), composed only of the areas of the classes of continuous built coverage and discontinuous built coverage of COS2018 [10], to which the resident population of each parish was assigned [8]. For this layer a population density per inhabited area was calculated, which, in turn, when multiplied by the total population of each parish results in a vector layer with the resident population in each polygon of urban area.

The vector layer of resident population per urban area was in turn overlaid on the simulations, extracting the simulated values inside each urban area polygon (for instance, in parish A with 3 urban areas, 3 sets of simulated values were obtained). With the distribution of simulated values inside each area we estimated the average exposure in each urban area. Finally, this information was used to calculate the average exposure for each parish, this average being weighted by the population of each urban area (within the same parish).

$$E_{kij} = \frac{\sum_{l=1}^{n_k} p_l * m_l(z)}{\sum_{l=1}^{n_k} p_l} \quad (9)$$

In eq. (9), E_{kij} is the estimated population exposure in parish k, in month i and year j, p_l is the resident population in urban area $l(l = 1, \ldots, n_k)$, and $m_l(z)$ is the mean of the empirical distribution of simulated values of z (variable NO_2) in urban area l, contained in parish k.

In order to establish a metric for comparing the average exposure with each of the previous years (2019 and 2018), non-parametric bootsrapping techniques were used using the empirical distribution of the estimated values in each parish (in each month and year). Confidence intervals were defined for E_kij using the 2.5% and 97.5% percentiles of the empirical distribution and the differ-

ences between exposure in 2020 and exposure in each of the previous years were compared. Differences are considered significant when the rejection areas of the distributions do not overlap (for a 95% confidence interval).

3.5. Software

Statistical analysis was performed in R language, supported by the integrated development environment (IDE) RStudio [11, 35]. Missing data imputation was performed using the MICE package of R [39]. Interpolation and geostatistical simulation was performed using the gstat package of R [15]. QGIS software was used as an aid to R to prepare raster and vector data [31].

4. Results

4.1. Descriptive statistics

4.1.1 NO_2

NO_2 presents a significant reduction in its concentrations in 2020 when compared to previous years. In the histograms of average daily values, it is visible in 2020 an increase in the frequency of observations with lower values, fig. 2. In the same figure, it is visible that the mean values of concentrations in the months analysed (represented by the dashed lines) decreased in 2020 when compared to 2019 and 2018. This, being more evident in the months of April to June. Observing the homologous variation of the monthly average values between the months of March and June, a reduction in the monthly averages of 32.99% (March), 41.99% (April), 32.84% (May) and 9.14% (June) between 2019 and 2020, table 1.

Table 1: Monthly averages of NO_2 concentrations in LMA.

	NO_2 $(\mu g/m^3/month)$			
Year	March	April	May	June
2018	17.88	19.64	18.71	16.96
2019	23.13	17.05	17.69	13.59
2020	15.50	9.89	11.88	9.63

By station type, the reduction in NO_2 concentrations is more pronounced at traffic stations compared to the reductions observed at background or industrial stations. In traffic stations there is also a marked narrowing of the variability around the median, and the absence of outliers in April, which corresponds to the first full month experienced under severe containment measures, fig. 3.

4.2. NO_2 concentrations and land use

Linear regression models were fitted to predict the expected value of NO_2 as a linear combination of land cover. The resulting residuals were then used in geostatistical modelling.

Models were fitted for each month and year, and

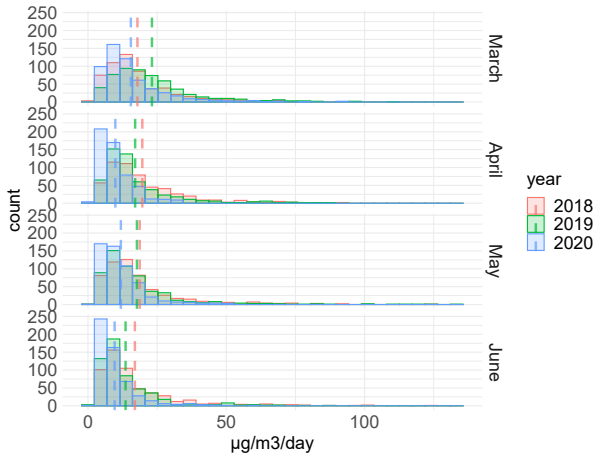


Figure 2: Histograms with empirical distribution of NO_2 in April-June, by year.

