Brigham Young University
BYU ScholarsArchive

International Congress on Environmental Modelling and Software

4th International Congress on Environmental Modelling and Software - Barcelona, Catalonia, Spain - July 2008

Jul 1st, 12:00 AM

A data mining approach to discover weather patterns contributing to PM10 exceedances

A. Sfetsos

D. Vlachogiannis

Follow this and additional works at: https://scholarsarchive.byu.edu/iemssconference

Sfetsos, A. and Vlachogiannis, D., "A data mining approach to discover weather patterns contributing to PM10 exceedances" (2008). *International Congress on Environmental Modelling and Software*. 155.
https://scholarsarchive.byu.edu/iemssconference/2008/all/155

This Event is brought to you for free and open access by the Civil and Environmental Engineering at BYU ScholarsArchive. It has been accepted for inclusion in International Congress on Environmental Modelling and Software by an authorized administrator of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.

iEMSs 2008: International Congress on Environmental Modelling and Software
Integrating Sciences and Information Technology for Environmental Assessment and Decision Making
4th Biennial Meeting of iEMSs, http://www.iemss.org/iemss2008/index.php?n=Main.Proceedings
M. Sànchez-Marrè, J. Béjar, J. Comas, A. Rizzoli and G. Guariso (Eds.)
International Environmental Modelling and Software Society (iEMSs), 2008

A data mining approach to discover weather patterns contributing to PM10 exceedances

Sfetsos A. [(a)], **Vlachogiannis D.** [(a)]

(a) Environmental Research Laboratory,
Institute of Nuclear Technology - Radiation Protection, NCSR "DEMOKRITOS",
15310 Aghia Paraskevi Attikis, Athens, Greece
(ts, mandy@ipta.demokritos.gr)

Abstract: The present study presents a data mining methodology to discover weather characteristics that are associated with exceedances of particulate matter (PM) with diameter less than 10 μm. The proposed approach is an eigenvector one consisting of a rotated principal components algorithm coupled with the fuzzy c-means clustering algorithm. This study examines only those days where at least one station exhibited an exceedance of the limit of a daily average value of 50 $\mu g/m^3$. The analysis was conducted using daily data from the monitoring network of the Grater Athens Area that were complemented with meteorological data from the NCEP Global Forecasting System (GFS). The analysis concluded that the PM10 exceedances in the Athens Area can be classified into 10 distinct types with varying spatial characteristics and weather contribution.

Keywords: data mining; climatology; PM10; exceedances.

1. INTRODUCTION

According to recent EU position papers (EC Directive 99/30/EC) PM is one of the top priority air pollutants that adversely influence human health (e.g. Kappos et al. [2004]). Climatological approaches have been reported with considerable success in relating atmospheric conditions to air pollution concentrations (e.g. Davis and Gay [1993]; Comrie [1994]. Eigenvector approaches have been applied in similar application that aim to identify group of days with similar characteristics, relating weather types to air quality patterns, such as those of Kasomenos et al. [1998] for the Attica basin, for Moscow (Shahgedanova et al. [1998]). Greene et al. [1999] developed the "Temporal Synoptic Index" to categorize weather in four US cities and related the resulting weather types to air pollution concentrations and human morbidity.

The Greater Athens Area (GAA) has a population of over 4.5 million inhabitants. The main pollution problem is attributed to the heavy traffic conditions (over 2.5 million registered vehicles and steadily decreasing infrastructures) taking into account that a large number of these vehicles are non-catalytic or old technology diesel engines (Grivas et al. [2008]). Space heating is the primarily done by fuel oil and only recently natural gas has been introduced in the energy mix with low penetration levels. Three main industrial zones exist in the region, which are located to the West (Thriassion plain including two oil refineries, cement plants and a steelworks factory), S–SE (electricity generating unit) and the North (secondary industrial operations and electricity production). According to a recent study by Mirasgedis et al. [2008], the external costs associated with industrial activities amount to 211M€ per year, representing 0.5% of the GDP and 4.1% of the Gross Value Added by the industrial activities in the area of interest in 2000 and are mainly associated with human mortality and morbidity primarily due to PM10 emissions.

This study presents the application of an eigenvector methodology for the identification of weather characteristics and the spatial analysis of exceedances of PM10 occurring in at least one monitoring location in the Greater Athens Area. According to Yarnal et al. (2001) eigenvector-based approaches consists of two phases. Once the available data are selected, a dimension reduction algorithm is applied so that only useful information from the original data set is conveyed. Then a clustering algorithm is applied to determine groups of data that have common variational properties.

