Statistical approach to the impact of air pollution on the otolaryngology system diseases

Barbara Jasiulis-Gołdyn* Dominik Nowakowski*

March 2021

Key words: Air pollution, Correlation analysis, Dependence modeling, Generalized linear model, Global health

Abstract

The World Health Organization considers pollen mixtures PM2.5 to be the most harmful to health among other types of atmospheric pollution. Particles with a smaller diameter more easily enter the body. The first contact of these dusts with the human body occurs in the respiratory tract.

Our main goal is to analyze the impact of air pollution indicator and meteorological data on the otolaryngological system on the basis of diagnosed diseases in residents of Wrocław (Poland).

Finally, we use R software to create a GLM (Generalized Linear Model) and semiparametric GLM to predict the number of temporary incapacitated workers and students unable to learn due to otolaryngological diseases based on the air pollution factors and meteorological data.

1 Introduction

Research is ongoing around the world on the impact of air pollution factors on human health. Our main goal is to create a model that explain the impact of environmental factors (mainly air pollutants) on the otolaryngology system, which is responsible for diseases of the throat, nose, ears and larynx. We investigate the impact of air quality factors on the daily number of cases, as our airways are primarily exposed to the negative effects of dust. Since the weather conditions seem to be significant, we also consider here the impact of the average air temperature or average humidity. We analyze data from the National Health Fund, which was collected in 2015 and are directly related to the city of Wrocław.

Our task will be to clarify the number of daily cases throughout the year. This will give us an overview of the year-round data compilation and we will be able to predict the intensity of the otolaryngological disease group in the coming years.

*Institute of Mathematics, University of Wrocław, pl. Grunwaldzki 2/4, 50-384 Wrocław, Poland

2 Optimalization

At the very beginning, we need to find a day in the past in which the concentration of a given factor has significantly influenced the number of diagnosed diseases of a particular day. Finally, we take into account data on the magnitude of factors from the past. For this reason we use Spearman's correlation coefficients. We investigate how many days back specific air quality and meteorological data should be taken into account to maximise the absolute value of Spearman's correlation between the number of diagnosed diseases and the factor in question. The results below indicate that the factor usually affects the number of cases with a maximum delay of three days. This applies to dust.

Variable	Spearman's correlation	shift days
As.PM10.	0.38663090	0
BaA.PM10.	0.76864594	3
BaP.PM10.	0.74069835	7
BbF.PM10.	0.76625105	3
BjF.PM10.	0.77488086	3
BkF.PM10.	0.74174381	3
Cd.PM10.	0.69833691	1
DBaH.PM10.	0.77021759	3
IP.PM10.	0.75942023	3
Ni.PM10.	-0.09422864	11
NOx.PM10.	0.27659045	5
O3	-0.56525307	12
Pb.PM10.	0.68218795	0
PM10	0.42772200	0
CO	0.51433960	12
NO2	0.15307808	5
C6H6	0.44700977	0
PM2.5	0.51059693	0
average air temperature	-0.77688749	2
average humidity	0.41433302	12
station-level pressure	0.16831916	0
sunshine	-0.44431827	14

Table 1: Air pollutants as variables with shifted day optimalization

Now let's look at the correlation structure of our data set. Since we have a counting variable here that takes values only from a set of natural numbers, we use Spearman correlation values, which use actions on the ranks. The relationship between variables can be seen in the Figure 1.
We choose explanatory variables, which are the most correlated with individual calls described by the ICD-10 codes and not correlated with another explanatory variables.

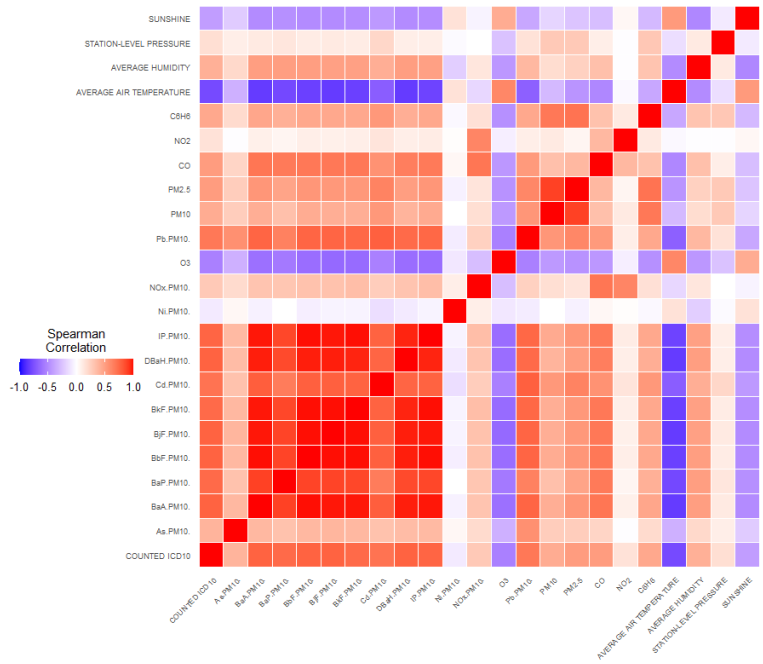

Figure 1: Matrix of Spearman correlation value for shifted days.