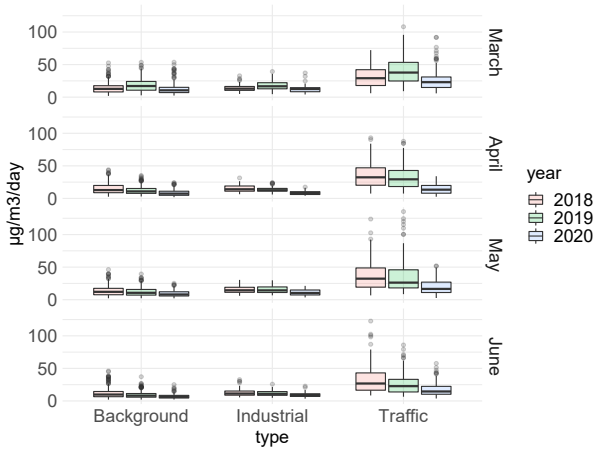


Figure 3: Boxplots of daily average concentrations of NO_2 for each month and year, by type of station.

the results obtained suggest that the impact of the "Transport" class is significant in the variation of the expected value for NO_2. The table 2 shows the results for the month of April:

Table 2: Estimated coefficients of the three regression models (1 per year) fitted for the month of April. The * symbol indicates a p-value < 0.05.

Year	Constant	Transports	Urban
2018	*12.10	*14.84	4.06
2019	9.02	*15.62	4.08
2020	*7.35	*5.32	1.43

By the analysis of the significance of the estimated parameters, it is observed that the soil class "Transports" has a significant impact (for a p-value $= 0.05$) in the variation of the expected value of the NO_2 in the three years analyzed. In the case of the "Urbanized" soil class, on the other hand, the results were not significant (i.e. p-value > 0.05). This pattern of results was repeated for the remaining months analysed.

The overall results of the adjustments obtained (by

month and year) can be summarized from the measure of the quality of adjustment provided by the coefficient of determination, R^2. Table 3 shows the results of the adjusted-R^2 obtained in the four months for the three years. It can be seen that the fitted linear models explain between 53% and 21% of the variability of NO_2 concentrations suggesting that the linear model fits well to measure the relationship between land use and average NO_2 concentrations.

Table 3: Adjusted-R^2 linear models used in the KED.

Year	March	April	May	June
2018	0.53	0.43	0.35	0.33
2019	0.37	0.43	0.24	0.35
2020	0.32	0.43	0.31	0.21

When analysing the distribution of the residuals for the year 2020 in table 4, it can be seen that the linear model fitted better in April than in the other months, since the median is closer to 0 (-0.2) and the range of values of the residuals (8.2) is the smallest among those analysed in 2020.

Table 4: Distribution of residuals generated by linear regression models for the year 2020.

Month	Min	1stQ	Median	3rdQ	Max
March	-8.5	-4.5	0.5	2.6	15.1
April	-3.9	-2.0	-0.2	1.9	4.3
May	-6.7	-2.6	-0.8	2.6	11.1
Junho	-8.1	-2.7	-0.2	2.1	14.7

4.3. Geostatistics

4.3.1 Sequential Gaussian Simulation

The parameters of the theoretical semivariograms adjusted in all the months and years analysed are presented in table 5 and indicate that major changes in the parameters occurred in the months of April, May and June(range and sill).

For March, the values of the estimated sills of the years 2018 and 2020 were much lower than those estimated for 2019. However, similar ranges are observed in 2019 and 2020 (9.1 km and 8.9 km respectively).

In April, the range value was maximum in 2020, and 2.5 times higher than the range estimated in 2018. On the other hand, the estimated sill for 2020 was 8-9 times lower than those estimated in previous years.

In May the estimated threshold for 2020 was 5 and 7 times lower than the estimated thresholds for 2018 and 2019 respectively. The estimated ranges in the three years ranged from 9 km to 12.5 km.

In June the range parameter varied between 9.2 km and 12.4 km and the sill between 26.7 and 90.9 $(\mu g/m^3)^2$.