2. METHODOLOGY DESCRIPTION

In the present study, the selected variables were transformed to have zero mean and untiy variance and thus common scaling. The scaling factors were determined using only those values that are above the regulatory limit, since these constitute the focus of the study and their potential impact is far more stressed.

The dimension reduction is achieved using a varimax rotated Principal Components Analysis (rPCA), where the number of new independent components is determined by setting a cut-off threshold in the estimated eigenvalues. In this study the *eigenvalue greater than one rule* suggested in Shahgedanova et al. [1998], was applied in order to determine the number of variables used. Several other alternatives were tested such as factor analysis and the positive matrix factorization, but did not return significant differences in the number of independent components nor their properties.

The selection of the clustering algorithm is very important in related studies, as it effectively identifies weather types. Following the work of Kalkstein et al. [1987] the average linkage clustering was favoured, although in recent years many authors shifted to the *k*-means algorithm and more recently to variations of the subtractive clustering algorithm. In the present study the fuzzy c-means algorithm (e.g. Besdek [1981]) was applied and the number of clusters was found using the "compactness and separation" criterion (Kim et al. [2001]).

3. AREA OF APPLICATION AND DATASETS USED

The weather conditions in GAA are mainly the results of interaction of large and local scale circulation systems. The climate of the GAA is typically Mediterranean with hot dry summers and wet mild winters. According to Ministry of Environment annual reports particulate matter is the only pollutant that exhibits persistent exceedances in GAA. The table below shows the annual mean concentration of PM10 in the GAA monitoring locations (MinEnv 2007).

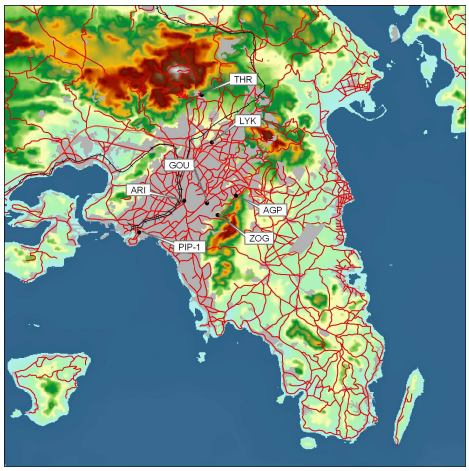

Figure 1. Monitoring Locations and GFS grid points.

Table 1. PM10 Annual Mean Concentration in μg/m^3 (MinEnv 2007)

YEAR	PIR-1	MAR	ZOG	LYK	AGP	ARI	GOU	THRA
2001	57	55	35	60	47	55	50	31
2002	63	69	35	62	38	55	53	34
2003	54	38	34	59	37	56	22	32
2004	56	29	33	63	39	58	18	33
2005	54	46	29	53	41	53		32
2006		48	26	59	34	57	34	27

Data for a six-year period (from 2001–2006) were used in this study. The air quality data stations, portrayed in Figure 1, are operating on a continuous basis under the supervision of the Greek National Air Pollution Monitoring Network. Only data from the stations of (i) Aristotelous (ARI), (ii) Agia Paraskevi (AGP), (iii) Lykovrisi (LYK), (iv) Thrakomakedones (THR) and (v) Zografou (ZOG) were used because they contained an adequate number of samples to conduct the analysis. The meteorological data were obtained from the NCEP reanalysis data set (http://dss.ucar.edu/datasets/ds083.2/) over the respective period of time on a daily basis at 00:00 and 12:00 hours. These were selected to have a general overview of the weather patterns and midday is period were highest values are observed at the suburban stations (MinEnv 2007). The selected variables are presented in Table 2 and their correlation coefficients in Table 3 are: the mixing layer height (MLH) which is crucially associated with dispersion of air pollutants, temperature (T) at 2m and the E-W (U) and N-S (V) components of wind speed

Table 2. Selected variables, and statistical properties

		Mean	STD	Units
00:00	***T***	290.2504	5.7588	K
	MLH	507.2484	384.8502	M
	U	1.3705	2.3071	m/s
	V	-1.6306	3.421	m/s
12:00	**T**	293.9041	5.9127	K
	MLH	1013.012	435.6832	M
	U	-1.3874	2.2443	m/s
	V	-0.5165	4.7	m/s
Daily	**ARI**	64.9458	25.9312	μg/m^3
	AGP	49.2586	27.4576	μg/m^3
	LYK	70.9209	27.3346	μg/m^3
	THR	36.2391	28.4849	μg/m^3
	ZOG	39.137	23.5763	μg/m^3

