COMPARISON OF HIERARCHICAL, NON-HIERARCHICAL AND NEURAL NETWORK CLUSTERING TECHNIQUES FOR THE CLASSIFICATION OF ATMOSPHERIC BACK TRAJECTORIES

P. Kassomenos[1], C. Papaloukas[2] , S. Vardoulakis[3], R. Borge[4], J. Lumbreras[4] ,S. Karakitsio[s2]
[1]University of Ioannina, Dept. of Physics, Lab of Meteorology, GR-45110, Ioannina, Greece.
[2]Univ. of Ioannina, Dept. of Biological Applications and Technology, GR-45110, Greece.
[3]Public and Environmental Health Research Unit, London School of Hygiene and Tropical Medicine, University of London, Keppel Street, London, WCIE 7HT, UK.
[4]Department of Chemical and Environmental Engineering, Technical University of Madrid, (UPM), Jose Gutierrez, Abascal 2, 28006, Madrid, Spain.

INTRODUCTION

Clustering techniques for the classification of air mass trajectories used in the past varied widely, based on different hierarchical and non-hierarchical approaches. Recently, Artificial Neural Networks have gained interest and are increasingly recognized as a useful statistical technique for the prediction and classification of both environmental and meteorological data. In this paper, we firstly introduce a method to define the appropriate number of clusters for the classification of atmospheric trajectories. Using the defined number of clusters we compare a hierarchical, a non-hierarchical clustering technique (K-means algorithm), and two Neural Network's Self Organizing Maps (SOM) to classify back trajectories.

Further to the weather types occurring in an area, it is important from an environmental point of view to know the source and path of air masses reaching this area. This can be achieved by classifying back trajectories into clusters. In order to analyse the influence of transport patterns on pollutant concentrations in the atmosphere, several multivariate techniques including statistical clustering methods can be applied to modelled back trajectories (*Dorling and Davis*, 1995; Borge et al., 2007). The time varying coordinates of the back trajectories can be used as the clustering variables, leading to the identification of distinct groups with similar characteristics, i.e. similar behaviour of their direction of approach and speed of passage over potential pollution source areas. The trajectory types are more readily interpretable in terms of the synoptic conditions that form them. Large scale circulation features are associated with certain trajectory clusters. Furthermore, trajectory clustering schemes are increasingly used to identify links between origin/path of an air mass and air quality.

In recent years, artificial neural network and fuzzy logic techniques have gained interest and are increasingly recognised as promising techniques for the prediction and classification of not only environmental but also meteorological data (*Hewitson and Crane*, 2002). The aim of this study is to: (a) Introduce a method for defining the appropriate number of atmospheric trajectory clusters, (b) apply several clustering techniques, examine their performance and compare the resulted back trajectory groups, (c) interpret the variability of daily PM_{10} averages recorded at three monitoring stations in Athens using the obtained trajectory clusters, and discuss the implications for air quality management.

DATA AND METHODOLOGY

5-day long kinematic back trajectories arriving in Athens, Greece (37.2 latitude, 23.47 longtitude), every day at 12.00 UTC during a four-year period (2001-2004) were used. These trajectories were calculated with version 4 of the model HYSPLIT developed by NOAA Air Resources Laboratory (*Rolph*, 2003). The back trajectories were computed at 500 m above

ground. In addition, daily mean PM_{10} concentrations estimated from hourly recorded values at three air quality monitoring stations in Athens were used. In this study we employed a variation of a graph-based method (*Salvador and Chan*, 2004), in order to define the appropriate number of back trajectory clusters. Three different clustering approaches, namely Hierarchical clustering, non-hierarchical K-means algorithm and Self-Organizing Maps, were used in order to classify individual back trajectories into groups.

RESULTS AND DISCUSSION

Following the above methodology, the appropriate number of clusters was set to six. The description of the origin and path of each one of the six clusters is presented below:

Group A. It has its origin either over Sahara desert or over the Gulf of Sidra or Tunis or the maritime area between Africa, Sicily and Peloponnesus. It is a rather slow moving mean centroid reaching Athens from southern directions when it passes over the sea carries particulates either from Sahara or from the sea (salt particulates) or both. In some cases it will be the result of local circulation or recirculation around Athens.