In the year 2020 (last row of the table 5) the estimated sill decreased from 37 to 7 $(\mu g/m^3)^2$ (81% decrease) between March and April. Thereafter a gradual increase of the sill was observed reaching 26.7 $(\mu g/m^3)^2$ in June. In the opposite direction, the range value increased from 8.9 km to 15 km (59%) between March and April. In the following months a stabilization of the range value was observed, in line with the evolution observed in previous years.

Table 5: Parameters of the fitted semivariogram in each month and year. a - range, in meters, C_1 - sill, in $(\mu g/m^3)^2$.

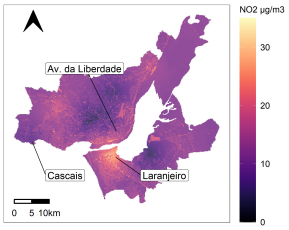
March			April		
Year	a	C_1	Year	a	C_1
2018	5256	38.42	2018	6000	61.59
2019	9167	86.49	2019	10158	65.84
2020	8881	36.97	2020	14997	6.98

May			June		
Year	a	C_1	Year	a	C_1
2018	8666	98.89	2018	12320	90.93
2019	12550	140.56	2019	9259	55.33
2020	10975	20.19	2020	12378	26.73

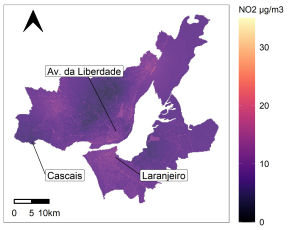
The fitted variogram models and the values of the residuals calculated at the different stations were used to generate 300 simulations with SGS algorithm and obtain estimates of the concentrations in the areas where the population resides. From this set of simulations, thematic maps representing the mean and interquartile range of the 300 simulations were generated. For the year 2020, these maps are represented in fig. 4 (mean), fig. 5 (interquartile range) and provide an estimate of the mean exposure and spatial uncertainty at each node of the simulation grid.

In the maps of the average of the simulations it is possible to observe the contour of the soil classes "Transports" and "Urban" fig. 4, being these zones signalled by the presence of higher values of NO_2 concentration. There is a reduction in the maximum levels observed from April onwards, a hotspot of high concentrations is always being identified in the Almada area.

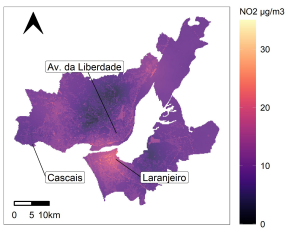
In the interquartile range maps, the areas with lower spatial uncertainty (i.e. smaller interquartile range) coincide with the areas closer to the stations, due to the presence of observed data. After March there is a reduction in the dispersion of the values of the 300 simulations. In addition to the reduction of the dispersion of the values occurred, there is also a large increase in the spatial continuity of the maps (especially in the month of



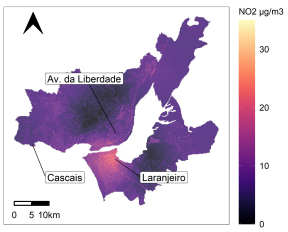
(a) Average of March 2020.



(b) Average of April 2020.



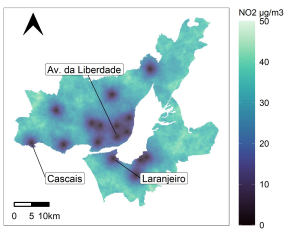
(c) Average of May 2020.



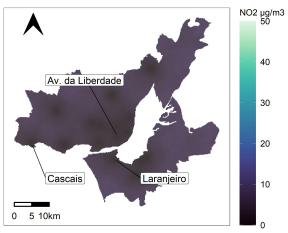
(d) Average of June 2020.

Figure 4: Maps of the local average NO_2 concentration from the 300 simulations generated by SGS.

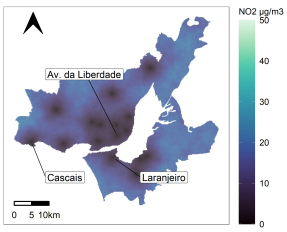
April), presenting a more uniform distribution values throughout the extent of the spatial domain as illustrated in fig. 5.



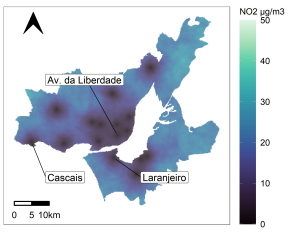
(a) Interquartile range March 2020.



