# Relating air Pollution and Respiratory Diseases Occurrences

M. F. Teodoro[a,b,c], J. N. Garcia[a], L. M. Coelho[a], M. G. Carvalho[d,e]

[a] *Escola Superior de Tecnologia Setúbal, Instituto Politécnico de Setúbal, Estefanilha,, 2910-761 Setúbal, Portugal*
[b] *CINAV, Escola Naval, Alfeite, 2810-001 Almada,Portugal*
[c] *CEMAT, Instituto Superior Técnico, Avenida Rovisco Pais, 1, 1049-001 Lisboa, Portugal*
[d] *Instituto Superior Técnico, Avenida Rovisco Pais, 1, 1049-001 Lisboa , Portugal,*
[e] *European Parliament, Brussels, Belgium*

**Abstract.** In this article we study the impact of air pollution on children's health in Portugal. In particular, we focus our attention on the city of Barreiro. We use the general linear methods (GLM), taking advantage of all the ease of handling and analyzing data in order to relate air quality and health. We relate levels of air pollution and incidence of entries from children with symptoms of respiratory problems in the pediatric urgency service at the hospital of Barreiro. It was not easy to obtain clear and unambiguous relations in particle dispersion, air quality and health. A set of models are estimated by GLM and validated using adequate tests and residual analysis. At the end of this process, we combine the best models. The work is still going on.

**Keywords:** Air quality, particle dispersion, health, generalized linear models
**PACS:** 92.60.Sz; 02.50.-r.; 89.60.Gg.

## INTRODUCTION

In last twenty years, air quality became an important issue due a significant increasing rate of respiratory problems, mostly in children, elderly and people with respiratory problems, diseases related to air pollution [1].

The urban traffic and industry produce a large variety of pollutants, especially CO (carbon monoxide), $NO_2$ (nitrogen dioxide), $NO_x$ (nitrogen oxides), VOCs (volatile organic compounds), $SO_2$ (sulfur dioxide) and PM (particulate matter). Direct emission, fires, some chemical reactions of gases are the main causes of existence of air pollutants.

The effects of particles on health are presented in several studies, [2,3]. Some articles show that children [4], elderly, and chronically ill patients [5], particularly respiratory diseases, are very sensitive to air pollution and are usually chosen as sample for the studies in this area.

In particular, children are the most vulnerable [4,6,7] to the effects of atmospheric pollution due various reasons, such as the time they spend outdoor and their immature anatomy and physiology of the respiratory system. Even more, children have higher rates of ventilation than grown up people and they are mostly mouth breathers, issue which conduces to an initial insufficient air filtration, increasing the entry of polluting particles that can cause irritation; stunting of children also increases their exposure to traffic emissions. All these factors contribute to frequent episodes of respiratory distress, even in the presence of lower concentrations of pollutants.

In [8], the authors present some models to estimate the concentrations of particulate matter with a diameter size smaller than $10\mu m$ ($PM_{10}$) in the portuguese city of Barreiro, by generalized linear models (GLM). The data is provided by outdoor air quality station. The measured concentration values of CO, $NO_x$, VOCs and $SO_2$ allow to estimate the $PM_{10}$ concentration. The estimated values are compared with the actual measured values of $PM_{10}$ concentrations in the air outside the city.

In continuation of work presented in [8], the children were the target population of the present study once they are the most vulnerable to the air pollution consequences.This paper reproduces a part of a study about the relation between atmospheric pollutant and children respiratory diseases and presents models which incidence of entries from children with symptoms of respiratory problems in the pediatric urgency service at the hospital in the portuguese city of Barreiro, by GLM. Data are provided by outdoor air quality station. The measured concentration values of CO, $NO_x$, VOCs and $SO_2$ allow to estimate the incidence of children entries. The estimated values are compared with the actual measured values of entries incidence from children with symptoms of respiratory problems in the pediatric urgency service at the Barreiro hospital.

Results show that the quality of models depends on temperature. Using some restrictions on temperature, we can get models with good performance, so some subsets of data are considered to estimate the best models.

