Application of the Data Mining Methods to Assess the Impact of Meteorological Conditions on the Episodes of High Concentrations of PM10 along the Polish – Czech Border

Leszek Ośródka*, *Ewa* Krajny, and *Marek* Wojtylak

IMWM-NRI, Department Modelling of Air Pollution, 10 Bratków Str., 40-045 Poland

Abstract. The paper presents an attempt to use selected data mining methods to determine the influence of a complex of meteorological conditions on the concentrations of PM10 (PM2.5) proffering the example of the regions of Silesia and Northern Moravia. The collection of standard meteorological data has been supplemented by increments and derivatives of measurable weather elements such as vertical pseudo-gradient of air temperature. The main objective was to develop a universal methodology for the assessment of these impacts, i.e. one that would be independent of the analysed pollution. The probability of occurrence (at a given location) of the assumed concentration level as exceeding the value of the specified distributional quintile was adopted as the discriminant of the incidence. As a result of the analyses conducted, incidences of elevated concentrations of air pollution particulate matter PM10 have been identified and the types of weather responsible for the emergence of such situations have also been determined.

1 Introduction and Aim of the Study

In spite of continuous measures undertaken by the European Union (EU) in order to improve air quality (overarching strategies: "*EU Thematic Strategy on Air Pollution*", "*EU Clean Air Policy Package*") and despite the implementation of many legal instruments (such as the "*Air Quality Directive*", or the "*National Emission Ceilings Directive*"), air pollution remains one of the most poignant problems in Central Europe [1]. Scientific research is proffering increasingly copious evidence of the negative impact of air pollution on the health of the population, both as a direct exposure and as an indirect result of exposure to pollutants accumulating in plants, soil or water [2]. The quality of atmospheric air in view of the presence of pollutants such as lead (Pb), sulphur dioxide (SO_2) or benzene (C_6H_6) has been systematically improving. On the other hand, a significant percentage of the population, especially in highly urbanised areas, is still affected by harmful levels of pollutants such as suspended particulate matter PM in the planetary boundary layer (O_3), nitrogen dioxide (NO_2) or polycyclic aromatic hydrocarbons such as

* Corresponding author: leszek.osrodka@imgw.pl

benzo(a)pyrene (BaP). The cause of poor air quality adversely impacting on health is the emission of substances from transport, industry, agriculture or households. The reduction of air pollution with substances such as particulate matter requires actions to reduce the natural or antropogenic emissions from different sources (point, area,, linear).

It is generally known that meteorological conditions determine the spreading of pollutants from their sources and their dispersion in the atmosphere. Constituting the most important and decisive meteorological factor here are the direction and speed of the wind, as well as the temperature of the air and precipitation. The wind affects the horizontal and vertical spreading of air pollution. The influence of air temperature on the level of pollution is both indirect and direct. Altitude-specific changes in the air temperature, or the so-called vertical thermal structure, shape the stability of the atmosphere. The more stable the thermal structure is – or as the temperature lowers, does not alter (isothermy) or even increases (inversion) with the altitude - the more unfavourable are the conditions for horizontal scattering of pollutants. Additionally, In wintertime, the air temperature "controls" the amount of pollutant emissions, thus indirectly influencing the concentration volumes. In turn, atmospheric precipitation causes diluting, reducing the level of the substance in the air. Insolation is also strongly impactful on photochemical pollution.

The aim of the study is to comprehensively assess the influence of the meteorological factors on the formation of episodes of high levels of particulate matter in areas with a high degree of anthropopressure.

A review of the reference literature identifies many works related to the study of the influence of meteorological conditions on the formation of high concentrations of air pollutants; however, most of them either constitute a description of such cases [3] or an undertaking to statistically analyse the influence of particular meteorological elements on air quality [4]. Nonetheless, considering the fact that the episodes of high concentrations of pollutants are not linearly dependent on particular meteorological elements but are the result of many factors, such assessments must be deemed as incomplete. Hence, in this paper an attempt has been made to use data mining methods that allow a more comprehensive assessment of the influence of many meteorological elements on air quality.