(b) Interquartile range April 2020.



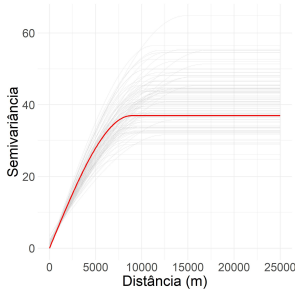
(c) Interquartile range May 2020.



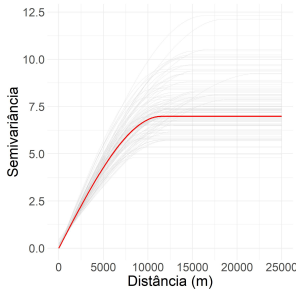
(d) Interquartile range June 2020.

Figure 5: Local interquartile range maps of the 300 simulations generated by SGS.

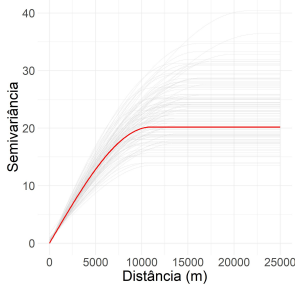
From the semivariograms of the 2020 simulations, we can also observe a decrease in semivariance between the months of March and April. Between April and June there is a gradual increase of this magnitude. There is a greater dispersion of the semivariance of the simulations (identified in grey) in values higher than the original semivariance (identified in red) fig. 6.



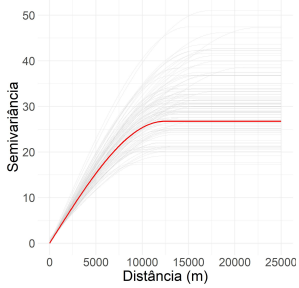
(a) Semivariograms for March 2020.



(b) Semivariograms for April 2020.



(c) Semivariograms for May 2020.



(d) Semivariograms for June 2020.

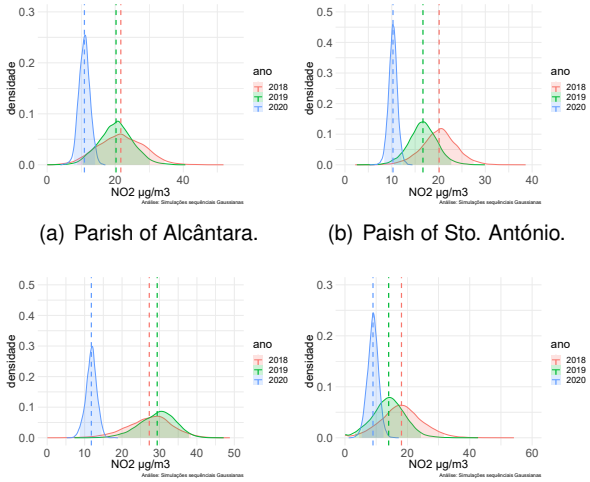
Figure 6: Theoretical semivariograms of the first 100 simulations generated by SGS (grey) and the variogram model fitted to the observed data (red)

4.3.2 Exposure analysis

For each month and each parish, the mean and the 2.5% and 97.5% percentiles (bootstrap method) of the empirical distribution of the 300 simulations were calculated. This approach made it possible to obtain a average exposure value and a 95% confidence interval for the mean.

In the density plots of the empirical distributions of the simulations of the NO_2 concentrations four parishes with sets of different characteristics (urbanization levels and location) are represented fig. 7. It is possible to observe from the graphs that, in the year 2020, the distributions of observations present lower mean values, with a range of values also lower, resulting in narrower confidence intervals (observations are less dispersed around the mean) when compared to the years 2018 and 2019.

In the parish of Sto. António, a reduction in 38.39% of the average concentrations to which the population was exposed between 2019 and 2020 occurred for April. There were reductions in



(a) Parish of Alcântara.

(b) Paish of Sto. António.

(c) Union of the parishes of Almada, Cova da Piedade, Pragal and Cacilhas.

(d) Union of the parishes of Oeiras, S. Julião da Barra, Pç. de Arcos and Caxias.

