Biostatistics 2021/2022
Instituto Superior Técnico

Project

Impact of air pollution on hospitalizations from respiratory disease

Professor Maria Rosário Oliveira

Augusto Marques (89789), Catarina Rodrigues (92434), Filipa Costa (92626),
Patrícia Roxo (92643), Maximilian Buxadé (101475)

ABSTRACT — *The present work aims to find and study the relationship between air pollution levels and the number of hospitalizations from respiratory disease, in Portugal. To fullfill this purpose, data from the last 5 years regarding levels of 6 air pollutants was retrieved, as well as the number of hospitalizations caused by respiratory disease. Data was pre-processed and an exploratory analysis was conducted. A strong dependence of hospitalizations over the age range of patients, as a well as a temporal dependency was discovered. A similar pattern appeared in the pollutants behavior over time. Nextly, an one-way ANOVA test (Kruskal-Wallis Test) was performed between levels of CO and NOx and hospitalizations for different age ranges/genders, both with and without missing values imputation. Lower p-values were reached with filled NAs, namely for higher ages. Finally, timeless regression models were created to access how much the air pollution affects the number of hospitalizations. To do this, variables transformation occured to maintain the linearity assumption, followed by feature selection via correlation and models construction. The best model obtained was the one done with stepwise forward and backward regression, after subjected to a box-cox transformation. The choosen model suggests that the air pollution increases the hospitalizations, namely for Madeira region, regardless of the age. Finally, predictions of hospitalizations were made using the model choosen, with relative errors below 20% for all age groups, which turned out to be surprising, since these were only computed using air pollution data.*

Keywords: Regression, Kruskal-Wallis Test, Hospitalizations, Air pollution

1 Introduction

Air pollution affects one's health, namely the respiratory and cardiovascular systems, in both acute and chronic ways, giving rise to diverse and long-term health problems [1]. These are caused by air pollutants such as nitrogen oxides (NO), ozone (O_3), carbon monoxide (CO), and particulate matter (PM), which in turn cause respiratory symptoms by damaging different parts of the respiratory tract through a variety of mechanisms. Although air pollution has an overall impact on population, some people are more vulnerable than others to its effects. Children living in polluted areas are more likely to suffer from coughs, wheezing, asthma and impaired lung function, as well as the elderly who are strongly affected due to reduced lung function that occurs with ageing.

Particulate matter (PM) is used as an indicator of air pollution and is typically classified according to particle size. $PM_{10\text{-}2.5}$ refers to coarse particulate matter with a diameter between 10 μm and 2.5 μm. Accordingly, $PM_{2.5}$ refers to particles with a diameter of less than 2.5 μm, which have the ability of penetrating deep into the lung, reaching the alveoli. Particulate matter that enters the lung epithelium can cause lung inflammation, intensifying pre-existing lung conditions and exacerbating asthma, as well as chronic obstructive pulmonary disease (COPD) [2]. Hence, reducing air pollution is essential, as it poses a serious threat to both health and the environment.

With this in mind, the goal of this project is to study how air pollution influences the number of hospitalizations, caused by respiratory disease, according to the age and gender of the patients, in the last 5 years. The role of the COVID-19 pandemic in this assessment will be studied as well.

Firstly, in section 2, data was pre-processed and submitted to a preliminary analysis. In section 3, Kruskal-Wallis test and regression models were set into place, in order to better estimate the correlation between pollution and hospitalizations. The first one was used to compare the distribution of hospitalizations with different levels of air pollution before COVID-19. The second one uses linear regression in order to find the general relationship between hospitalizations and pollutants, before and during the COVID-19 season. Finally, the conclusions of this study are given in section 4.

2 Materials and Methods

2.1 Data Preparation

To start the assessment, a search for data regarding the two parameters in hand was made (levels of air pollutants and hospitalizations resulting from respiratory disease). Note that only Portuguese related data was retrieved, implying that the work in hand does not establish an assessment for other countries, aside from Portugal. Thus, data regarding hospitalizations was taken from the Portuguese national health system official website, and Portuguese air pollution data was taken from the European Environment Agency (EEA) [3][4]. Both datasets share two common variables, through which they would later be merged: `Region` (Lisboa e Vale do Tejo, Centro, Norte, Algarve, Madeira) and `Period` (from January 1^{st} 2017 until December 31^{th} 2020).

For the air pollution dataset, the levels of air pollution were all converted into $\mu g/m^3$. Then, only 6 pollutants were considered, namely $\mathtt{PM_{10}}$, $\mathtt{O_3}$, $\mathtt{NO_x}$ as $\mathtt{NO_2}$, `NO`, $\mathtt{PM_{2.5}}$ and `CO`. Next, given that data was provided with an hourly measurement, for a given region, month and pollutant, the monthly mean and maximum values were computed. This was done to all 5 regions, 60 months (5 years) and 6 pollutants, generating two new variables per pollutant. Note that the data for a given region was provided separately according to the city in which it was measured. To ensure that all cities contributed equally to the total pollution measures of a given region, the values were added iteratively until there was no data left for that region, and divided by the total number of records read. Missing values imputation was performed by filling them with the median value of air pollution level, of the correspondent region.

On the other hand, for the hospitalizations dataset the preparation was simpler. The variables considered, besides `Period` and `Region`, were `Age Range`, `Gender` and `Hospitalizations`. Next, Açores and Alentejo region data were discarded, since the air pollution dataset did not contain data for these two regions. Then, `Age Range` variable was mapped into integers, with age ranges $[0,1[, [1,5[, [5,15[, [15,25[$ merged into one, since the statistical measures for these age groups were similar (Figure 1). Additionally, `Hospitalizations` variable was transformed in order to get its values per 100 000 inhabitants, as it facilitates their reading and posterior analysis. Note that the original dataset had `Hospitalizations` records per health care facility (hospital, clinic, etc.), therefore these were joined according to the same region. This dataset did not contain missing values.

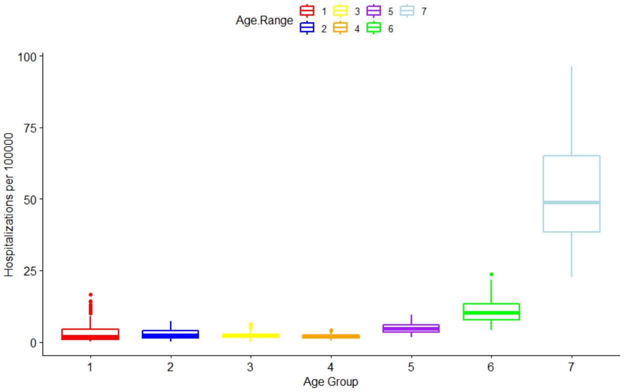

Fig. 1: Distribution of hospitalizations per age range variables, in the original hospitalizations dataset, retrieved from SNS website.