This article presents the results of quantification of the above-mentioned relations between suspended particulate matter PM10 in the planetary boundary layer and the weather in the cross-border area of Silesia (PL) and Moravia (CZ). A collection of standard meteorological data has been supplemented with increments and derivatives of measurable weather elements such as the vertical pseudogradient of air temperature.

1.1 Data and the Scope of the Study

The analysis of weather conditions responsible for episodes of high concentrations of air pollutants: suspended particulate matter and ground ozone was carried out for the area of the Polish - Czech border area in the Silesian and Moravian regions. On the Czech side, this region included the Moravian-Silesian region comprising the Ostrava-Karviná agglomeration as well as the Bruntál, Frydek-Mistek, Nový Jičín and Opava counties. On the Polish side, it was a region of the Silesian Voivodeship encompassing the Upper Silesian agglomeration, and the rural counties of Bielsko, Cieszyn, Racibórz, Rybnik, Wodzisław, Żywiec, followed by the town counties of Bielsko-Biala, Jastrzębie-Zdrój, Rybnik and Żory. The total study area amounted to 10 370 km^2, of which 52.5% was the Czech territory inhabited by 38.0% of the population in relation to the total number of persons for the entire area. In turn, the population density of 648.9 persons /km^2 was twice the size on the Polish side of the border [5, 6]. The described regions of the Silesian and Moravian-Silesian voivodships are among the most urbanised and industrialised in Poland and in the Czech Republic. This area is very diverse in terms of physical, geographical and

socio-economic conditions. It is located in four provinces: the Central European Lowlands, the Polish Highlands, the Czech Massif and the Western Carpathians [7, 8]. In terms of its climate, it is located in a temperate zone with a characteristic annual course of four seasons. As part of the study, a collection of measurement data covering a period from October 2006 to March 2011 was analysed. The meteorological data was obtained from the measurement network of national meteorological services, four of which belong to the Czech Hydrometeorological Institute (ČHMÚ): Ostrava-Mošnov, Ostrava-Poruba, Lučina, Červená, Lysá hora and three to the Polish Institute of Meteorology and Water Management National Research Institute (IMGW-PIB): Bielsko-Biała, Katowice, Racibórz. The table 1 below shows the air quality monitoring stations included in the analyses.

Table 1. Air quality PM10 on monitoring stations used in the analysis.

Station name		**Type of station / area**	**Localization** Latitude N / Longitude E *degrees (°), minutes ('), seconds (")* / Elevation above sea level *(m a.s.l.)*
CZ - ČHMÚ	Bohumín	B/S	49°54'15" / 18°20'50" / 200
	Český Těšín	B/U	49°44'56" / 18°36'35" / 285
	Frýdek-Místek	B/S	49°40'18" / 18°21'04" / 290
	Havířov	B/U	49°47'28" / 18°24'25" / 260
	Karviná	B/S	49°51'50" / 18°33'05" / 238
	Opava - Kateřinky	B/U	49°56'42" / 17°54'34" / 255
	Orlová*)	B/U	49°52'32" / 18°26'01" / 266
	Ostrava - Českobratrská (hot spot)*)	T/U	49°50'23" / 18°17'24" / 215
	Ostrava - Fifejdy	B/U	49°50'21" / 18°15'49" / 220
	Ostrava – Poruba (ČHMÚ)*)	B/S	49°49'31" / 18°09'33" / 242
	Ostrava - Přívoz*)	I/U	49°51'23" / 18°16'11" / 207
	Ostrava - Radvanice (ZÚ)*)	I/S	49°48'25" / 18°20'21" / 263
	Ostrava - Zábřeh	B/U	49°47'46" / 18°14'50" / 236
	Studénka	B/R	49°43'15" / 18°05'22" / 231
	Třinec - Kosmos	B/U	49°40'05" / 18°40'40" / 320
	Věřňovice*)	B/R	49°55'29" / 18°25'22" / 203
PL - Silesian Voivodship Inspector of Environmental Protection	Bielsko-Biała	B/U	49°48'36" / 19°01'37" / 365
	Cieszyn	B/U	49°44'00" / 18°38'20" / 353
	Dąbrowa Górnicza	B/U	50°19'36" / 19°13'52" / 293
	Gliwice	B/U	50°16'36" / 18°39'18" / 236
	Godów*)	B/U	49°55'19" / 18°28'17" / 205
	Katowice-Muchowiec	B/U	50°15'36" / 18°58'30" / 273
	Rybnik	B/U	50°06'36" / 18°30'58" / 245
	Tychy	B/U	50°06'00" / 18°59'24" / 252
	Wodzisław Śląski	B/U	50°00'36" / 18°27'18" / 271
	Zabrze	B/U	50°18'00" / 18°47'53" / 257
	Żywiec	B/U	49°41'36" / 19°12'22" / 352