Figure 7: Density plots of the empirical distributions of the simulations of NO_2 concentrations in 2018, 2019 and 2020, in the month of April, in four different parishes.

the order of 59.85% for the Union of parishes of Almada , 45.92% in Alcântara and 35.83% in the in Union of parishes of Oeiras, table 6.

Table 6: Average population exposure to observed concentrations of NO_2 in the month of April.

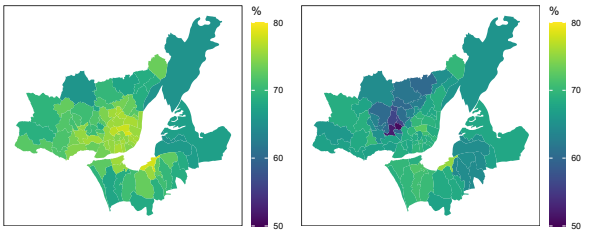
Alcântara				**Sto. António**			
Year	$P_{2.5\%}$	Mean	$P_{97.5\%}$	Year	$P_{2.5\%}$	Mean	$P_{97.5\%}$
2018	8.99	21.64	34.45	2018	12.36	20.13	27.57
2019	10.53	20.23	30.17	2019	10.53	16.67	22.78
2020	7.82	10.94	14.14	2020	8.42	10.27	12.08
Almada				**Oeiras**			
Year	$P_{2.5\%}$	Mean	$P_{97.5\%}$	Year	$P_{2.5\%}$	Mean	$P_{97.5\%}$
2018	14.31	27.27	38.42	2018	5.57	18.12	31.91
2019	19.09	29.39	37.94	2019	3.24	14.01	24.49
2020	8.71	11.8	14.57	2020	5.42	8.99	12.4

The following maps in fig. 8 provide a summarized representation of the reduction of NO_2 concentrations for the 86 parishes in the month of April of 2020, by comparing this interval with April 2018 and April 2019.

5. Discussion

The descriptive statistical analysis for the concentrations of NO_2 measured at the monitoring stations showed a clear reduction in the concentrations of this pollutant in 2020 [24, 17, 27] when compared with previous years. This analysis provided more evident results in March and April, which are likely to be related to the effects of the initial "shock" (and fear) caused by the appearance of a pandemic and the implementation of the severe lockdown measures imposed by the national authorities.

Fitting the linear regression models with the linear land use predictor helped to increase the accuracy



(a) Reduction in 2020 compared to 2018.

(b) Reduction in 2020 compared to 2019.

Figure 8: Reduction by parish, of the 95% confidence intervals of the empirical distributions for the month of April.

of the geostatistical modelling of NO_2, which takes advantage of the fact that there is land cover over the entire spatial domain providing available auxiliary information that is well related to NO_2 emissions [12, 5]. The linear modelling results reinforced the idea that the state of emergency experienced in 2020 caused a marked reduction in anthropogenic emissions and will have positively affected population exposure levels to NO_2. The observation of this phenomenon is not surprising and the results obtained are in line with the existing literature [3].

The results obtained from the adjustments of the semivariograms were considered as positive. During the year 2020, between March and April there is an increase in the range of the semivariogram and a significant reduction in the semivariance, resulting in a smoothing of the concentrations. The increase in the range of the semivariogram indicates that there is correlation with data at greater distances, while the decrease in semivariance indicates less variability in the data [37, 29]. The results of the semivariograms accompanied by the observed decreases in the averages reinforce the impact of confinement effects. Both of these observations occur during the final phase of March and the whole of April, coinciding with the implementation of more restrictive lockdown measures, namely in the form of restrictions on commuting and the implementation of teleworking [24]. These measures resulted in a decrease in the number of vehicles in circulation and the shutdown of several services and industries [22], which are factors that contributed to the increase in local variability. The reduction of the range and semivariance in the following months are also identified as expected results, because their gradual increase represent the lifting of some restriction measures indicating the gradual return to normality.

We then verified that the averages of the 300 SGS resulted in maps similar to those generated by KDE. Similarly, the variance of the 300 SGS generated maps where we observe lower levels of variance in the municipality of Lisbon, due to the

greater proximity between observed data. From the maps of the interquartile range we verified between April and June, a smoothing of the values generated by the simulations and also a higher probability of occurrence of lower values. In areas that lack of data observation, extremely high values of variance were generated for the years of 2018 and 2019, which can be explained by the limitation of the residuals obtained through the linear regression models, which increase the local variability in this type of areas [37, 29]. In the year 2020 there is a tremendous attenuation of the variance, which can be explained due to the increase in spatial continuity and the reduction in observed concentrations.