With the two datasets cleaned, both were merged into one, by region and period (month). Note that for different age ranges and genders, the pollution levels remain the same. Moreover, different datasets were computed from these merge. Six datasets were created - with and without gender; with and without COVID-19 pandemic period; and, finally, with and without missing values imputation. The merged dataset description (from which the previous 6 datasets were created) can be seen in detail in the appendix of this paper.

2.2 Preliminary Analysis

2.2.1 Outliers and Distributions

The goal of our project is to study the impact of air pollutant levels in hospitalizations, so, in order to have a better understanding of the dataset, an outlier detection and correspondent distribution were computed through boxplots.

As it can be seen in both Figures 2 and 3 there are some values out of the interquartile range, however, since one is dealing with medical observations, any value can be important for the study, and as all the values seem to be not too far from each other, it was decided to not remove any value. Note that in Figure 2 there seems to be a great number of outliers. In fact, a black line of outliers can be identified, which indicate that those records do not, in reality, represent outliers, since they are not clearly apart from the variable's distribution.

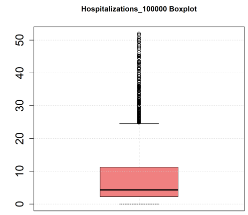

Fig. 2: Boxplots of `Hospitalizations` per 100.000 residents.

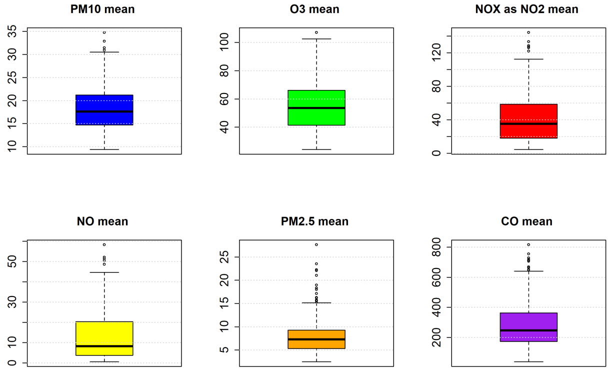

Fig. 3: Boxplots of mean levels of air pollutants.

Now, observing the boxplots in figure 4 for the maximum levels of air pollutants it is clear that there is an outlier in PM_{10} with value 1000. This value differs drastically from all others, as it deviates from normality and can cause anomalies in the results obtained by the linear regression models. Consequently, it was decided to remove it.

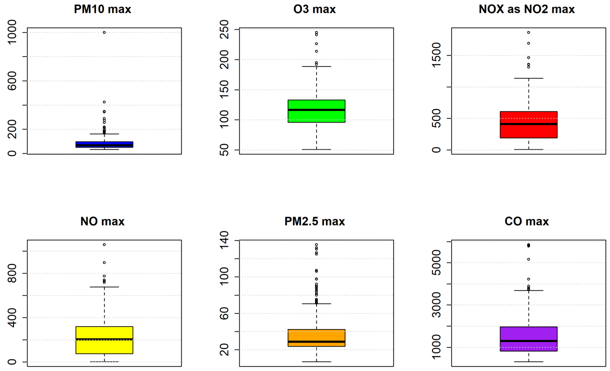

Fig. 4: Boxplots of air pollutants max

2.2.2 Temporal Dependency

For a brief analysis of the dataset, Figure 5 shows the mean levels of the different pollutants over time, in each region (LVT, Centro, Norte, Algarve, Madeira).

Analysing Figure 5 it is clear that the mean values of NO_2, NO and CO are higher in Norte region, which makes sense since the northern region is a large anthropogenic region, where road traffic and energy production are predominant [5]. Nevertheless, the mean levels of O_3 are high, which makes sense because levels of O_3 are higher in places without NO pollutants since they react with each other to form NO_2. This happens in rural areas, with less pollution, like Madeira and Algarve [6].

After asserting the general behaviour of each pollutant, the next step was to examine hospitalizations over time. Moreover, it made sense to analyze these categorized by age. It is well known that elders have a much more debilitated immune system and therefore are far more at risk of disease in comparison to other age groups. Conversely, children and young adults are known to be the healthiest age group, as it is too soon for the majority of (possibly preexisting) health conditions to manifest or develop. The data supported this heterogeneity, as it can be seen in Figure 6.

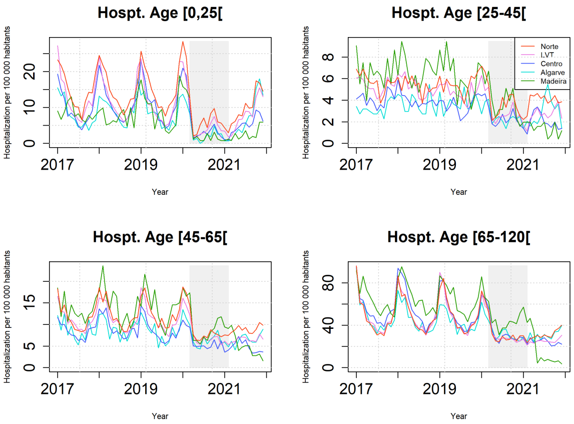

Fig. 6: Hospitalizations per 100 000 residents for different age groups. The gray rectangle represents the time of the COVID-19 pandemic.

For hospitalizations, the oldest age group [65,120[stands out among the rest, being approximately 15 times as high as the hospitalizations for age group [45,65[, which in turn is about 2.5 times higher than those for age group [25,45[but curiously about the same as the hospitalizations for age group [0,25[.

Initially the analysis was done under the assumption that male and female individuals might have behaved differently. However little disparity was found, thus the shown graphs do not take gender into account.