type of station: T – traffic, I – industrial, B – background; type of area station: U – urban, S – suburban, R – rural; *) - complementary station; ZÚ station (Public Health Institute Ostrava).

1.2 Methodological Principles of Data Analysis

Although meteorological conditions determine the formation of high concentrations of particulate matter, however, the existing dependencies are not linear. Taking this into consideration, data mining methods were used to evaluate the influence of meteorological factors on the quality air [9, 10]. The main aim was to elaborate a universal methodology, i.e. one that would be independent of the analysed pollution. The probability of occurrence of a given concentration level in excess of the value of a particular quantile distribution was assumed as the discriminant of the episode.

1.2.1 Identification of Episodes High Concentrations of PM10

First, the measurement data were unified, replacing the absolute values of PM10 with the probability of occurrence of a given concentration at the station. For data sequences $x_1,\ldots,x_n$ representing average daily PM10 concentrations the probabilities were calculated from the moving averages, according to the following formula:

$$p_i = \frac{\overline{\{x \in \{x_i,\ldots,x_n\}; x \geq x_i\}}}{n} \tag{1}$$

where: symbol $\overline{=}$ denotes the cardinality (numerical amount) $i=1,\ldots,n$.

Thus, the concentrations were characterised by the frequency of their occurrence together with the higher values. Subsequently, average probability for each day was calculated for all stations actively measuring on that day, resulting in an averaged (area average) sanitary air quality. The lower probability corresponded to the higher levels of the analysed pollution:

$$\Pr(day, month, year) = \frac{\sum_{k=1}^{N} p_k}{L_s} \tag{2}$$

where: L_s – the number of measuring stations operating on a given day.

1.2.2 Identification of Meteorological Situations Associated with High Concentrations of PM10

The characteristics of the daily meteorological situation were obtained by averaging of a number vector consisting of 59 hourly elements and the sum of the daily precipitation. The weather for each day was represented by the p vector, as per the following formula:

$$p^{(i)} = [\tilde{p}_1^{(i)},\ldots,\tilde{p}_{59}^{(i)}] \quad i=1,\ldots,N. \tag{3}$$

Thus, for the period from October 2006 – March 2011 (N) for PM, a collection of 2008 cases was obtained. In order to standardise the data, a transformation was performed for each coordinate of the meteorological vectors:

$$p_j^{(i)} = \frac{\widetilde{p}_j^{(i)} - \widetilde{p}_j^{\min}}{\widetilde{p}_j^{\max} - \widetilde{p}_j^{\min}} \qquad where$$
$$\widetilde{p}_j^{\min} = \min_{i=1,..,N} \widetilde{p}_j^{(i)} \qquad \widetilde{p}_j^{\max} = \max_{i=1,..,N} \widetilde{p}_j^{(i)} \tag{4}$$

In this way, data was prepared in the form of vectors $p = [p_1, ..., p_{59}]$ with coordinates in the [0,1] interval, which were used for further analyses.

