Modeling the impact of air pollution on the respiratory system diseases

Bożena Cegiełka*, Barbara Jasiulis-Gołdyn* *

Abstract

Smog is a serious problem in most big urban areas. We rarely realize the consequences of being in polluted air, while mixture of air pollutants can seriously endanger human health. Bronchitis, pneumonia and asthma are only some of the respiratory diseases that are associated with the effects of smog. Polluted air also makes it difficult for people to breathe properly. We analyze the relationship between respiratory diseases and smog based on data from Wrocław (Poland) regarding calls for ambulance services of 15107 individuals and indicators of air pollution and meteorological data in 2016. The results of our analyzes are optimized for the selection of explanatory variables for models using the Spearman coefficient values. A novel approach proposed here is to use generalized linear models by optimizing shifted air pollution data to predict the number of ambulance calls on a given day. Finally, the best generalized linear model with logarithm linking function was fitted to analyzed data.

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

1 Introduction

Wrocław (Poland) is one of the city with the most polluted air in the winter months. This fact has many social and economic consequences, including health losses for the city's inhabitants. Ambulance calls, which we consider here are strictly connected with the economic losses, since the patients take the sick leaves and consistently employers lose money. Even if you think that this problem does not apply to you, with the butterfly effect, it may turn out that even a resident of another part of the World pays for it by being exposed to climate change caused by ingredients chemicals of polluted air.

In this paper we analyze a database of ambulance calls, weather and pollution data from Wrocław. We have completed minor meteorological and pollution data gaps linearly.

In Section 2 we carry out a correlation analysis and calculate the values of Spearman correlation coefficients between the number of ambulance calls and environmental indicators. At the next step we examine after what time the greatest impact of pollution or meteorological factors on respiratory diseases is

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

noticeable. Finally in Section 3 we create a model that can be used to predict the number of ambulance calls to patients with respiratory diseases based on weather conditions and the level of pollution. For this purpose, we use a generalized linear model for count data.
All analyzes carried out in this paper were made using the R software environment.

2 Correlation analysis

The Spearman correlation coefficient is significant if you consider monotonic dependence structure. We use here correlation analysis to investigate whether there is any relationship between the number of ambulance calls and air pollution or meteorological data. In order to optimize the number of significant environmental factors and air pollutants it is important to look at shifted data. Intuitively, illness should not occur immediately after exposure to harmful substances, but only after some time.

Definition 2.1 *The ρ - Spearman correlation coefficient has the following formula:*

$$r_s = 1 - \frac{6 \cdot \sum_{i=1}^{n} d_i^2}{n(n^2-1)},$$

where n is the number of observations, $d_i = R(x_i) - R(y_i)$ is the difference between the i^{th} rank for the variable x, and the i^{th} rank for the variable y. The rank determines the position on which the observation is located after sorting the data.
The natural question is after how many days the given pollution or weather conditions have the greatest impact on health. For this purpose, we calculated the Spearman correlation coefficients for shifted environmental data for the selected number of days and the number of ambulance calls. Table 2 lists the optimized number of days to improve the Spearman coefficients. The values of pollutants meteorological data were shifted to the point where the greatest Spearman correlation occurred.

3 Generalized linear models

Generalized linear models (GLMs) were formulated by John Nelder and Robert Wedderburn ([5]) as a way of unifying various other statistical models, including linear regression, logistic regression and Poisson regression. GLMs are one of the most useful modern statistical tools, because they can be applied to many different types of data. GLM is a generalization of ordinary linear regression and for this model, it is acceptable that the response variables have a non-normal distribution.
Before proceeding to the construction of the model, with the help of tools in the R environment, we will present methods for assessing the quality of model fit. We use the following measures to compare the quality of a set of statistical models to each other. We want to choose the right number of predictors and avoid overfitting the model. For this purpose, we want to minimize Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC).
We arrive at 3 GLMs (BACKWARD, FORWARD, STEPWISE methods) with Information Crterions described in Table 1. Finally, in Table 2 we present the

parameters estimated by the MLE method for GLM described in Table 1 and lists of significant variables connected with air pollutants and weather factors for data analysis based on database of ambulance calls from Wrocław in 2016.

Table 1: Characteristics of created models with quality measures

method of variables selection	BACKWARD	FORWARD	STEPWISE
Poisson regression model	$4.57837 + \sum_{i=1}^{23} \beta_i x_i$	$5.26886 + \sum_{i=1}^{17} \beta_i x_i$	$5.46231 + \sum_{i=1}^{16} \beta_i x_i$
Link function	log	log	log
AIC	2405.2	2406.1	2403.9
BIC	2498.865	2476.343	2470.215
RSS	13528.63	13882.57	13896.94
Residual deviance	329.67	342.57	342.34
Degrees of freedom	342	348	349
Spearman correlation*	0.7196309	0.694976	0.6949126

NOTE: *- Spearman's correlation between the values fitted by the model and the real number of ambulance calls

Analyzing the presented results, we draw several conclusions. The differences in fitting the created models for all three variable selection methods are not significant. The model created using the backward method of variable selection is the model with the largest number of variables, which, however, positively affects the value of residual deviance and RSS which are the lowest among all models. The model with the least number of variables, i.e. the one created using the stepwise method, has the smallest value of both information criteria. For the lists of estimated parameters (by the MLE method) and explanatory variables describing air pollutants see Table 2 at the end of the paper after References.
Acknowledgements. This paper is partially supported by the project "First order Kendall maximal autoregressive processes and their applications", Grant no POIR.04.04.00-00-1D5E/16, which is carried out within the POWROTY/ REINTEGRATION programme of the Foundation for Polish Science co-financed by the European Union under the European Regional Development Fund.