A clear drop in hospitalizations can be seen in the gray area among all age groups - the period corresponding to start of the COVID-19 pandemic. Despite having numerous consequences for the respiratory system, COVID-19 is not labeled as a respiratory disease and so the societal context experienced at the time can explain this decrease in hospitalizations [7]. Note that, according to the Portuguese National Health System, every hospitalization and death caused by COVID-19 has a special diagnosis code, even though these patients actually die or get hospitalized due to respiratory conditions such as COPD, pneumonia or breathing problems. So, despite being related to respiratory illnesses, data regarding these patients' clinical history is marked as a COVID-19 diagnostic, since it was the precursor of those same illnesses [8]. This explains in detail why the drop in hospitalizations across the whole Portuguese land occurs. Furthermore, people were advised, and at times obliged, to stay and quarantine at home, avoiding all kinds of travels. This included attending medical appointments (which were often postponed indefinitely) and visiting the hospital, given

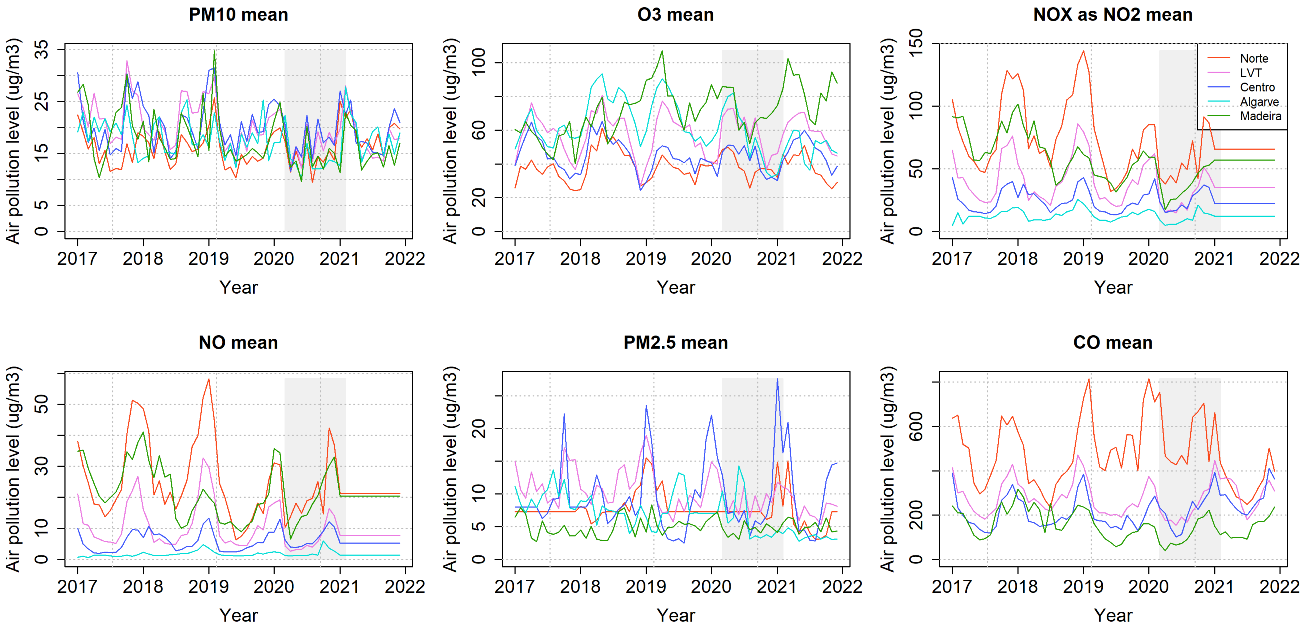

Fig. 5: Mean levels of the different air pollutants over time in each region of Portugal. The gray rectangle represents the time of the COVID-19 pandemic.

that the national healthcare system was overloaded handling COVID-19 patients.

This same context also explains the consequent rise in hospitalizations right after the gray region ends. With social restrictions lifted and attenuated, people could finally be admitted for care, and during the pandemic, when they were indeed admitted, they couldn’t stay for long.

Aside from the erratic evolution of the Madeira islands, the relative number of hospitalizations per 100.000 residents generally exhibited the same behaviour and even decreased for the most at risk age group, [65,120[. This somewhat matches the pollutants’ measurements. For pollutants and hospitalizations and alike, they exhibit the same close to cyclic behaviour, with a visible but not necessarily always high decrease during the COVID-19 pandemic period.

2.2.3 Correlation Analysis

Under the reasonable assumption that deaths and hospitalizations are linked to one another and that usually a death due to respiratory disease is preceded by a hospitalization, these observations motivate directly assessing the correlations between pollutants and hospitalizations. This analysis was also carried out per age group, because despite close in behaviour, they don’t always follow the same trend, as previously mentioned. In Figure 7 there is a summary of the results found.

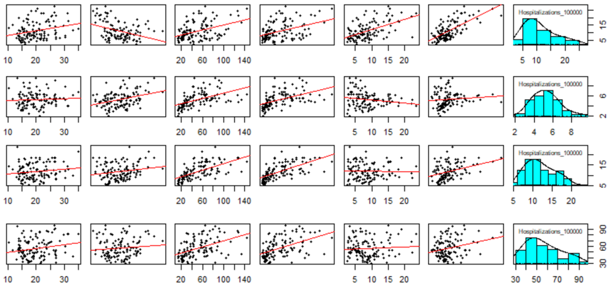

Fig. 7: From left-to-right: correlation between $\texttt{PM}_{10}$, $\texttt{O}_3$, $\texttt{NO}_x$ as $\texttt{NO}_2$, $\texttt{NO}$, $\texttt{PM}_{2.5}$ and $\texttt{CO}$ mean and the number of hospitalizations per 100 000 residents per age group. From top-to-bottom: correlations for age group [0,25[and respective relative hospitalizations density, correlations for age group [25,45[and respective relative hospitalizations density, correlations for age group [45,65[and respective relative hospitalizations density and correlations for age group [65,120[and respective relative hospitalizations density.

$\texttt{PM}_{10}$ and $\texttt{O}_3$ mean- the first 2 columns - do not appear to be at all correlated with hospitalizations due to respiratory disease in age 6 and 7. $\texttt{NO}_x$ as $\texttt{NO}_2$ and $\texttt{NO}$ mean - the next 2 columns - already exhibit a relatively pronounced positive correlation with the number of hospitalizations on all age groups. Do note that these 2 gases behave very similarly as they belong to the same family of pollutants: $\texttt{NO}_x$ encompasses precisely $\texttt{NO}_2$ and $\texttt{NO}$. Moreover, this positive correlation makes sense. $\texttt{NO}_x$ are typically considered to be the most relevant gases for air pollution.

At last, $\texttt{PM}_{2.5}$ and $\texttt{CO}$ mean surprisingly seem to only affect the youngest age group. In the remaining age groups, the relation between $\texttt{PM}_{2.5}$ mean and hospitalizations is so thin that its best fitting curve’s slope even changes signs from age groups [25,45[and [45,65[to age group [65,120[, thus proving that there is no correlation between them. In the case of `CO` mean, it can be argued that the apparent positive correlation for the older age groups came from the data being more densely gathered on the left of the graph with its outliers on the right 'pulling' the slope upwards. The scatter plot overall does not support a positive linear relation between the `CO` measurements and hospitalizations.