5.1. Limitations

The use of data imputation methods is a solution found to minimize the impact of missing data in the analysis, and the method chosen may not have been the most appropriate, because it is known that imputation algorithms are sensitive (in terms of performance) to the characteristics of the dataset [26]. PMM was chosen as the imputation algorithm due to its versatility, robustness and simplicity of use [38], having been successfully applied to air pollutant data in [32, 16].

Air quality monitoring stations provide little spatial coverage of the study area, causing difficulties in accurately representing spatial variability. The problem was overcome by using an auxiliary variable with coverage over the entire spatial domain (COS2018). This type of approach is common and often applied by other methodologies such as the *Land use Regression* (LUR models), and supported by existing literature [36, 18, 13, 28, 20]. In order to use land cover as an auxiliary variable, a simplification of COS2018 was used, assessing the presence/absence of each class in the area of influence of the monitoring stations. Although these procedures led to a reduction of available information (due to the simplification) and to an increase of subjectivity (due to the classification) about land cover, the solution found allowed to integrate the impact of the main sources of NO_2 emissions in urban environment (transports, industries or population).

The use of COS2018 for the modelling of NO_2 concentrations was extremely relevant, the presence of high residuals was verified. These suggest that there is the omission of other relevant variables for explainability. A way to fill this gap would be the use of other explanatory variables (temperature, humidity, distance to the nearest road, distance to the nearest park). The inclusion of these variables was not possible due to time constraints.

In areas where there are no air monitoring stations, interpolation of physically impossible values (negative concentrations) was verified. These interpolated values arise when the Kriging estimator is applied to areas where there are ”extreme” values, i.e. larger, in modulus, than those in its neighbourhood, fig. These occurrences make the Kriging estimator unstable, leading to the interpolation of values outside the expected range [29, 9, 37]. On the other hand, the extreme residuals are the result of the weakness of the linear regression models, worsening in the months of May and June (when the models explain less the concentration of NO_2). The solution for the negative concentrations was to replace these occurrences by 0, this being a usual procedure (although not desirable, being painful when there is no large-scale occurrence) used successfully in [29, 4, 2, 34].

6. Conclusion

This study made it possible to carry out a temporal and geostatistical assessment of air quality in the LMA.

In general, through the analysis of the temporal evolution of the concentrations of NO_2, it is possible to state that, during the period analysed, there was a general improvement in the context of this pollutant. The emission of air pollutants has decreased as a result of the application of restrictive measures at national and local level.

COVID-19, and the measures imposed by the Portuguese government to prevent and contain virus transmission, confirmed and corroborated the scientific evidence linking air pollution with human activity. The change in daily life imposed by the pandemic has strongly exposed some of the main sources of air pollution, such as the transport and industry sectors. It became evident, especially in the month of April, that the restriction of socialization and the limitation of economic activities had an almost instantaneous effect in reducing the concentrations of air pollutants. From the general data analysis carried out, there was a significant reduction in the concentrations of NO_2.

NO_2 allowed a clearer interpretation of the impact of COVID-19 on air quality since, given its short lifetime in the atmosphere, its variation of concentrations in the atmosphere resulted directly from the reduction of anthropogenic activities. In the month of April 2020, a reduction of 42% in the values measured at the LMA stations was recorded, compared with the previous year.

The basic problem to be solved with geostatistics is the characterisation of the spatial distribution and the evaluation of uncertainty measures, taking into account the variability of the spatial phenomenon, the quality of the observations, the type of geostatistical model and the degree of knowledge about

the phenomenon. Therefore, in this work we can identify the quality of the available observations as the major limiting factor.
The monitoring network used is, for the most part, located in the urban centres of the LMA, allowing the objective of the work to have been, to a certain extent, fulfilled. It was possible to achieve for 2020, at the parish level and with reasonable levels of uncertainty, an estimate of the concentrations of NO_2 during the four months of lockdown.
The integrated geostatistical approach adopted in this thesis allowed us to estimate, with relative levels of success, the behaviour of NO_2 concentrations in the most populated areas in the centre of the LMA. This analysis may contribute to a discussion of future strategies for the improvement of large-scale monitoring in the LMA. In addition, we consider that one of the strengths of this work was the unique opportunity to analyze a period where there was a suspension of anthropogenic activities on a global scale, which allowed the study of the impact of sectors such as transport and industry, in the concentrations of NO_2 in the LMA.