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] Chief Inspectorate of Environmental Protection, http://www.gios.gov.pl (2 August 2019).

[5] Nelder J., Wedderburn R., *Generalized Linear Models* , Journal of the Royal Statistical Society. Series A (General), Vol. 135, No. 3/1972, p. 370-384.

Table 2: Estimated parameters for created models

x_i	shift(days) optimized	BACKWARD β_i	FORWARD β_i	STEPWISE β_i
DsWrocWybCon_BkF.PM10._mean	15	-	0.0526702	0.0529895
DsWrocWybCon_BbF.PM10._mean	15	0.0188377	-	-
DsWrocWybCon_BjF.PM10._mean	18	-	-	-
DsWrocWybCon_BaA.PM10._mean	15	-0.0273202	-	-
DsWrocWybCon_BaP.PM10._mean	15	-	-	-
temperature_mean	17	-0.0060180	-	-
DsWrocWybCon_IP.PM10._mean	15	-	-0.0536273	-0.0539423
DsWrocWybCon_DBahA.PM10._mean	13	-	-	-
temperature_max	13	-	-0.0073691	-0.0080720
sensible_temperature	17	-	-	-
DsWrocWybCon_C6H6_min	9	-	-	-0.0168097
temperature_min	12	-0.0050892	-	-
DsWrocWybCon_C6H6_mean	29	-	-0.0073514	-
DsWrocWybCon_C6H6_max	31	-	-	-
DsWrocBartni_O3_max	28	-	-0.0007691	-0.0006843
DsWrocWybCon_O3_max	27	-0.0012807	-	-
DsWrocBartni_O3_mean	31	0.0036058	-	-
DsWrocWybCon_O3_mean	31	-	-	-
DsWrocWybCon_Pb.PM10._mean	31	-	-	-
daily_jump_of_humidity	11	-	-	-
air_humidity_min	27	-	-	-
DsWrocAlWisn_PM2.5_max	9	0.0006978	0.0005943	0.0007880
DsWrocWybCon_PM2.5_max	27	0.0009229	-	-
DsWrocWybCon_Cd.PM10._mean	30	0.0894244	0.0514380	0.0685657
DsWrocWybCon_SO2_mean	26	-	-	-
DsWrocWybCon_PM2.5_mean	8	-0.0014359	-	-
DsWrocWybCon_CO_mean	14	-0.1676498	-	-
DsWrocWybCon_SO2_min	31	-	-	-
DsWrocAlWisn_PM2.5_mean	10	-	-	-
DsWrocBartni_NO2_min	26	-	-	-
DsWrocAlWisn_CO_mean	0	-	-	-
DsWrocWybCon_PM2.5_min	3	-	-	-
daily_jump_of_temperature	13	-0.0059996	-	-
air_humidity_mean	27	-	-	-
DsWrocAlWisn_CO_min	0	-	0.1120757	0.1091200
DsWrocWybCon_Ni.PM10._mean	2	-	0.0211369	-
DsWrocWybCon_CO_max	31	0.0746910	-	-
DsWrocWybCon_NO2_min	26	-	-	-
DsWrocWybCon_NOx_min	26	-	-	-
DsWrocWybCon_PM10_mean	31	-0.0003768	-0.0002448	-0.0002699
DsWrocAlWisn_PM2.5_min	10	-	-	-
DsWrocWybCon_CO_min	7	-	-	-
DsWrocWybCon_NOx_mean	27	-	-	-
DsWrocBartni_NOx_min	26	-	-	-
DsWrocAlWisn_CO_max	27	-	-	-
DsWrocBartni_NO2_mean	27	-	-	-
DsWrocWybCon_NO2_mean	27	-	-	-
DsWrocWybCon_As.PM10._mean	0	-	-	-
DsWrocWybCon_SO2_max	23	-	-	-
wind_speed_mean	7	0.0142198	0.0153561	0.0159375
DsWrocAlWisn_NOx_mean	27	0.0009294	0.0008694	0.0008365
DsWrocWybCon_O3_min	31	-	-	-
DsWrocAlWisn_NOx_max	24	-	-	-
daily_jump_of_pressure	19	-	-	-
DsWrocWybCon_NOx_max	27	-	-	-
DsWrocBartni_NOx_mean	27	-	-	-
DsWrocBartni_O3_min	27	-0.0019613	-	-
DsWrocAlWisn_NO2_max	22	-	-	-
DsWrocBartni_NOx_max	27	-	-	-
air_pressure_max	20	0.0028493	0.0024775	0.0023944
wind_speed_max	31	-	-	-
DsWrocAlWisn_NO2_mean	1	-0.0026473	-0.0026814	-0.0025940
DsWrocBartni_NO2_max	14	-	-	-
air_pressure_min	0	-	-	-
air_humidity_max	30	-	-	-
air_pressure_mean	0	-0.0036883	-0.0035261	-0.0036003
DsWrocWybCon_NO2_max	27	-	-	-
DsWrocAlWisn_NO2_min	25	-	-	-
DsWrocAlWisn_NOx_min	31	0.0008829	-	-