As a rule of thumb, variables are only considered to be correlated to one another if their correlation exceeds the threshold of 0.80. Since NOX as NO2 mean and NO mean fulfill this threshold (in Figure 8), one of them will be discarded from the models, keeping the one that has more correlation with the Hospitalizations variable. For the remaining age groups, please consult the Appendix.

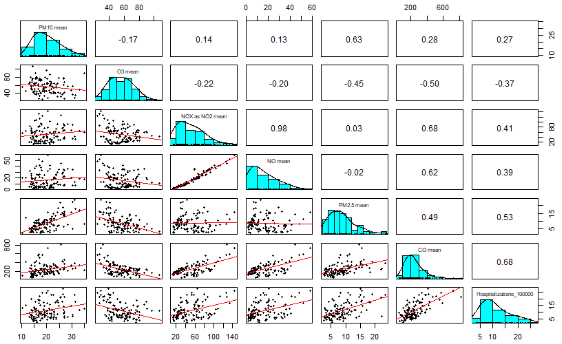

Fig. 8: Correlation between each pollutant and relative number of hospitalizations per 100 000 residents for age group [0,25[.

Moreover, it jumps out that some pollutants seem to exhibit some kind of correlation with others. This is to be expected, as we know they chemically interact in the atmosphere and thus the presence of one might affect the presence of another (positively or negatively). Although interesting, this chemical interaction is outside the scope of this work.

2.2.4 Missing Values Analysis

As previously said in the Data Preparation section, missing values were imputed regarding the air pollutant variables. Nonetheless, a closer analysis to the records with missing values inside these variables was conducted, in order to infer over possible patterns amongst them. Note that the conclusions taken onwards are valid for both mean and maximum levels of air pollution.

For starters, PM_{10} and CO levels are missing for Algarve region. CO is not measured at all for Algarve region, having the second greatest number of missing values (480). Note that, in function of this, the plots over time regarding the CO air pollution levels do not present data for Algarve region. PM_{10} missing values only appear between January and February 2020, being a minority (16). Regarding NO_x as NO_2 and `NO` levels, these have the same number of missing values (496), for the same records, which is expected since `NO` is inside the NO_x family compounds. Interestingly, the missing values for these two pollutants happen only in January, April, May and June of 2017 for Algarve, and over all months of 2021, across all regions. As mentioned before, it is expected to have low values of NO_x family compounds for more rural areas, thus explaining the tendency for a greater number of missing values of these compounds in Algarve (not so rural when compared to Centro, LVT or even Norte). Finally, ozone (O_3) did not contain any missing values, and $PM_{2.5}$ levels had no clear pattern, along region or period, though the majority of missing values (424) seems to be present namely in Centro and Norte region.

3 Results and Discussion

3.1 Kruskal-Wallis Test

The goal of this section is to observe if the distribution (or central tendency) of hospitalizations is the same for different levels of pollution. A division into three levels of pollution was performed, where the three levels were chosen in a way that the amount of data for each one is the same. This was done by separating them into the 33% and 66% quartiles. Furthermore, as the correlation analysis showed, the most correlated pollutants for hospitalizations were `CO` and NO_x, so the test was done only to these two pollutants.

Unfortunately, the data does not meet the criteria for a parametric ANOVA test, specifically the normality assumption and the homoscedasticity one. As a result of this, the non-parametric Kruskal-Wallis Test was performed. The results in this section have to be interpreted cautiously, because independence between the three pollution groups has to be assumed for the test. However, this is not the case, since the Hospitalizations seem to have a time-dependent autocorrelation. This weakens the meaningfulness of the results. The test was performed for the Age groups 1, 5 and 6.

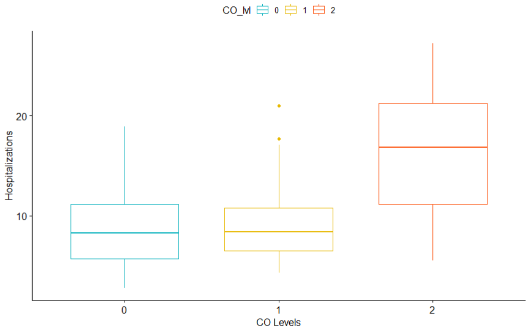

Fig. 9: Hospitalizations per 100.000 residents of age group 1 for different levels of `CO`.

Looking at the boxplots shown in Figures 9 and 10, one might suspect that the initial supposition that hospitalizations increase when pollution increases is correct. Especially for `CO`, a steep increase in hospitalizations for the highest pollution level can be observed. Indeed this relation is also seen in the other two age groups. After performing the test, the resulting p-values, as seen in Table 1, indicate that the null hypothesis, H_0, can be discarded, i.e. the three populations do not seem to follow the same distribution, at the 5% significance level. Given the previously discussed boxplots, it seems that the median increases for increased pollution. Even taking into consideration that the results given may not be as meaningful, because the assumptions are not completely met, there is still indication that the initial supposition is true. Furthermore, a lower p-value for the higher age groups is observed. This can be interpreted as a stronger difference in the central tendency for these groups. Therefore one might suspect a stronger relationship between hospitalizations and air pollution for increased ages. The test did not have significantly different results for male or female population.

Performing the test on the data which had the missing values filled as discussed previously, results in lower p-values. This is expected, since considering hospitalizations for periods without pollution data makes it possible to analyze more data, which stabilizes the distribution of hospitalizations for each pollution level and therefore reinforces the observed difference between them.

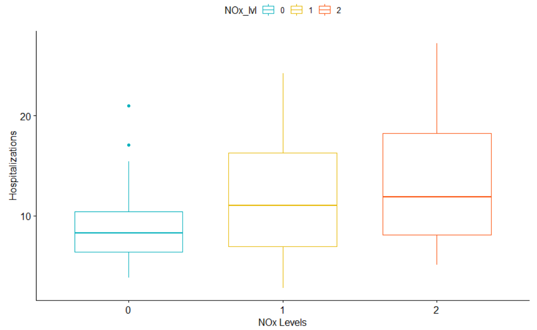

Fig. 10: Hospitalizations per 100.000 residents of age group 1 for different levels of $\mathtt{NO}_x$ measurements.

3.2 Regression

Regression models were used to determine how much the air pollution affects the number of hospitalizations per 100 000 residents per age group. Age group 7 was removed because the data had an extremely high variance, as can be observed in the boxplot of figure 1. This might be explained by the larger width of this age group as well as the many regional differences among the elderly population. Therefore it was decided to focus on the regression for the other groups.

3.2.1 Model Selection

As a first approach, a full model including the pollution variables, months, and regions was defined. Because hospitalizations are significantly reliant on time as discussed in 2.2.2, this model only gave importance to months and regions (4). As a consequence of this, it was decided to omit the time dependency and construct regression models without the month variable, given that the main goal is to identify the influence of air pollution alone.