References

[1] APA. A rede de medição, 2019. REDE QUALAR, visitado 6/10/2021.

[2] M. Beauchamp, L. Malherbe, C. de Fouquet, L. Létinois, and F. Tognet. A polynomial approximation of the traffic contributions for kriging-based interpolation of urban air quality model. *Environmental Modelling & Software*, 105:132–152, jul 2018.

[3] J. D. Berman and K. Ebisu. Changes in U.S. air pollution during the COVID-19 pandemic. *Science of The Total Environment*, 739:139864, oct 2020.

[4] J. J. Carrera-Hernández and S. J. Gaskin. Spatio temporal analysis of daily precipitation and temperature in the Basin of Mexico. *Journal of Hydrology*, 336(3-4):231–249, apr 2007.

[5] S. M. CCDRLVT, Luisa Nogueira. Avaliação da qualidade do ar ambiente na região de lisboa e vale do tejo em 2019, 2020. CCDRLVT, visitado 16/09/2021.

[6] N. Cressie. Fitting variogram models by weighted least squares. *Journal of the International Association for Mathematical Geology 1985 17:5*, 17(5):563–586, jul 1985.

[7] N. Cressie and D. M. Hawkins. Robust estimation of the variogram: I. *Journal of the International Association for Mathematical Geology 1980 12:2*, 12(2):115–125, apr 1980.

[8] I. N. de Estatística. Censos 2021, 2021. Resultados Preliminares, visitado 4/10/2021.

[9] C. V. Deutsch. Correcting for negative weights in ordinary kriging. *Computers & Geosciences*, 22(7):765–773, aug 1996.

[10] F. M. DGT, Mário Caetano. Especificações técnicas da carta de uso e ocupação do solo (cos) de portugal continental para 2018, 2019.

[11] T. R. Foundation. What is r? visitado 15/10/2021.

[12] T. Fu, Joshua S.; Godish. *Air quality*. CRC Press, fourth edition. edition, 2014.

[13] N. L. Gilbert, M. S. Goldberg, B. Beckerman, J. R. Brook, and M. Jerrett. Assessing Spatial Variability of Ambient Nitrogen Dioxide in Montréal, Canada, with a Land-Use Regression Model. *http://dx.doi.org/10.1080/10473289.2005.10464708*, 55(8):1059–1063, 2012.

[14] P. Goovaerts and D. Goovaerts. *Geostatistics for Natural Resources Evaluation*. Applied geostatistics series. Oxford University Press, 1997.

[15] B. Gräler, E. Pebesma, and G. Heuvelink. Spatio-temporal interpolation using gstat. *The R Journal*, 8:204–218, 2016.

[16] S. J. Hadeed, M. K. O'Rourke, J. L. Burgess, R. B. Harris, and R. A. Canales. Imputation methods for addressing missing data in short-term monitoring of air pollutants. *Science of the Total Environment*, 730:139140, aug 2020.

[17] B. M. Hashim, S. K. Al-Naseri, A. Al-Maliki, and N. Al-Ansari. Impact of COVID-19 lockdown on NO2, O3, PM2.5 and PM10 concentrations and assessing air quality changes in Baghdad, Iraq. *Science of the Total Environment*, 754(2):141978, 2021.

[18] G. Hoek, R. Beelen, K. de Hoogh, D. Vienneau, J. Gulliver, P. Fischer, and D. Briggs. A review of land-use regression models to assess spatial variation of outdoor air pollution. *Atmospheric Environment*, 42(33):7561–7578, oct 2008.

[19] E. S. R. Institute. Layers, 2016. ERSI, visitado 16/11/2021.

[20] S. Janssen, G. Dumont, F. Fierens, and C. Mensink. Spatial interpolation of air

pollution measurements using CORINE land cover data. *Atmospheric Environment*, 42(20):4884–4903, jun 2008.