4. RESULTS

The application of the r-PCA approach returned 4 new variables (each named as PCi, i=1…4) that explained approximately 77.64 % of the variance of the initial data. The variables that had a significant contribution to each loading are described in Table 4. It may be observed that the resulting loadings are organized into well related variables at every interval on the day. PC1(define) is dominated by the PM. PC2 is related to the T parameters at both time intervals, whereas PC3 is dominated by the N-S and the E-W wind components. Finally, PC4 is dominated by the MLH parameters at both intervals and the E-W (*U*) component at 12:00. Table 3 provides an explanation of the grouping of similar variables into a new loading since they exhibit the highest correlation, in excess of 0.7. The correlation coefficient of the resulting factors is below 0.2 for every possible combination of the resulting four variables.

Table 3. Correlation Coefficient of examined variables

	T	**MLH**	**U**	**V**	**T**	**MLH**	**U**	**V**	**ARI**	**AGP**	**LYK**	**THR**	**ZOG**
00:00													
T	1.00	-0.19	-0.01	-0.17	0.94	-0.04	-0.30	-0.19	-0.15	0.19	-0.12	0.07	0.15
MLH	-0.19	1.00	-0.16	-0.31	-0.31	0.47	0.20	-0.38	-0.18	-0.11	-0.01	-0.07	-0.05
U	-0.01	-0.16	1.00	0.31	0.08	0.03	0.31	0.28	-0.06	-0.12	0.00	-0.02	-0.17
V	-0.17	-0.31	0.31	1.00	-0.07	0.00	0.34	0.72	0.28	0.16	0.24	0.20	0.12
12:00													
T	0.94	-0.31	0.08	-0.07	1.00	-0.01	-0.27	-0.04	-0.15	0.21	-0.09	0.10	0.14
MLH	-0.04	0.47	0.03	0.00	-0.01	1.00	0.32	-0.07	-0.18	-0.03	-0.05	0.00	-0.02
U	-0.30	0.20	0.31	0.34	-0.27	0.32	1.00	0.28	0.07	-0.12	0.17	-0.02	-0.12
V	-0.19	-0.38	0.28	0.72	-0.04	-0.07	0.28	1.00	0.29	0.17	0.27	0.27	0.13
Daily													
ARI	-0.15	-0.18	-0.06	0.28	-0.15	-0.18	0.07	0.29	1.00	0.58	0.79	0.58	0.62
AGP	0.19	-0.11	-0.12	0.16	0.21	-0.03	-0.12	0.17	0.58	1.00	0.66	0.73	0.87
LYK	-0.12	-0.01	0.00	0.24	-0.09	-0.05	0.17	0.27	0.79	0.66	1.00	0.65	0.73
THR	0.07	-0.07	-0.02	0.20	0.10	0.00	-0.02	0.27	0.58	0.73	0.65	1.00	0.79
ZOG	0.15	-0.05	-0.17	0.12	0.14	-0.02	-0.12	0.13	0.62	0.87	0.73	0.79	1.00

Table 4. Rotated Factor Loadings

		PC1	**PC2**	**PC3**	**PC4**
00:00	T	0.0052	-0.654	-0.0542	-0.0184
	MLH	-0.0311	0.1418	-0.2626	-0.5576
	U	0.1618	-0.1205	0.4638	-0.103
	V	-0.0707	0.0236	0.5402	0.0435
12:00	T	0.0117	-0.6731	0.0458	-0.0017
	MLH	0.0011	-0.1238	0.0487	-0.6791
	U	0.0134	0.1244	0.3688	-0.4241
	V	-0.0823	0.0361	0.522	0.1217
Daily PM10	ARI	-0.521	0.1104	0.0372	0.1061
	AGP	-0.4509	-0.1659	-0.0428	-0.0241
	LYK	-0.5673	0.0373	0.0415	-0.0837
	THR	-0.2281	-0.0555	0.0118	-0.0267
	ZOG	-0.3348	-0.0882	-0.0566	-0.0294

The number of clusters that were found using the compactness and separation criterion was equal to 10 (Figure 2). Then depending on the resulting cluster centers, all data were attributed to the nearest center using the Euclidean distance. The resulting percentage appearance of each resulting cluster is estimated and displayed on Figure 3.