After the initial model was built, an analysis to inspect the relation between the hospitalizations and other variables was done. By means of this analysis, it was possible to observe whether some transformation of the predictor variables would improve the models. Some pollutants appeared to be better explained by a non-linear relationship, such as a logarithmic function. So the respective data was transformed by age to maintain the linearity assumption between the dependent and independent variables. The variables' transformation can be seen in the Appendix (Table 11). Furthermore, the variables that were highly correlated, i.e. more than 80%, between each other were removed, leaving just the one with greater correlation with the hospitalizations to minimize the risk of multicolinearity. Variables with a correlation of less than 20% with the hospitalizations were also removed because they did not appear to explain much of the hospitalization data. The final variables are given by the previous correlation analysis. Finally, the data was divided into a training and a testing set, where the last month before the pandemic, namely February of 2020, was set to be the test data.

Recall that the first dataset for which models were constructed had the data with missing pollution data removed, as well as genders joined. The following models were considered:

1. Model with a subset of variables chosen in accordance with the correlation analysis;
2. Model with a subset of variables considering the correlation analysis with partly transformed data;
3. Model obtained by stepwise forward and backward regression on each considered age group, denoted by $(m1_1, m1_5, m1_6)$ respectively;
4. Model obtained by stepwise forward and backward regression on all variables with partly transformed data;
5. Model obtained by Best Subsets Technique.

The models in 3 and 5 were the approaches that gave better adjusted R^2 values. Both these models were similar to each other, therefore model 3 was arbitrarily chosen to run further analysis on.

As a test, it was decided to fit the regression model to the dataset considering genders as well. As it was to be expected, the adjusted R^2 value was similar, since the gender did not present meaningful differences as observed in the preliminary analysis. It is noticeable that the adjusted R^2 value for the female population is slightly lower than for males. One could argue that the air pollution can better explain the hospitalizations of the male population, but considering that the difference is small no conclusions are established here.

Tab. 1: Kruskal-Wallis Test p-value results.

Missing Values Imputation	Gender	Pollutant	Age Range 1	Age Range 5	Age Range 6
No	M/F	CO	1.18e-08	0.0494	0.001
		NOx as NO2	0.005	3.39e-10	3.38e-11
	M	CO	1.17e-08	0.00189	0.0002
		NOx as NO2	0.0327	8.65e-10	7.56e-11
	F	CO	6.69e-08	0.0207	0.00989
		NOx as NO2	0.0000627	2.02e-07	7.86e-09
Yes	M/F	CO	7.76e-12	0.0009	0.0001
		NOx as NO2	1.09e-08	<2.20e-16	<2.20e-16

Because there was essentially no difference on the performance of the model considering gender, they were once again discarded and models for the dataset with imputed missing values were constructed. These models are denoted by $m2_1, m2_5$ and $m2_6$, and were once again obtained by the same approach as in 3. The adjusted R^2 results for models $m2_1, m2_5$ and $m2_6$ were slightly better than those for $m1_1, m1_5$ and $m1_6$. Diagnosis analysis was performed on all these models, $m1_1, m1_5, m1_6, m2_1, m2_5$ and $m2_6$. Finally, Table 2 shows a resume of the criteria of these 6 models.

3.2.2 Diagnosis

In this subsection the assumptions regarding the residuals of the regression are analyzed. The following tests were performed to achieve this:

- To test the **homoscedasticity of residuals variance**, the Breusch-Pagan Test was used, where the null hypothesis states that the residuals have constant variance.
- To test the **normality of the residuals** assumption, the Shapiro-Wilk Test was used, where the null hypothesis states that the residuals follow a normal distribution.
- To test the **autocorrelation between the residuals**, the Durbin-Watson Test was used, where the null hypothesis states that the autocorrelation between the residuals is zero.
- To check the lack of **multicolinearity** of the predicting variables, the VIF was calculated and interpreted.

Moreover, the results of the performed tests are displayed in the Table 2 for the corresponding models. Since none of the models pass the Breusch-Pagan Test, it is concluded that the residuals have non-constant variance. To overcome this, a box-cox transformation in the dependent variable was performed, such that the variance of the residuals would be constant. This approach resolved the issue for the models $m1_1, m1_5$ and $m1_6$. On the other hand, the remaining models still had a varying variance. Additionally, the normality of the residuals was verified for all the models, except for the models of age group 5, with the box-cox transformation. Regarding the autocorrelation of the residuals, the Durbin-Watson test returns values between 0 and 4, where 2 means no autocorrelation, values above 2 translate into positive autocorrelation, and below 2, negative. Thus, we conclude that the residuals have a weak negative autocorrelation. Finally, almost no multicolinearity was found, since the VIF values are all below 5 except for $m2_1$. That model has multicolinearity for $\mathtt{NO_x}$ as $\mathtt{NO_2}$ mean and NO mean.

In order to select the best model, a few validation measures were considered for the constructed models. The results can be seen in Table 2. From Table 2, it can be concluded that the models who behave the best are the ones with the box-cox transformation. These are the models that will be used to make the predictions.

From these models and their coefficients given in 5, 6 and 7, some observations can be made regarding age. All regions, except Madeira, have no or only low significance, so they

Tab. 2: Results of Breusch-Pagan Test, Shapiro-wilk Test and Durbin-Watson Test for nine models considered to verify different assumptions; Results of performance measures R^2_{adj}, AIC, BIC and PRESS

Models	**Assumptions**			**Criteria**			
	Breusch-Pagan	**Shapiro-Wilk**	**Durbin-Watson**	R^2_{adj}	**AIC**	**BIC**	**PRESS**
	p-value	**p-value**	**Statistic**				
m1_1	0.0479	0.5501	1.649294	0.6669	600.4024	618.4274	1348.416
m1_5	0.0002	0.2968	1.74426	0.5151	341.5253	361.9598	131.2312
m1_6	0.0072	0.1833	1.242642	0.5017	552.2741	570.2991	849.3292
m2_1	0.0386	0.422	1.575872	**0.6838**	776.4013	807.3405	1681.2140
m2_5	1.438e-06	0.1692	1.681207	**0.537**	417.1705	442.7705	145.4846
m2_6	4.019e-05	0.0601	1.449121	**0.5934**	671.2531	691.3804	815.8275
m1_1.box.cox	0.0169	0.2918	1.692048	0.5935	**-51.10679**	**-28.3079**	**4.0783**
m1_5.box.cox	0.132	7.706e-05	1.683823	0.5168	**-125.4552**	**-109.8833**	**2.1038**
m1_6.box.cox	0.5266	0.7667	1.277544	0.5500	**-156.2279**	**-138.2028**	**1.6071**

do not seem to have an influence on hospitalizations. The geographical and political circumstances make Madeira the most distinct compared to the other regions, so it is expected that this can be seen in the hospitalizations, with it being the most significant of regions. Further, positive coefficients for all pollutants, except O_3 max, were found. This supports the hypothesis that an increase in air pollution, makes the Hospitalizations increase. O_3 exhibits a negative correlation towards the other pollutants, and higher levels of it are mostly found in rural areas, therefore even the negative impact of it is to be expected. Lastly, `CO mean` and `NO`$_x$ `mean` had the overall highest significance. This could mean that they have the most effect on the health of the population.