1.2.3 Determining the Weather Patterns Responsible for the Occurrence of Smog Episodes

The $P:=\{p^{(1)},\ldots,p^{(N)}\}$ sequence of daily probabilities of occurrence of a given pollution level was sorted in an ascending order. It was then assigned a sequence of weather vectors corresponding to the given sanitary air quality. In the next stage, grouping of set elements was performed. In order to allow clustering of the most similar meteorological situations, a limit distance $r > 0$ was determined. The value of the r parameter was determined based on the clustering index referred to as the Davis-Bouldin index. For the start, the vector of the weather for which the worst sanitary air quality occurred was assumed as the pattern. And this weather was the first one to be included in the k^{th} group. Examining the P sequence starting from the second position, the k^{th} group was subsequently enlarged by those $p^{(l)}$ whose Manhattan distance from $p^{(1)}$ was lower than the set r value; afterwards, the $p^{(l)}$ vector was removed from the P sequence.

$$\sum_{j=1}^{59} \left| p_j^{(1)} - p_j^{(l)} \right| < r \quad \text{l=2,..,M.} \tag{5}$$

Thus, groups of several of numerical amount from several to dozens of vectors were designated. However, there have been cases - especially for smog situations with the highest concentration of pollutants (unusual weather) - in which the resultant group was one-element. On the other hand, a certain imperfection of the Davis-Bouldin index is constituted by the fact that it achieves a minimum of zero for one-element groups. Therefore, the proposed modification of this index for the purposes of these analyses entailed an insertion of the $2/\overline{\overline{S_i}}$ component where S_i represents the numerical amount of the i^{th} group:

$$DB_{\text{mod}} = \frac{1}{n}\sum_{i=1}^{n} \frac{2}{\overline{\overline{S_i}}} + \max_{j,j\neq i} \frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \quad c_x - \text{centroid of the } x^{\text{th}} \text{ cluster} \tag{6}$$

where: $\overline{\overline{x}}$ – an average distance of the elements of the x^{th} cluster from its centroid, x=1,…,n (the number of all clusters), and $d(c_i,c_j)$ is the centroid distance. Grouping was performed many times for different r values, assuming as appropriate such r_{min} for which DB_{mod} (r_{min}) reached the lowest value.

Due to the large variability of meteorological conditions, which resulted in the small number of groups containing meteorological situations responsible for the highest concentration of pollutants, it was decided to cluster similar groups into larger aggregates. For this purpose, a c centroid was calculated for each group, or, in other words, a vector composed of 59 coordinates formed by averaging the same coordinates of the weather vectors belonging to the given group. Then proceeding as described above, the resulting clusters were combined into groups, which were referred to as weather patterns. As a result for each pollutant, three weather patterns (types) were assigned for each PM level.

2 Results of the Study and Discussion

The result of the conducted research was a clustering of common features (meteorological elements) with high mean area concentrations of PM pollutants, followed by the determination of the types of weather responsible for such situations.

2.1 Particulate Matter PM

Table 2 shows for selected weather stations the characteristics of the basic meteorological elements that distinguish separate types of weather responsible for particulate matter PM episodes.

Table 2. Meteorological characteristic of the types of weather responsible for the situations of high PM10 concentrations.