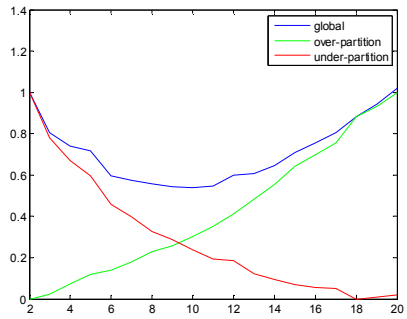

Figure 2. Number of clusters

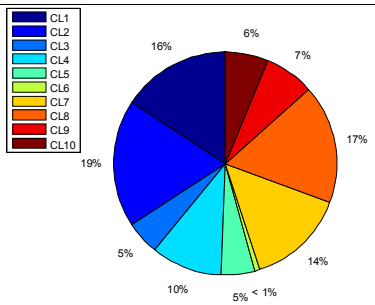

Figure 3. Data distribution in clusters

The resulting clusters are presented in Table 5. The Table shows the corresponding values of the examined variables of each cluster and the nearest real date from the modeling data. These dates are only indicative and they closely, but not entirely, match the observed conditions, and they are intended to be used in future detailed studies with numerical weather and atmospheric pollution models.

Table 5. Cluster characteristics

	00:00				12:00			
	T	**MLH**	**u**	**v**	**T**	**MLH**	**u**	**v**
	C	m	m/s	m/s	C	m	m/s	m/s
CL1	22.95	451.45	0.96	-3.60	26.42	988.04	-2.83	-3.17
CL2	21.82	233.49	2.40	-0.17	26.27	931.95	-1.67	1.57
CL3	16.99	856.40	4.28	1.50	21.57	1825.04	2.68	2.67
CL4	11.17	980.44	-0.80	-5.34	13.25	1134.71	-1.64	-5.72
CL5	16.82	361.70	1.51	1.44	21.12	967.05	-0.84	3.93
CL6	16.18	963.77	-2.79	1.48	19.29	1434.79	-0.85	4.24
CL7	10.78	422.38	1.95	0.53	14.74	859.71	-0.31	2.55
CL8	15.66	273.84	0.85	-1.60	19.29	627.40	-2.51	-0.01
CL9	22.02	1082.28	-0.59	-7.28	24.16	1508.53	-2.40	-8.76
CL10	10.55	743.61	3.18	0.97	14.64	1385.54	1.81	2.44

Table 5. Continued

	ARI	**AGP**	**LYK**	**THR**	**ZOG**	**"Matching Day"**
	$\mu g/m^3$	$\mu g/m^3$	$\mu g/m^3$	$\mu g/m^3$	$\mu g/m^3$	
CL1	51.38	48.39	57.32	30.57	37.07	18-Jun-02
CL2	65.25	57.15	71.25	42.71	43.40	31-Aug-04
CL3	53.42	42.52	71.98	36.74	32.22	15-Nov-04
CL4	58.31	34.62	63.99	21.77	30.63	9-Mar-03
CL5	107.98	96.09	119.23	84.89	80.46	21-Apr-01
CL6	228.84	248.52	267.71	236.54	221.95	17-Apr-05
CL7	73.06	39.55	74.68	31.09	31.26	29-Nov-04
CL8	69.90	46.32	68.95	32.18	36.44	12-Nov-04
CL9	43.03	48.83	57.68	30.46	40.94	12-Jun-03
CL10	57.67	27.38	66.43	22.32	20.63	10-Jan-01

The following paragraph shows a detailed analysis of the resulting clusters

CL1: Corresponds to warm conditions with low to moderate winds that change from NE to NW. High pollution levels are located in the city center (ARI) and LYK, which is located near the main highway, that are mainly from primary emissions.

CL2: Corresponds to sea breeze conditions which are frequently occurring during summertime. Low morning winds change into south components. High values are found in the city center (primary emissions) and in the stations nearest the mountains (THR, AGP) from transportation.

CL3: Corresponds to clear days (high MLH during midday) and moderate winds from SE. The peak is located near LYK, which is a combination of primary emissions and secondary from transportation from the city center.

CL4: Corresponds to cold days with predominant strong NW winds and relatively high MLH. The higher pollution values are found in LYK and ARI, which is mainly associated with primary production of PM10 from diverse sources.

CL5: Corresponding to very high PM10 levels in all stations, approximating a value between 80 and 100 $\mu g/m^3$ for all stations in the GAA. The other important is the dominant south wind component during all hours.