3.2.3 Prediction

The models chosen were used to perform predictions on the last month before the pandemic (February 2020), and will therefore be used to predict the number of Hospitalizations during the COVID period, without the pandemic effect. A detailed view of the results can be found in the Appendix. The best results of this February 2020 prediction were obtained for the age group 6, where the worst relative error was 6.39% for Madeira and the best was for the Centro Region with 0.22%, just 0.06 more hospitalizations per 100 000 residents than the true value. The model performed the worst for the age group 1, with a relative error of 19.11% for the Madeira region.
Finally the models were used to predict the hospitalizations of 10 months of the pandemic. Since age group 6 performed the best on the test data, specially for the Centro Region, only the results for this age and region will be exposed in the report. The remaining ones can be seen by running the code provided.

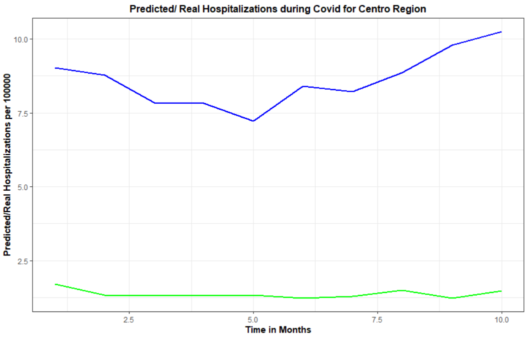

Fig. 11: Hospitalizations per 100 000 residents of age group 6, for the Centro Region, from March 2020 until December 2020.

Figure 11 corroborates that the real hospitalizations due to respiratory diseases during the COVID-19 pandemic (in green) decreased a lot, as measured with respect to the baseline predicted by the fitted regression model (in blue).

4 Conclusions

The aim of this report was to investigate the effect of air pollution in respiratory diseases, at a national level. Unfortunately, the official data available posed many challenges and so the results given by all statistical tests applied are to be taken with a grain of salt. Regardless, it was still very much possible to infer a positive link between two singled out pollutants, NO_2 and CO, and respiratory diseases.

To best model the effect of these pollutants, multiple distinct models were proposed and iteratively chosen for further improvement and testing. The performance of these models was then measured by running it against real data.

Moreover, these models were also taken as a baseline to study the sudden drop in hospitalizations during the COVID-19 pandemic seen in the data. The predicted values from the models show that during the pandemic, the hospitalizations due to respiratory disease decreased significantly, which is a positive silver lining to the situation.

Other interesting links found include suggestions that men's health may be more affected by pollution than women's; infants through young adults are susceptible to more air pollutants than other age groups; and that the older half of the working population, aged [45,65[, being the best modelled group, are thus presumably less exposed to other random factors that could influence their respiratory system.

Future works could and should included parametric tests, provided that the available data improves, as well as a more in-depth look at the behaviour of outliers. A possible explanation for all these outliers, may have to do with climate change, which would also be interesting to study, although it would require decades and decades of data, preferably, not constraint to Portuguese territory. Further, a time dependency analysis with temporal series remains to be done.

Moreover, ideally, as the pandemic becomes less and less relevant in the current days, the same tests run in this report would satisfy the homoscedasticity of residuals hypothesis, thus allowing for more reliable conclusions, without added work.

References

[1] S. Rajagopalan, S. Al-Kindi, and R. Brook, "Air pollution and cardiovascular disease.," *Journal of the American College of Cardiology*, vol. 72, no. 17, pp. 2054–2070, 2018.

[2] T. C. B. Institute. "Risk factors and symptoms of air pollution on respiratory health." (accessed: 08.06.2022). (2019), [Online]. Available: https://www.thecleanbreathinginstitute.com/evidence/risk-factors/.

[3] S. N. de Saúde. "Morbilidade e mortalidade hospitalar - transparência." (accessed: 25.05.2022). (2021), [Online]. Available: https://transparencia.sns.gov.pt/explore/dataset/morbilidade-e-mortalidade-hospitalar/table/?sort=periodo.

[4] E. E. Agency. "Download of air quality data." (accessed: 25.05.2022). (2019), [Online]. Available: https://discomap.eea.europa.eu/map/fme/AirQualityExport.htm.

[5] F. N. Francisco Ferreira. "Rumo à descarbonização e à poluição zero na região norte." (accessed: 10.06.2022). (2021), [Online]. Available: https://www.ccdr-n.pt/storage/app/media/2021/CCDRN_JUN2021_final.pdf.

[6] B. I. E. Agency. "Why are ozone concentrations higher in rural areas than in cities?" (accessed: 10.06.2022). (), [Online]. Available: https://www.irceline.be/en/documentation/faq/why-are-ozone-concentrations-higher-in-rural-areas-than-in-cities.

[7] J. Sarma. "Is covid-19 really a respiratory disease? no, it is a vascular illness, claims new study." (accessed: 10.06.2022). (2021), [Online]. Available: https://www.thehealthsite.com/news/is-covid-19-really-a-respiratory-disease-no-it-is-a-vascular-illness-claims-new-study-834748/.

[8] A. C. do Sistema de Saúde LP. "Circular normativa - codificação clinica de doentes internados com diagnóstico de covid-19." (accessed: 10.06.2022). (2020), [Online]. Available: https://www.acss.min-saude.pt/wp-content/uploads/2020/05/Circular-Normativa-5_2020_Codif-Covid-19.pdf.

Appendix

Dataset Description

Tab. 3: Variables belonging to the used dataset, and their description.