Group B. Initially the air mass is over the wider area of Western Ukraine, South Poland, Slovakia, Austria and Hungary. Then it crosses Eastern Balkans (in some cases western Balkans) arriving in Athens after its passage through Northern Greece or the Aegean Sea. It is a rather slow moving regime.

Group C. The origin of the air mass is over Russia. The air mass is moving towards the areas of Crimea peninsula and Black sea, passes over eastern Thrace, North Aegean Sea reaching Athens from north-eastern directions.

Group D. This regime is a fast moving one. It takes out from north Germany, Poland or Scandinavia and it crosses central-west Europe arriving in Athens after its passage over the Balkans.

Group E. It is the faster moving regime. Its origin is over mid-Atlantic. It passes over British isles, France, Northern Italy and the Adriatic Sea arriving in Athens from north west.

Group F. The origin of the air mass is over the Pyrenees Mountains (Gulf of Lion) or Western Mediterranean, it crosses South Italy reaching Athens from the west. It is a rather slow moving mean back trajectory.

In general, at 500m height, the Hierarchical and K-means clustering approaches present quite similar results (Fig. 1). The only difference found was the origin of Group B which was over Austria in K-means, but over Ukraine in the Hierarchical approach. The two SOM classifications (for 1X6 and 2X3 dimensions) produced nearly identical results and quite similar with those of the Hierarchical approach. The results estimated with different clustering approaches show that there is larger variability between Hierarchical and the other two approaches than between K-means and SOM. Specifically the Pearson correlation coefficient (R) ranged between 0.429-0.369 for the pairs of Hierarchical with the other three approaches, while the respective correlation coefficients between K-means and SOM ranged between 0.577-0.548. Also the variability of the percentage of days attributed to each cluster was very high between Hierarchical and the other three techniques, but smaller between K-means and SOM. Groups A, D and F were found to be more stable.

Both SOM approaches, did not present differences in the distribution of days in different trajectory clusters. The higher differences were detected in Group F in SOM 2X3. R ranged between 0.588 and 0.703 for SOM 2X3 and between 0.607 and 0.703 for SOM 1X6. It is interesting to compare the results of the two SOM alternative procedures. It was found that the two SOM techniques (with dimensions 2X3 and 1X6) produced very similar results (R = 0.852).

Since the focus of this work was to reveal the impact of different atmospheric transport regimes on air quality in Athens, an examination of daily PM_{10} concentrations corresponding to different back-trajectories clusters was carried out.