Meteo.element (24h mean values)	**Meteorogical station**	**Cool half of the year (Oct – Mar)**	**Winter season (Nov – Feb)**	**Episode PM10**					
				Type I		**Type II**		**Type III**	
		Probability		**Size of data set / Probability**					
		0.412	**0.403**	**16**	**0.020**	**78**	**0.142**	**35**	**0.111**
Mean area concentration 24-hr daily PM10		**60.1**	**64.5**	**258.0**		**105.2**		**115.5**	
Air temperature [^oC]	Ostrava-Mošnov	2.7	0.8	-13,0		-0.2		-3.1	
	Červená	0.1	-1.8	-8.6		-1.7		-3.1	
	Lysá hora	-2.3	-3.8	-8.2		-3.5		-4.1	
	Bielsko-Biała	2.8	0.9	-10.9		0.3		-3,0	
	Katowice	2.4	0.5	-12.2		-0.4		-2.3	
	Racibórz	2.8	0.9	-13.8		-0.4		-2.7	
Wind speed [$m \cdot s^{-1}$]	Ostrava-Mošnov	3.9	4.1	1.9		2,0		2,0	
	Červená	4.4	4.5	3.6		3.4		3.2	
	Lysá hora	8,0	8.4	6.7		8.5		6.8	
	Bielsko-Biała	3.4	3.5	1.4		2.3		1.9	
	Katowice	2.8	2.9	1.5		2,0		2,0	
	Racibórz	3.6	3.7	1.7		2,0		1.8	
Calm daily [%]	Ostrava-Mošnov	7.5	6.5	25.5		15.2		18.8	
	Červená	0.1	0.1	0.3		0.2		0.1	
	Lysá hora	0	0	0		0		0	
	Bielsko-Biała	4.8	5,0	19,0		6.8		6,0	
	Katowice	7.5	6,0	17.4		11.5		12.4	
	Racibórz	3.8	3.6	9.4		7.5		8.3	
Wind direction [degrees]	Ostrava-Mošnov	237	236	220		23		39	
	Červená	241	240	184		149		76	
	Lysá hora	248	247	118		164		113	
	Bielsko-Biała	207	210	131		80		88	
	Katowice	235	220	135		85		90	
	Racibórz	220	215	133		86		111	
Height of the cloud base [km]	Ostrava-Mošnov	5.6	5.3	6.4		5.1		4.6	
	Katowice	5.6	5.4	8.2		5.4		5.8	

Meteo.element (24h mean values)	**Meteorogical station**	**Cool half of the year (Oct – Mar)**	**Winter season (Nov – Feb)**	**Episode PM10**					
				Type I		**Type II**		**Type III**	
		Probability		**Size of data set / Probability**					
		0.412	**0.403**	**16**	**0.020**	**78**	**0.142**	**35**	**0.111**
Relative air humitity [%]	Ostrava-Mošnov	84,0	86,0	77.9		89.6		89.2	
	Katowice	83.7	86,0	78.7		86.6		86.3	
Atmospheric pressure a.s.l. [hPa]	Ostrava-Mošnov	1017.7	1017.9	1036.0		1010.3		1029.2	
	Katowice	1017.5	1017.6	1035.8		1010.8		1029.3	
Effective snnshine duration [hr]	Ostrava-Mošnov	25.2	17.8	47.3		12.1		29.7	
	Katowice	24.9	17.4	44.5		17.9		29.4	
Cloud cover [octant]	Ostrava-Mošnov	5.9	6.3	2.8		6.6		4.7	
	Katowice	5.8	6.1	2.1		6.4		4.9	
Vertical pseudogradient of temperature [ºC·100 m^{-1}]	Ostrava-Mošnov - Červená	0.5	0.5	-0.9		0.3		0	
	Ostrava-Mošnov - Lysá hora	0.5	0.4	-0.4		0.3		0.1	
Atmos. precipitation [mm]	Ostrava-Mošnov	1.1	1.1	0		1.1		0	
	Katowice	1.7	1.8	0.1		1.5		0	