CL6: Corresponds to extremely high PM10 concentrations which occurs only a very few number of times (percentage < 1%). However pollution values in all stations is in the area of 200 $\mu g/m^3$.

CL7: Corresponds to cold days that are partly or totally clouded (low MLH values during all day). Also stationary conditions (low winds) with a mildly increasing SW component in the direction. The hotspots in the GAA are the city center (ARI) and LYK.

CL8: Corresponds to mild days that are partially cloudy or overcastted. The wind is low and blowing from NE in the morning changing to W. The associated with primary emissions are above the limit of 50 $\mu g/m^3$, but also AGP has values near that due to transportation.

CL9: Corresponds to warm days with clear skies (high MLH). The wind has strong component from the North, which is typical weather during summer days (etesian winds). The peak is found in the eastbound part of city (AGP, ZOG) and LYK.

CL10: Corresponds to cold and clear days with high MLH values. The winds are low and have an east component in addition to an increasing south one. The locations of primary emissions (ARI, LYK) are the only locations with high concentrations.

5. CONCLUSIONS

This study presented a data mining approach for analyzing the correspondence between exceedances of the national regulatory limit of PM10 and weather conditions in the Greater Athens Area. A rotated PCA approach was used coupled with the fuzzy c-means algorithm to discover closely matching patterns of air pollution and weather conditions, in addition to spatial diversifications between stations. The analysis revealed the presence of ten distinct weather / air pollution patterns.

REFERENCES

Bezdec, J.C., *Pattern Recognition with Fuzzy Objective Function Algorithms*, Plenum Press, New York, 1981

Comrie, A.C., A synoptic climatology of rural ozone pollution at three forest sites in Pennsylvania, *Atmospheric Environment*, 28 (9), 1601-1614, 1994.

Davis, E.R., and D.A. Gay, A synoptic climatological analysis of air quality in the Grand Canyon national park. *Atmospheric Environment,* 27 (5), 713-727, 1993.

Grivas, G., A. Chaloulakou, and P. Kassomenos, An overview of the PM10 pollution problem, in the Metropolitan Area of Athens, Greece. Assessment of controlling factors and potential impact of long range transport, *Science of the Total Environ*ment, 389(1), 165-77, 2008.

Greene, J.S., L.S. Kalkstein, H. Ye, and K. Smoyer, Relationships between synoptic climatology and atmospheric pollution at 4 US cities. *Theoretical Applied Climatology*, 62, 163-174, 1999.

Kalkstein, L.S., G. Tan, and J.A. Skindlov, A comparison of three clustering procedures for use in synoptic climatological classification. *Journal of Climatology & Applied Meteorol*ogy, 19, 717-730, 1987.

Kappos, A.D., P. Bruckmann, T. Eikmann, N. Englert, U. Heinrich, P. Höppe, E. Koch, G.H.M. Krause, W.G. Kreyling, K. Rauchfuss, P. Rombout, V. Schulz-Klemp, W.R. Thiel, and H.E. Wichmann, Health effects of particles in ambient air, *International Journal of Hygiene and Environmental Health*, 207 (4), 399-407, 2004.

Kassomenos, P.A., H.A. Flocas, S. Lykoudis and M. Petrakis, Analysis of mesoscale patterns in relation to synoptic conditions over an urban Mediterranean basin, *Theoretical and Applied Climatology,* 59, 215-229, 1998.

Kim, D.J., Y.W. Park, and D.J. Park, A novel validity index for determination of the optimal number of clusters. *Institute of Electronics, Information and Communication Engineers Transactions of Information Systems*, E84-D, 281-285, 2001.

MinEnv 2007, The Air Pollution in Athens 2006 Annual Report, published by Ministry of Environment, March 2007, (in Greek).

Mirasgedis, S., et al. Environmental damage costs from airborne pollution of industrial activities in the greater Athens, Greece area and the resulting benefits from the introduction of BAT, *Environmental Impact Assessment Review*, 28 (1), 39-56, 2008.

Shahgedanova, M., T.P. Burt, and T.D. Davies, Synoptic climatology of air pollution in Moscow, *Theoretical Applied Climatology,* 61, 85-102, 1998.

Yarnal, B., A.C. Comrie, B.J. Frakes, and D.P. Brown, Developments and prospects in synoptic climatology, *International Journal of Climatology*, 21, 1923-1950, 2001.