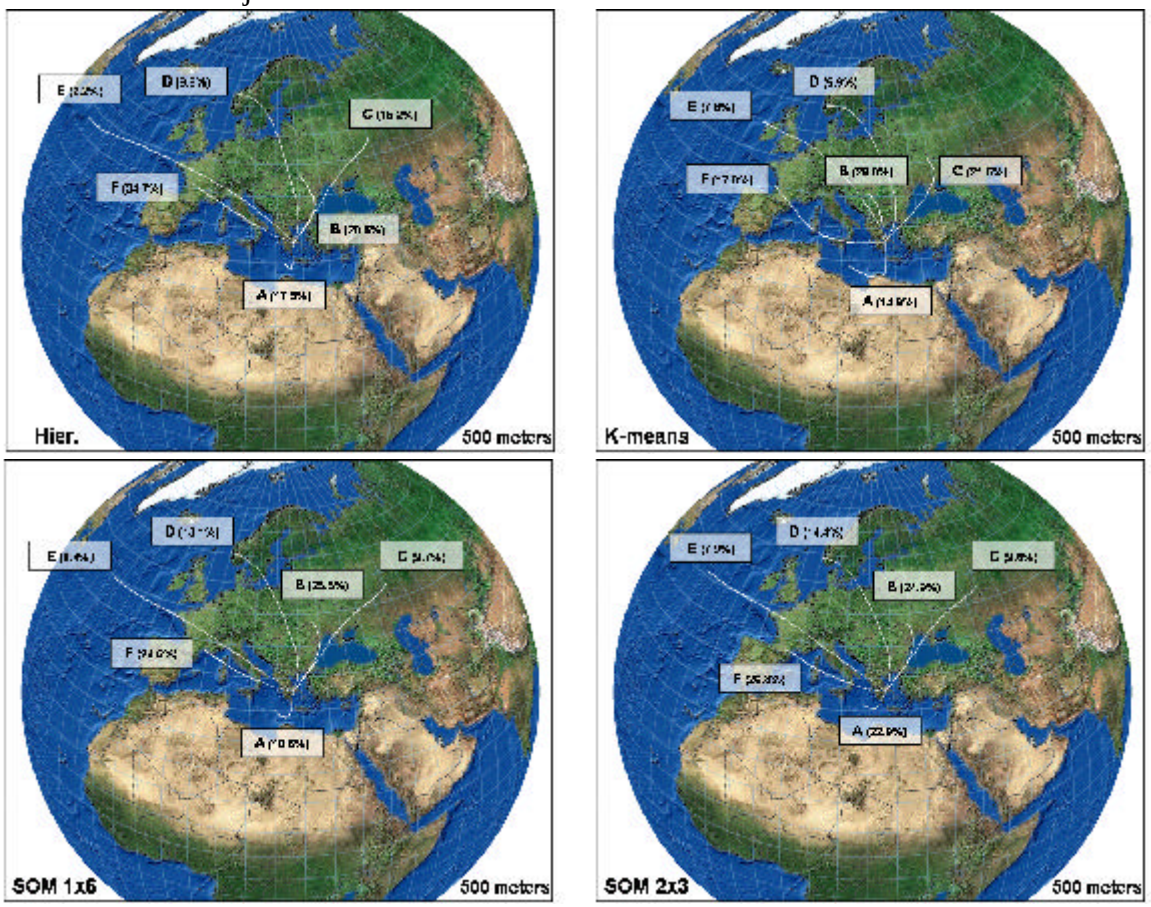

Fig. 1; Centroids of cluster analysis at 500m for (a) Hierarchical, (b) K-means (c) SOM 1X6 and (d) SOM 2X3 approach. Percentage of occurrence for each cluster is shown in ( ).

Table 1. Mean PM_{10} concentrations (μgr/m^3) per cluster mean trajectory arriving in Athens at 500m for Hierarchical, K-means, SOM 1X6 and SOM 2X3 at three air quality stations .

Transport Regime	Hierarchical			K-means			SOM 1X6			SOM 2X3		
at 500-m	THR	LYK	PEI	THR	LYK	PEI	THR	LYK	PEI	THR	LYK	PEI
A	**39.2**	66.0	**65.2**	**42.8**	69.0	**66.0**	**40.8**	67.9	61.2	**41.9**	**69.9**	**66.0**
B	29.0	55.4	58.8	32.6	64.1	62.9	*24.3*	70.0	**66.7**	29.6	55.4	56.9
C	26.0	*46.8*	51.5	*26.7*	*48.8*	53.0	27.8	51.0	*52.2*	27.6	*50.2*	53.9
D	*22.6*	52.9	*49.7*	28.7	48.9	*51.4*	29.4	*50.2*	54.0	*24.9*	52.1	*52.4*
E	36.7	**76.1**	54.8	24.5	63.9	54.8	30.5	55.5	56.4	29.0	68.2	55.5
F	38.6	69.5	61.7	42.4	**71.5**	58.8	40.3	**70.3**	55.6	39.7	68.7	60.6

Three PM_{10} monitoring stations were selected. The first of semi rural character is located in the northern western periphery of the city (THR). Since in the vicinity of this station there are no significant source of particulates, it is assumed that it can give indications of possible long range transport. Additionally, one suburban background (LYK) and one traffic oriented (PEI) stations were selected to check for potential contributions from long range transport to the observed PM_{10} concentrations.