Variable	Description
Period	Datetime variable in the format %Y-%m-%d
Region	Region of Portugal[1]
Age Range	Discretization of age ranges by integers[2]
Gender	Patients gender (M or F)
PM10 mean	Mean level of particles with diameter $< 10\ \mu m$
PM10 max	Maximum level of particles with diameter $< 10\ \mu m$
O3 mean	Mean level of ozone
O3 max	Maximum level of ozone
NOX as NO2 mean	Mean level of nitrogen dioxide
NOX as NO2 max	Maximum level of nitrogen dioxide
NO mean	Mean level of nitrogen oxide
NO max	Maximum level of nitrogen oxide
PM2.5 mean	Mean level of particles with diameter $< 2.5\ \mu m$
PM2.5 max	Maximum level of particles with diameter $< 2.5\ \mu m$
CO mean	Mean level of carbon monoxide
CO max	Maximum level of carbon monoxide
Hospitalizations_100000	Hospitalizations per 100 000 residents

[1] Norte, Centro, LVT (Lisboa e Vale do Tejo), Algarve e Madeira. Does not include Açores and Alentejo regions.

[2] An age range of 1 represents ages between 0 and 25 years old. An age range of 5 represents ages between 25 and 45 years old, 6 between 45 and 65 years old, and 7 between 65 and 120 years old.

Correlation Plots

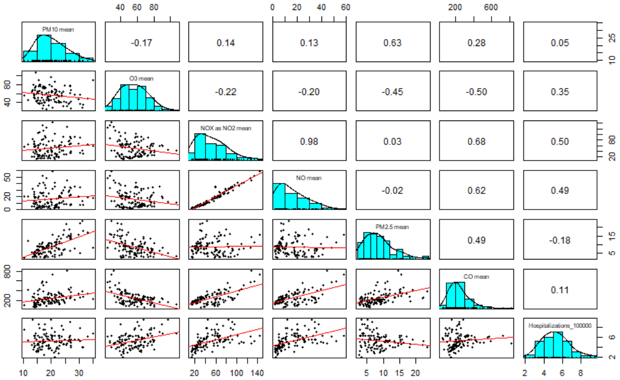

Fig. 12: Correlation between each pollutant and relative number of hospitalizations per 100 000 residents for age group [25,45[.

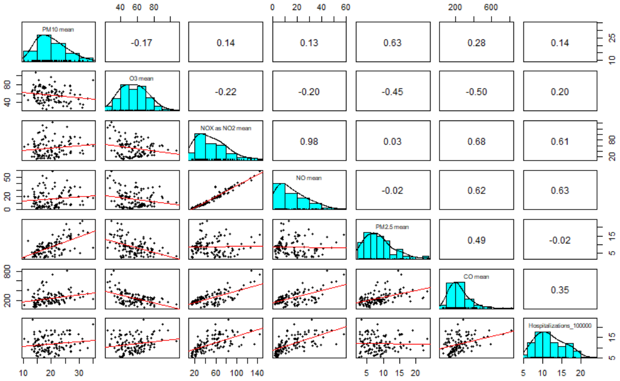

Fig. 13: Correlation between each pollutant and relative number of hospitalizations per 100 000 residents for age group [45,65[.

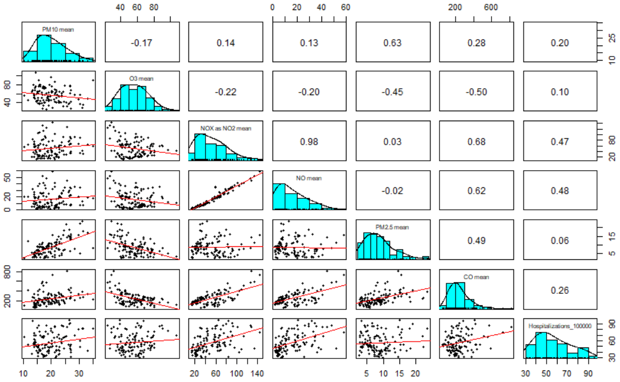

Fig. 14: Correlation between each pollutant and relative number of hospitalizations per 100 000 residents for age group [65,120[.

Regression Tables

Tab. 4: Model with months with a significance (relevance) level interval of [0, 0.001] (***),]0.001, 0.01] (**),]0.01, 0.05] (*),]0.05, 0.1] (.) and]0.1,1] ().

	Estimate	Standard Error	t value	Pr($T > \|t\|$)	Relevance
(Intercept)	4.936	1.072	4.604	1.42×10^{-05}	***
Region LVT	1.620e	4.622×10^{-01}	3.504	0.000729	***
Region Madeira	2.805	7.864×10^{-01}	3.567	0.000592	***
Region Norte	1.986	8.916×10^{-01}	2.227	0.028536	*
PM10 mean	1.636×10^{-02}	3.463×10^{-02}	0.472	0.637857	
PM10 max	3.911×10^{-05}	2.836×10^{-03}	0.014	0.989029	
O3 mean	9.861×10^{-04}	1.486×10^{-02}	0.066	0.947247	
O3 max	1.082×10^{-03}	5.675×10^{-03}	0.191	0.849186	
NOX as NO2 mean	1.932×10^{-02}	2.763×10^{-02}	0.699	0.486142	
NOX as NO2 max	-1.969×10^{-03}	2.211×10^{-03}	-0.891	0.375541	
NO mean	-4.813×10^{-02}	5.392×10^{-02}	-0.893	0.374560	
NO max	2.499×10^{-03}	3.909×10^{-03}	0.639	0.524368	
PM2.5 mean	1.174×10^{-02}	5.540×10^{-02}	0.212	0.832655	
PM2.5 max	-5.347×10^{-04}	8.383×10^{-03}	-0.064	0.949297	
CO mean	-7.581×10^{-04}	2.678×10^{-03}	-0.283	0.777807	
CO max	-1.601×10^{-04}	2.451×10^{-04}	-0.653	0.515347	
Month 2	-8.584×10^{-01}	5.063×10^{-01}	-1.695	0.093609	.
Month 3	-4.031×10^{-01}	5.755×10^{-01}	-0.700	0.485512	
Month 4	-1.814	5.943×10^{-01}	-3.051	0.003030	**
Month 5	-1.251	6.164×10^{-01}	-2.030	0.045494	*
Month 6	-1.828	6.909×10^{-01}	-2.646	0.009681	**
Month 7	-2.336	6.575×10^{-01}	-3.552	0.000622	***
Month 8	-2.3121	$7.512 \times 10^{-01}1$	-3.078	0.002798	**
Month 9	-2.074	6.490×10^{-01}	-3.196	0.001948	**
Month 10	-1.135	5.864×10^{-01}	-1.936	0.056122	.
Month 11	-1.513	5.290×10^{-01}	-2.859	0.005325	**
Month 12	-5.110×10^{-01}	4.935×10^{-01}	-1.036	0.303286	

Tab. 5: Stepwise model without months for age group [0,25[, with a significance (relevance) level interval of [0, 0.001] (***),]0.001, 0.01] (**),]0.01, 0.05] (*),]0.05, 0.1] (.) and]0.1,1] ().