## THE METHOD

### GLM APPROACH

In 1972, the authors of [9] published with some detail the idea of GLM as a powerful method in Statistics, standardizing the theoretical and applied points of view about all the structure of linear regression developed until that time.

Due to the large number of models, and simplicity of development associated with rapid computational analysis, the GLM have been playing an important role in statistical analysis. The idea of GLM is the establishment of a functional relation between the variable to predict (dependent variable) and a set of other exogeneous variables (explanatory variables or covariates). This relation allows to predict the dependent variable. The dependent variables and the explanatory variables can be of any type: continuous, discrete, dichotomous, quantitative, qualitative, stochastic, non-stocastic. The response variable discrete can also be a proportion, be positive, have a nonnormal random component. At 1935, Bliss proposed the probit model to proportions; at 1944 Berkson developed the logistic regression, log-linear models for contingency tables were introduced by Birch at 1963. In 1972, Nelder and Wedderbrun proved that all these models are particular cases of a general family of models: the generalized linear models. In GLM, the random component of models belongs to exponential family and a transformation of expected value of response variable is related with explanatory variables.

The simplest models, where the explanatory variables are nonrandom and the disturbances are gaussian white noise, which are estimated by ordinary least squares, can be extended for more general models in which the disturbances are autocorrelated, heteroscedastic, nongaussian, etc, or when some of the explanatory variables are stocastic. Then, linear regression models can be estimated by generalized least squares.

In GLM, a vector $\mathbf{X}$ with p covariates ($\mathbf{X}=(X_1, X_2, \ldots X_p)$) can explain the variability of the variable of interest Y (response variable). The data are in the form $(y_i, \mathbf{x}_i)$, $i = 1, \ldots, n$, as result of observation of $(Y, \mathbf{X})$ n times.

In GLM, the response variable follows an exponential family distribution [10,11] with expected value

$$E(Y) = \mu \quad (\text{or } E(Yi) = \mu_i, \quad i = 1, \ldots, n).$$

It is also defined a differentiable and monotone link function g, such that $g(\mu) = \mathbf{Z}\beta$.

The link function g relates the random component with the systematic component of response variable Y, where $\beta = (\beta_1, \beta_2, \ldots \beta_p)^T$ is the vector of unknown parameters to be estimated and $\mathbf{Z} = (Z_1, Z_2, \ldots Z_p)$ (sometimes, we consider $\mathbf{Z} = (x_1, x_2, \ldots x_p)$.

In usual ordinary regression, we consider the link function g the identity function is $g(\mu_i) = \mu_i$ and

$$\mu_i = E(Y_i) = \mathbf{X}i\beta, \quad i = 1, \ldots, n.$$

There are different link functions in GLM. When the random component of response variable has a Poisson distribution, the link function is logarithmic and the model is log-linear. In log-linear model, each parameter $\beta_i$ is the effect of variable $X_i$ in $g(\mu_i)$.

The GLM methodology can be summarized in three steps:

1. Models formulation: identify response variable distribution, select the preliminaries covariates and specification matrix, select the link function;
2. Models adjustment: estimation of model parameters, application of suitability measures of estimates;
3. Selection and validation of models: selection of variables, outliers diagnostics, residual analysis and interpretation.

## EMPIRICAL APPLICATION

The purpose of the work was to study the relationship between the number of children admitted in urgency of the pediatric urgency service of the Barreiro´s Hospital, with symptoms of respiratory problems and atmospheric pollution levels.

In an effort to use the largest quantity of information, we use as covariates several outside air pollutant concentrations and some meteorological variables.

**THE DATA**

In this article, we estimate models which relates estimates the number of children admitted in an urgency of the pediatric service and air pollutant concentration from CO, $NO_2$, $NO_x$, $O_3$ and $SO_2$ (in $\mu g/m^3$), $PM_{10}$ concentration and meteorological variables as air temperature (T,ºC), relative humidity (RH,%) and wind velocity (WV, m/s).