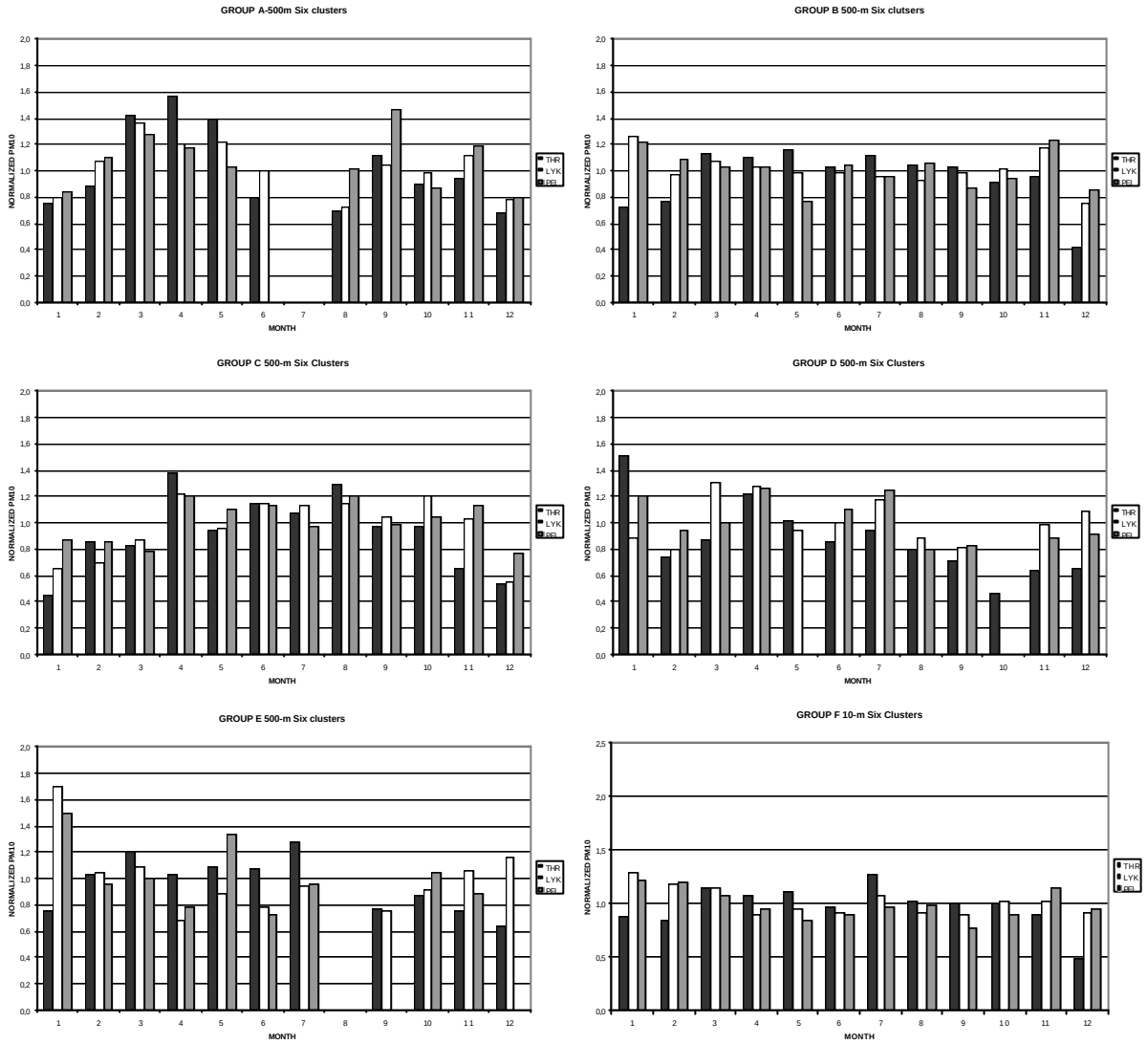

Fig. 2; Monthly distribution of the normalized PM_{10} values per Group computed with the K-means approach at 500m above ground.