[21] F. Karagulian, C. A. Belis, C. F. C. Dora, A. M. Prüss-Ustün, S. Bonjour, H. Adair-Rohani, and M. Amann. Contributions to cities' ambient particulate matter (PM): A systematic review of local source contributions at global level. *Atmospheric Environment*, 120:475–483, nov 2015.

[22] L. Li, Q. Li, L. Huang, Q. Wang, A. Zhu, J. Xu, Z. Liu, H. Li, L. Shi, R. Li, M. Azari, Y. Wang, X. Zhang, Z. Liu, Y. Zhu, K. Zhang, S. Xue, M. C. G. Ooi, D. Zhang, and A. Chan. Air quality changes during the COVID-19 lockdown over the Yangtze River Delta Region: An insight into the impact of human activity pattern changes on air pollution variation. *Science of The Total Environment*, 732:139282, aug 2020.

[23] R. Menezes, P. Garcia-Soidán, and M. Febrero-Bande. A comparison of approaches for valid variogram achievement. 2003.

[24] L. Menut, B. Bessagnet, G. Siour, S. Mailler, R. Pennel, and A. Cholakian. Science of the Total Environment Impact of lockdown measures to combat Covid-19 on air quality over western Europe. *Science of the Total Environment*, 741:140426, 2020.

[25] C. J. Morgan. Theoretical and practical aspects of variography: in particular, estimation and modelling of semi-variograms over areas of limited and clustered or widely spaced data in a two-dimensional South African gold mining context. jan 2012.

[26] T. P. Morris, I. R. White, and P. Royston. Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology 2014 14:1*, 14(1):1–13, jun 2014.

[27] L. Y. K. Nakada and R. C. Urban. COVID-19 pandemic: Impacts on the air quality during the partial lockdown in São Paulo state, Brazil. *Science of The Total Environment*, 730:139087, aug 2020.

[28] E. V. Novotny, M. J. Bechle, D. B. Millet, and J. D. Marshall. National satellite-based land-use regression: NO2 in the United States. *Environmental Science and Technology*, 45(10):4407–4414, may 2011.

[29] T. Pei, C. Z. Qin, A. X. Zhu, L. Yang, M. Luo, B. Li, and C. Zhou. Mapping soil organic matter using the topographic wetness index: A comparative study based on different flow-direction algorithms and kriging methods. *Ecological Indicators*, 10(3):610–619, may 2010.

[30] M. Pereira, A. Soares, J. Almeida, and C. Branquinho. Geostatistical models for air pollution. 01 2000.

[31] QGIS. About qgis. visitado 15/10/2021.

[32] M. E. Quinteros, S. Lu, C. Blazquez, J. P. Cárdenas-R, X. Ossa, J. M. Delgado-Saborit, R. M. Harrison, and P. Ruiz-Rudolph. Use of data imputation tools to reconstruct incomplete air quality datasets: A case-study in Temuco, Chile. *Atmospheric Environment*, 200:40–49, mar 2019.

[33] M. C. Ribeiro, P. Pinho, C. Branquinho, E. Llop, and M. J. Pereira. Geostatistical uncertainty of assessing air quality using high-spatial-resolution lichen data: A health study in the urban area of Sines, Portugal. *Science of the Total Environment*, 562:740–750, aug 2016.

[34] L. D. Rizo-Decelis, E. Pardo-Igúzquiza, and B. Andreo. Spatial prediction of water quality variables along a main river channel, in presence of pollution hotspots. *Science of The Total Environment*, 605-606:276–290, dec 2017.

[35] RStudio. About rstudio, 2021. visitado 15/10/2021.

[36] P. H. Ryan and G. K. Lemasters. A Review of Land-use Regression Models for Characterizing Intraurban Air Pollution Exposure. *https://doi.org/10.1080/08958370701495998*, 19(SUPPL. 1):127–133, 2008.

[37] A. Soares. *Geoestatística para as ciências da terra e do ambiente*, volume 3rd. IST Press, 2014.

[38] S. van Buuren. *Flexible Imputation of Missing Data. Second Edition*. Chapman & Hall/CRC Press, Boca Raton, FL, 2018.

[39] S. van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3):1–67, 2011.

[40] WHO. Ambient (outdoor) air pollution, 2021. WHO, visitado 4/11/2021.