An analysis of the results (Table 2) showed that the highest concentrations of particulate matter PM10 were observed in the weather type no. I, which is characterised by the lowest probability of occurrence and occurs at very low mean daily air temperatures (at all stations < -10 °C). Such low temperatures are accompanied by a very low wind speed (except for plateaus) and a long windless period. In weak wind conditions, there is also a dominant inflow from the southern sector. Among other weather patterns, this one is characterised by the lowest relative air humidity, the highest insolation, no precipitation or precipitation of a small daily sum thereof, as well as 24-hour inversion of temperature. The weather type no. II episodes that account for the highest number of dust episodes, but with the lowest mean regional concentration values occur at slightly negative air temperatures with relatively high wind velocities considering the dust episode conditions and the lowest number of silences among all the differentiated weather patterns. Dominant here is a clear influx of air masses from the north and northeast, with high humidity and low insolation. A slightly positive vertical pseudogradient of temperature indicates a slight inversion of temperature during these episodes. Episodes of this type feature atmospheric precipitation with relative frequency. The weather type no. III, which accounts for less than 30% of the dust episodes, is characterised by a mean daily air temperature of about -4.0°C, the highest wind speeds among all other types of weather (excluding the plateaus), the influx of air masses from the east, and a relatively high total radiation with no precipitation. Its characteristic feature is isothermy. As can be seen from the results shown in Table 1, with the weather type no. I, the probability of occurrence of an episode of an at least regional scale is almost total. In

type no. II, the probability of occurrence of an episode amounts to about 60%. In turn, in type no. III, the probability is estimated at 75%.

3 Summary

Areas which have been anthropogenically strongly transformed and thus those which are marked with dominance of industrial and urban infrastructure are particularly vulnerable to the occurrence of high concentrations of particulate matter base. The identification and subsequent testing of the development phases of such episodes had been hitherto performed using classical synoptic analysis methods or descriptive statistics. In contrast, this paper proposes a comprehensive approach to the subject - the use of data mining methods, which, according to the authors, allows for a more complete indication of a group of meteorological factors which influence high concentrations of particulate matter. The applied methodology helped delineate several groups of meteorological situations responsible for the formation of episodes of high concentrations of PM10. It seems that this methodology - despite the fact that it does not simply refer to the origin of such situations - does allow for an objective identification (i.e. such that is independent of the researcher's intuition) of a set of meteorological factors that are conducive to their engendering. This method may therefore constitute an important auxiliary tool in both short-term air quality forecasting, and - what is perhaps even more important – it may create an element of an air quality management system in areas of particular vulnerability.

The scientific study was elaborated within the framework of the project entitled "Air Quality Information System in the Polish - Czech Border Area in the Region in Silesia and Moravia", which was founded by the Operational Program for Cross - Border Cooperation between the Czech Republic and the Republic of Poland (POWT RCz – RP 2007-2013) (registration number CZ.3.22/1.2.00/09.01610) and co-financed by the ERDF (European Regional Development Fund).

References

1. EEA Report, Air quality in Europe –2017 report. No 13/2017
2. WHO, Ambient air pollution: A global assessment of exposure and burden of disease. 2016
3. J. He, S. Gong, Y. Yu, L. Yu, H. Mao, C. Song, S. Zhao, H. Liu, X. Li, R. Li, *Air pollution characteristics and their relation to meteorological conditions during 2014–2015 in major Chinese cities.* Environ. Pollut., **223**, (2017)
4. K. Juda-Rezler, M. Reizer, M. and J.P Oudinet, *Determination and analysis of PM10 source apportionment during episodes of air pollution in Central Eastern European urban areas: The case of wintertime 2006*. Atmos. Environ. **45,** 36, (2011)
5. Central Statistical Office (GUS), Area and Population in the Territorial Profile in 2017. Statistical Information and Elaborations. Warszawa 2017, http://stat.gov.pl/en
6. Czech Statistical Office (CZSO), https://www.czso.cz (accessed 15.11.2017)
7. J. Demek, P. Mackovčin P. (eds.) et. al., Hory a nížiny - zeměpisný lexikon ČR. Agentura ochrany přírody a krajiny ČR - AOPK ČR (2006)
8. J. Kondracki, *Geografia regionalna Polski*. WN PWN Wyd.3, Warszawa 2011
9. A. Łachwa, *Rozmyty świat zbiorów, liczb, relacji, faktów, reguł i decyzji.* AOW Exit, Warszawa 2001
10. Osowski, *Sieci neuronowe*. WNT, Warszawa 1997