The mean PM_{10} concentrations in the semi rural station THR are higher on days attributed to Group A mean trajectory centroids in all four clustering approaches (Table 1). These high values indicated possible long range transport of particulates either from Sahara desert or from the Mediterranean Sea or both. The second higher values are associated with Group F mean trajectory centroids. Group F describes back trajectories having their origin over the western Mediterranean and arriving in Athens from the west. The lowest PM_{10} values are associated with Groups C and D mean back trajectories. The origin of the trajectories belonging to these two groups is either Russia or Central Europe.

The mean PM_{10} concentrations in the suburban background station LYK showed that in the majority of the cases Group A is associated with the highest or second highest daily values. The lowest values are associated again with Groups C and D (Table 1).

Concerning the traffic oriented station PEI in all the approaches the higher PM_{10} values are associated with Group A (with the exception of SOM 1X6 in which group A is associated with the second highest values). The second highest values of PM_{10} are associated with Group F. Finally, the lowest values are associated with Groups C and D.

We also compared the variation of average PM_{10} concentrations between corresponding groups obtained using the different clustering approaches, but we did not found significant differences. The atmospheric transport regime described by Group A back trajectories is associated with the highest values of PM_{10} and produces consistent results in almost all three clustering approaches and stations. It is supposed that a significant transportation of particulates could also happen with Group F back trajectories.

Monthly distribution of the recorded PM_{10} concentrations (Figure 2) showed that for Group A in all clustering approaches the highest normalized PM_{10} values for THR were detected during spring months (while for the other two stations, PEI and LYK, a lower increase was also detected). Previous studies have also indicated the higher frequency of this regime during spring months (Reference?). The transport regime representing by Group C (etesian winds) shows higher normalized PM_{10} values during June, July, August and April. A lower increase in PM_{10} is also detected for PEI and LYK stations during the same months. Group D and Group F have their peak during January and July, respectively.

CONCLUSIONS

From the above analysis the following conclusions could be drawn:

- The Hierarchical approach seems very sensitive to fast and slow moving clusters, thus results should be interpreted with caution.
- K-means produces more consistent results with SOM, although the variability of the percentages of occurrence is also small.
- Neural network SOM methods performed significantly better than the Hierarchical clustering. SOM 1X6 seems to have slightly better performance than SOM 2X3.
- Air masses having their origin over Sahara Desert and western Mediterranean Sea are associated with high values of mean PM_{10} in all cases. Air masses with such origins could be characterized as slow moving.
- On the contrary, fast moving air masses having their origin over Russia, Ukraine or Central and North Europe are associated with the lowest daily mean PM_{10} in Athens.

ACKNOWLEDGEMENTS:

The authors gratefully acknowledge the NOAA Air Resources Laboratory (ARL) for the provision of the FNL-HYSPLIT data, the HYSPLIT transport and dispersion model, and the READY web site (http://www.alr.noaa.gov/ready.htm) used in this study.

REFERENCES

Borge R., Lumbreras J., Vardoulakis S., Kassomenos P., Rodríguez E., 2007. Analysis of long-range transport influences on urban PM10 using two-stage atmospheric trajectory clusters. Atmospheric Environment (In Press)

Dorling S. R. and T. D. Davis, 1995. Extending cluster analysis-synoptic meteorology links to characterize chemical climates at six northwest European monitoring stations. Atmospheric Environment, 29, 145-167.

Hewitson B.C. and R.G. Crane, 2002. Self-Organizing maps: applications to synoptic climatology. Climate Research, 22, 13-26.

Rolph, G.D., 2003. READY: Real time Environmental applications and Display system. NOAA Air resources Laboratory. (hhtp://www.arl.noaa.gov/ready.html).

Salvador S. and P. Chan, 2004: Determining the number of clusters/segments in Hierarchical clustering/segmentation algorithms. 16th IEEE International Conference on Tools with Artificial Intelligence (ICTAI), 576-584.