Values of concentration of some air pollutants were supplied by the network monitoring stations managed by Regional Coordination and Development Commission Lisbon and Tagus Valley (CCDR - LVT). $SO_2$, NO, $NO_2$, $NO_x$, $PM_{10}$, CO and $O_3$ (ozone) pollutants are measured by these stations.

The meteorological data were collected by the rules: the station measured hourly meteorological conditions, as wind speed and direction, temperature and relative humidity; available data was statistically treated according to the Portuguese Meteorological Institute directives, i.e., for wind direction the most predominant direction of the day is considered the mean for that day; for temperature and wind speed a daily average is calculated and for relative humidity, the value measured at 9 a.m. represents the average value for the day; however there are some periods without data available, making more difficult to apply a statistical model; the daily average considered was the maximum value obtained by the 8 hour mean in the 24 hours of the day.

SPSS and MATLAB software were used to manage the data and construct the developed GLM models.

**THE MODEL SELECTION**

There were tried different link functions g. The log-linear case conduced to better models.

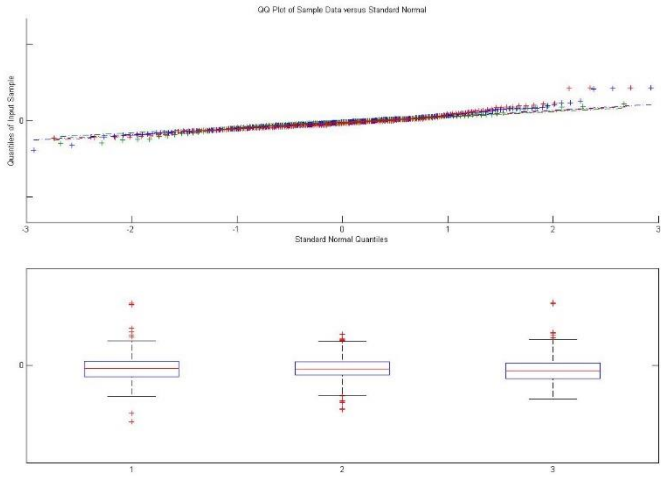
After the pre-selection of potential explanatory variables, different models were estimated and compared. The stepwise selection was used to get the best explanatory variables. Notice that interaction between covariates was taken into account.

The ´best´ models were estimated using, respectively, fractions of all sample and all sample. All estimated models were validated and its suitability verified. The rejection of the model can be suggested if some details do not agree with the expected behavior, e.g. the signs of the coefficient of some explanatory variable. In general, the validation of a model consists in a set of actions described with detail in [9]:

1. Linear restrictions to parameters tests were done, being the big majority asymptotic tests, likelihood ratio LR), Wald, Lagrange multiplier (LM), which under some hypothesis are assymptotically chi-square distributed;
2. To check the assumption of the asymptotic gaussian distribution of independent, homocedastics and no correlated residuals (normalized innovations). The following plots were made: residuals versus time, cumulative residuals versus time (detection of structural changes), cumulative squared residuals versus time, qq-plots, histograms, etc. Also, tests about autocorrelation, heteroscedasticity, bias and skewness of residuals were necessary. Equally, nonparametric tests were used, like Kolmogorov-Smirnov;
3. Other goodness adjustment statistics were analyzed, like the variance of the estimation error, the sum of squared errors, the mean of absolute error, the determination coefficient $R^2$ and the adjusted determination coefficients;
4. To decide between two selected and validated models it is necessary to compare the values obtained for statistics presented in **3.** Also, when models have a different number of parameters, it is used the Akaike information criterium (AIC), corrected Akaike information criterium (AICC), Bayes information criterium (BIC) and the maximum of likelihood.

Initially, models were estimated with the complete sample but the results were not good enough. As in [8], the estimated models provided better estimation performances when the initial data were split in two sets: i) when maximum daily temperature is greater than 25°C; ii) when maximum daily temperature is less or equal than 25°C. The fraction of all sample for higher temperatures lead to better models.