	Estimate	**Standard Error**	**t value**	**Pr($T > \|t\|$)**	**Relevance**
(Intercept)	7.756233	2.053786	3.777	0.000262	***
CO mean	0.037099	0.004107	9.034	7.80×10^{-15}	***
PM2.5 max	0.069606	0.14040	4.958	2.69×10^{-06}	***
O3 max	-0.076236	0.11867	-6.424	3.75×10^{-09}	***
NOX as NO2 mean	-0.058693	0.017279	-3.369	0.000958	***
O3 mean	0.064723	0.025380	2.550	0.012183	*

Tab. 6: Stepwise model without months for age group [25,45[, with a significance (relevance) level interval of [0, 0.001] (***),]0.001, 0.01] (**),]0.01, 0.05] (*),]0.05, 0.1] (.) and]0.1,1] ().

	Estimate	**Standard Error**	**t value**	**Pr($T > \|t\|$)**	**Relevance**
(Intercept)	2.836398	0.694623	4.083	8.64×10^{-05}	***
LVT	0.821973	0.351644	2.338	0.02129	*
Madeira	1.395817	0.528663	2.640	0.00953	**
Norte	0.450575	0.626935	0.719	0.47391	
NOX as NO2 mean	0.019029	0.007168	2.655	0.00916	**
O3 mean	0.026568	0.010670	2.490	0.01433	*
O3 max	-0.006292	0.003966	-1.587	0.11557	

Tab. 7: Stepwise model without months for age group [45,65[, with a significance (relevance) level interval of [0, 0.001] (***),]0.001, 0.01] (**),]0.01, 0.05] (*),]0.05, 0.1] (.) and]0.1,1] ().

	Estimate	**Standard Error**	**t value**	**Pr($T > \|t\|$)**	**Relevance**
(Intercept)	7.659543	1.539239	4.976	2.49×10^{-06}	***
LVT	1.497985	0.805604	1.859	00.0657	.
Madeira	5.600851	0.689694	8.121	8.57×10^{-13}	***
Norte	0.138761	1.391963	0.100	0.9208	
CO mean	0.016411	0.003439	4.772	5.81×10^{-06}	***
O3 max	-0.017274	0.008935	-1.933	0.0558	.

Tab. 8: Stepwise model without months for age group [0,25[with missing values filled, with a significance (relevance) level interval of [0, 0.001] (***),]0.001, 0.01] (**),]0.01, 0.05] (*),]0.05, 0.1] (.) and]0.1,1] ().

	Estimate	**Standard Error**	**t value**	**Pr($T > \|t\|$)**	**Relevance**
(Intercept)	6.4289073	1.9158648	3.356	0.00102	**
CO max	0.0009328	0.0005395	1.729	0.086036	.
O3 max	-0.0759220	0.0110054	-6.899	1.77×10^{-10}	***
Region LVT	-1.7840803	1.0782079	-1.655	0.100280	
Region Madeira	-5.0011780	1.0345847	-4.834	3.54×10^{-06}	***
Region Norte	-0.3689191	1.3107614	-0.281	0.778787	
PM2.5 max	0.0619065	0.0156303	3.961	0.000120	***
O3 mean	0.1274975	0.0282879	4.507	$1.40e \times 10^{-05}$	***
CO mean	0.0164281	0.0046408	3.540	0.000547	***
NO max	0.0036073	0.0024535	1.470	0.143784	

Tab. 9: Stepwise model without months for age group [25,45[with missing values filled, with a significance (relevance) level interval of [0, 0.001] (***),]0.001, 0.01] (**),]0.01, 0.05] (*),]0.05, 0.1] (.) and]0.1,1] ().

	Estimate	Standard Error	t value	Pr($T > \|t\|$)	Relevance
(Intercept)	2.719662	0.625656	4.347	2.65e-05	***
Region LVT	0.985820	0.269229	3.662	0.000355	***
Region Madeira	2.211966	0.293102	7.547	5.32e-12	***
Region Norte	1.391306	0.452013	3.078	0.002511	**
CO mean	0.002424	0.001117	2.169	0.031789	*
O3 max	-0.009556	0.003132	-3.052	0.002728	**
O3 mean	0.024697	0.008334	2.963	0.003580	**
PM10 mean	0.031139	0.020747	1.501	0.135647	

Tab. 10: Stepwise model without months for age group [45,65[with missing values filled, with a significance (relevance) level interval of [0, 0.001] (***),]0.001, 0.01] (**),]0.01, 0.05] (*),]0.05, 0.1] (.) and]0.1,1] ().

	Estimate	Standard Error	t value	Pr($T > \|t\|$)	Relevance
(Intercept)	3.961923	1.244794	3.183	0.00179	**
NO mean	0.117121	0.021255	5.510	1.65e-07	***
O3 mean	0.136726	0.015745	8.684	8.57e-15	***
O3 max	-0.038805	0.006707	-5.786	4.47e-08	***
CO mean	0.008410	0.001923	4.373	2.37e-05	***
PM2.5 max	0.024475	0.009214	2.656	0.00881	**

Tab. 11: Variables Transformation. The variables that are filled with ”-” are the ones to be discarded from the model due to correlation analysis

Age	PM10 mean	PM10 max	PM2.5 mean	PM2.5 max	O3 mean	O3 max	NOx mean	NOx max	NO mean	NO max	CO mean	CO max
1	linear	-	linear	linear	log	log	log	-	-	linear	linear	linear
5	-	-	-	-	linear	linear	log	-	-	log	-	-
6	-	linear	-	-	-	linear	-	log	-	log	log	log

Tab. 12: Prediction Results for Age group 1

Region	Predicted Value	True Value	Relative Error (%)	Predicted Hospitalizations	True Hospitalizations
Centro	1.831422	2.076797	11.815074	11.65244	18.89941
LVT	1.833913	2.036144	9.932095	11.70645	17.37143
Madeira	1.587051	1.962055	19.112840	7.58458	14.96387

Tab. 13: Prediction Results for Age group 5

Region	Predicted Value	True Value	Relative Error (%)	Predicted Hospitalizations	True Hospitalizations
Centro	1.136795	1.267205	10.291166	3.815968	4.600812
LVT	1.367814	1.307810	4.588092	5.348408	4.885714
Madeira	1.474002	1.387610	6.225957	6.310162	5.513005

Tab. 14: Prediction Results for Age group 6

Region	**Predicted Value**	**True Value**	**Relative Error (%)**	**Predicted Hospitalizations**	**True Hospitalizations**
Centro	1.736631	1.732766	0.223025	9.809973	9.742896
LVT	1.865747	1.940492	3.851893	12.425202	14.342857
Madeira	1.883410	2.012065	6.394182	12.847867	16.539014