3 Generalized linear models and data analysis

At the beginning, we use GLM (Generalized Linear Model) with the Poisson distribution. Next, we will extend the model by adding a nonlinear factor that will be time-related creating such called semiparametric GLM (for more details on the construction see [6]). By comparing these two models, we use the AIC and BIC information criteria with values presented below:

Model	AIC	BIC
GLM	14486.36	14609.81
semiparametric GLM	6149.657	6350.27

Table 2: Value of information criteria for both models. A better model turns out to be semiparametric GLM than the classical GLM because it has smaller values for both criteria.

The results of our analyzes one can find in the Figure 2. Unfortunatelly, the explicit formula for the semiparametric model is much more complicated but easy to describe by numeric methods. It follows that we present here the GLM model for analyzed data having friendly analytical formula. The prediction of daily number of cases at time t (notation DNofC_t) described by GLM is given by the following formula (with parametres estimated by the MLE method):

$$\begin{aligned}\mathrm{DNofC_t} = \exp\{&6.418 - 0.089\ \mathrm{BaA_{t-3}} + 0.0065\ \mathrm{BaP_{t-7}} + 0.2\ \mathrm{BbF_{t-3}}\\ &- 0.33\ \mathrm{BjF_{t-3}} - 0.058\mathrm{BkF_{t-3}} + 0.047\ \mathrm{Cd_{t-1}} + 0.049\ \mathrm{DBaH_{t-3}}\\ &- 0.001\ \mathrm{O3_{t12}} + 0.32\ \mathrm{Pb_t} - 0.001\ \mathrm{PM10_t} + 0.001\ \mathrm{PM2.5_t}\\ &+ 0.022\ \mathrm{CO_{t-12}} - 0.02\ \mathrm{AAT_{t-2}} + 0.0009\ \mathrm{AH_{t-12}}\\ &- 0.0002\ \mathrm{sunshine_{t-14}} + 0.001\ \mathrm{STP_t} + 0.006\ \mathrm{C6H6_t}\}\end{aligned}$$

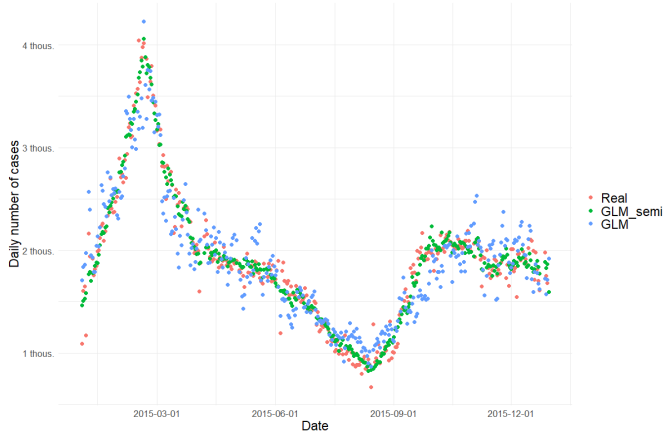

Figure 2: Prediction of daily number of cases in Wrocław in 2015 based on environmental factors and air pollutants.

References

[1] World Health Organization, https://www.who.int/air-pollution/news-and-events/how-air-pollution-is-destroying-our-health (27 August 2019).

[2] Yu O., Sheppard L., Lumley T., Koenig J., Shapiro G., *Effects of Ambient Air Pollution on Symptoms of Asthma in Seattle-Area Children Enrolled in the CAMP Study* , Environmental Health Perspectives, Vol. 108, No. 12/2000.

[3] Castanas E., Kampa M., *Human health effects of air pollution*, 2007.

[4] R. E. Arku, M. Brauer, S. H. Ahmed, K. F. AlHabib, A. Avezum, J. Bo, T. Choudhury, A. ML. Dans, R. Gupta, R. Iqbal, N. Ismail, R. Kelishadi, R. Khatib, T. Koon, R. Kumar, F. Lanas, S. A. Lear, L. Wei, P. Lopez-Jaramillo, V. Mohan, P. Poirier, T. Puoane, S. Rangarajan, A. Rosengren, B. Soman, O. T. Caklili, S. Yang, K. Yeates, L. Yin, K. Yusoff, T. Zatonski, S.Yusuf, P. Hystad *Long-term exposure to outdoor and household air pollution and blood pressure in the Prospective Urban and Rural Epidemiological (PURE) study*, Environmental Pollution 262 (2020)

[5] P. Hystad, A. Larkin, S. Rangarajan, K. F. AlHabib, Á. Avezum, K. B. T. Calik, J. Chifamba, A. Dans, R. Diaz, J. L. du Plessis, R. Gupta, R. Iqbal, R. Khatib, R. Kelishadi, F. Lanas, Z. Liu, P. Lopez-Jaramillo, S. Nair, P. Poirier, O. Rahman, A. Rosengren, H. Swidan, L. A. Tse, L. Wei, A. Wielgosz, K. Yeates, K. Yusoff, T. Zatonski, R. Burnett, S. Yusuf, M. Brauer *Associations of outdoor fine particulate air pollution and cardiovascular disease in 157436 individuals from 21 high-income, middle-income, and low-income countries (PURE): a prospective cohort study*, The Lancet Planetary Health Vol. 4, Issue 6, E235-E245, 2020

[6] J. Harezlak, D. Ruppert, M. Wand *Semiparametric Regression with R*, Springer (2018)