When estimated models have similar performance, the simplest models in the sense of less explanatory variables were considered. When all described validation and selection techniques were complete, we need to be able to choose correctly between different models with the same explanatory performance. The best models with similar performance conduce to estimates which are combined using a weighted mean by some criteria. The estimated models and some numerical results will be provided in the presentation of this article. Next plots are the example of what should not happen, where the normality of residuals is not present. They correspond to some initial tries, but they are not selected models. We can see some examples of the ‘’bad’’ behavior of residuals in the sense of normality. Notice the qqplots evidenciate heavy tails and box and whiskers plots detect outliers.




The tables with the main results of the best models, without heavy tails and outliers and with good performance will be presented in the complete article and in presentation.

## CONCLUSIONS AND ONGOING WORK

In [8] has been found by GLM models that estimate the urban atmospheric $PM_{10}$ in city of Barreiro based on the values of the atmospheric concentrations of other gaseous pollutants (CO, $NO_2$, $NO_x$, $O_3$ and $SO_2$) and values of meteorological variables, namely air temperature, relative humidity and wind velocity.
In the present work we could get good models which estimate the entries on pediatric urgency services of Barreiro's hospital using as explanatory variables the pollutants concentration and meteorological variables. As in [9], the best results are obtained with a subset of all sample, when daily temperature is greater than 25ªC. We can conclude that daily outdoor temperature affects the incidence of respiratory occurrences.

## ACKNOWLEDGMENTS

The authors acknowledge Comissão de Coordenação e Desenvolvimento Regional de Lisboa e Vale do Tejo (CCDR-LVT) and Instituto de Metereologia (IM) by the provided information.

## REFERENCES

1. Who, *Guidelines for indoor air quality: selected pollutants*, World Health Organization, 2010.
2. K. L. Timonen, J. Pekkanen, "Air Pollution and Respiratory Health among Children with Asthmatic or Cough Symptoms". *American Journal of Respiratory and Critical Care Medicine*, **156**, 1997, pp. 546–552.
3. F. Wei, W. Hu, J. Teng, R.S. Chapman, "Relation analysis of air pollution and children's respiratory system disease prevalence". *Chin. Environ. Sci.*, **20** (3), 2000, pp 220–224.
4. J. Pekkanen, S. T. Remes, T. Husman, M. Lindberg, M. Kajosaari, A. Koivikko, L. Soininen,"Prevalence of ashma symptoms in video and written questionnaires among children in four regions of Finland", *Eur Respir J.,* **10**, 1997, pp 1787-1794.
5. R. P. Peng, F. Dominici, R. Pastor-Barriuso, J. M. Zeger S.,Samet, *Seasonal Analyses of Air Pollution and Mortality in 100 U.S. Citie*, Berkeley Electronic Press, 2004.
6. E.E.A., *Environment and Health*, EEA report 10/2005, Denmark – Copenhagen, 2005.
7. E. P.A., *Child-specific exposure factors handbook*. U.S. Environmental Protection, 2002.
8. J. N. Garcia, M. F. Teodoro, L. M. Coelho, M. G. Carvalho, "Empirical Study of Air Quality in Barreiro *City*", in *International Conference of Mathematical methods on Science and Engineering-2014 Athens*, Edited by T. Simos et al., AIP Conference Proceedings (accepted).
9. J. A. Nelder, R. W. M. Wedderburn. "Generalized linear models". *J R Stat Soc A*, **35**, 1972, pp 370-384.
10. G. M. S. Conceição, P. H. N. Saldiva, J. M.Singer, "Modelos MLG e MAG para análise da associação entre poluição atmosférica e marcadores de morbi-mortalidade: uma introdução baseada em dados da cidade de São Paulo". *Revista Brasileira de Epidemiologia* **4**, (3), 2001, pp 206-219.
11. M. A. Turkman, G. Silva, *Modelos Lineares Generalizados da teoria à prática*, Sociedade Portuguesa de Estatística, Lisboa, 2000.