20th International Conference on
Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes
14-18 June 2020, Tartu, Estonia

A NOVEL MODELLING-BASED METHOD FOR AIR QUALITY ZONING. APPLICATION TO THE MADRID REGION

Rafael Borge [1], Daeun Jung[1], Iciar Lejarraga[1], Enrique Crespo[2], Ricardo Vargas[2], David de la Paz[1], Jose María Cordero[1], Jaime S. Gallego[2]

[1]Environmental Modelling Laboratory, Department of Chemical & Environmental Engineering, Universidad Politécnica de Madrid (UPM), José Gutiérrez Abascal 2, 28006. Madrid, Spain
[2]Consejería de Medio Ambiente, Ordenación del Territorio y Sostenibilidad, Área de Calidad Atmosférica, C/ Alcalá, 16, 28014, Madrid, Spain

Abstract: According to the European Air Quality Directive (AQD) (2008/50/EU), Member States should establish and periodically review zones and agglomerations to control and manage air quality throughout their territory. Madrid currently considers seven zones of air quality based on administrative, geographic and land use criteria. This zoning was proposed in 2014 and needs to be reviewed according to the AQD. However, there is a lack of standardized methodologies based on scientific criteria to define the new zoning.
In this study, we developed and applied a new methodology based on mesoscale air quality modeling and statistical cluster analysis. We rely on a CMAQ 5.0.2 with 1 km^2 resolution air quality simulation over the whole Madrid region for the year 2015. From the 1-hour resolution model outputs, we computed all the relevant indicators for NO_2, $PM_{2.5}$, PM_{10} and O_3 as defined in the Directive 2008/50/EU. Representative statistics (mean, median, quartiles, etc) were calculated from the individual values of the grid cells overlapped for each of the 179 municipalities in the region. Then, the most informative parameters were selected as classification variables according to the result of a Principal Components Analysis (PCA). Subsequently, a k-means clustering algorithm was applied to identify municipalities with similar air quality features that could be homogeneous zones. The number of zones (or clusters) was defined through a series of tests that inform of the distribution of variance within and between clusters and the statistical significance of the differences found for each of the legal AQ indicators. The efficiency of the zoning is done by analyzing WCSS (Within Cluster Sum of Squares) in comparison to the total Sum of Squares. Besides, it is important to check the spatial continuity of municipalities within a cluster and the distribution of air quality monitoring stations (47 in the region) among them.
Following this methodology, we compared two alternative zonings to the current one: i) an optimal zoning from the statistical point of view and ii) one under consideration by the Madrid Greater Region air quality service.

Key words: *Air quality zoning, CMAQ, K-mean clustering, Representativeness, Zoning assessment*

INTRODUCTION

The European Air Quality Directive (AOD) (2008/50/EU) indicates that air quality zoning of the territory should be revised every 5 years. Air Quality (AQ) is supposed to be homogenous within AQ zones, so the whole zone is in a non-compliance situation if there is a single station that exceeds the legal limit values.

Madrid is located in the center of Spain and currently has seven zones established on important factors from the air quality point of view such as administrative divisions (179 municipalities), geographic data and land use, meteorological conditions, and emission sources. There is a dense AQ monitoring network, with 48 AQ monitoring stations in the region The last review for AQ zones of Madrid was 2014, therefore it is necessary to be revised. However, there is a lack of standardized methodology to define and revise the current zoning.

The principal aim of this study is to assess the current zoning of Madrid from the AQ point of view, considering the main pollutants (NO_2, $PM_{2.5}$, PM_{10} and O_3), which have a major impact on human health and vegetation and are relevant regarding AQ regulation. Then, we intend to identify an alternative that includes more homogenous AQ zones. Also, we perform a statistical comparison of these two zoning

schemes (the current and the alternative), and a third one zoning currently under consideration by the Madrid Greater Region air quality service. Moreover, we try to contribute to the definition of objective criteria that may be used elsewhere for the definition of homogeneous AQ zones.

METHODOLOGY AND RESULTS

As a first step, we obtain the spatial and temporal distribution of the target pollutants in 2015 through the Comunity Multiscale Air Quality (CMAQ) modeling system with a 1 x 1 km^2 over the Madrid region. This system consists of three models, i) WRF, the Weather Research and Forecasting (WRFV 3.7.1) model (Borge *et al.*, 2008; Skamarock and Klemp, 2008), ii) the Sparse Operator Kernel Emission (SMOKEV 3.6.5) model processes the emission inventory for the Region (Borge *et al.*, 2014; UNC, 2015) and lastly, iii) the Chemistry Multiscale Air Quality (CMAQV 5.0.2) (Byun and Schere, 2006), an Eulerian transport-chemical model. Model setup details can be found in Borge *et al.* (2018). We use this modeling system to calculate the legal parameters of all relevant pollutants according to the AQD:

- Annual mean NO_2 concentration,
- Annual mean PM_{10} concentration,
- Annual mean $PM_{2.5}$ concentration,
- 99.8th percentile of hourly NO_2 concentration,
- 93.2th percentile of eight-hourly O_3 concentration,
- 90.4th percentile of daily PM_{10} concentration,
- AOT 40 (Accumulated dose of ozone Over a Threshold of 40 ppb) for the period May – July.

The following seven statistical variables are calculated for each one of the 179 municipalities, considering all the grid cells overlapped: mean, median, quartile 25 (Q1), quartile 75 (Q3), interquartile range (IQR), standard deviation (sd) and coefficient variation (cv).

Then, a Principal Component Analysis (PCA) is performed to identify the most statistically relevant metrics for each parameter because (Lletı *et al.*, 2004) studied that the results can be changes if there are irrelevant variables for the analysis. Consequently, the classification variables selected are mean, median, Q1 and Q3 for all the parameters, and additionally, cv is relevant for PM_{10}, $PM_{2.5}$ and annual mean NO_2 concentration. This selection guarantees a good separation for the alternative zoning in the clustering process.

These variables are used for the k-mean clustering analysis (MacQueen, 1967; Hartigan and Wong, 1979) known as *unsupervised learning* to separate homogenous groups, following the steps:

- Assign randomly *k* numbers of clusters,
- Calculate the distances between centroid points and other observations in *k* clusters,
- Reassign the mean value of the distances to new centroid points of clusters,
- Repeat until finding adequate centroid points that have a minimum distance with the observations in a cluster.

Among others, three methods are employed to specify the optimal number of groups o *clusters*: Elbow, Silhouette (Kaufman and Rousseeuw, 2009) and Gap statistic (Tibshirani, Walther and Hastie, 2001). We performed two cluster classifications: one for NO_2, PM_{10}, and $PM_{2.5}$, and the other for O_3 due to the different characteristics between pollutants. The optimal number obtained is five in the case of the three pollutants and four as for O_3 and this zoning is referred to as optimal zoning in this study. Then, the three zonings are compared: the current, the optimal (one for O_3 and another one for all other pollutants) and the proposed (Figure 1).

For the comparison to assess the three alternative zoning options, boxplot graphs are generated for concentration dispersions of each zone and each municipality. In an ideal classification, the groups would have very similar mean values with a small dispersion, given by a reduced IQR and absence of extreme values far from the mean (*outliers*). As shown in Figure 2, the separation seems better in the optimal zoning than in the others as the IQR values of zones are comparably small. Zone 1, which corresponds to Madrid city, tends to have high concentrations as for NO_2, but low respecting O_3, having great IQR

values for both pollutants. However, it was considered as a single zone due to administrative reasons and considering that it has its own AQ network.

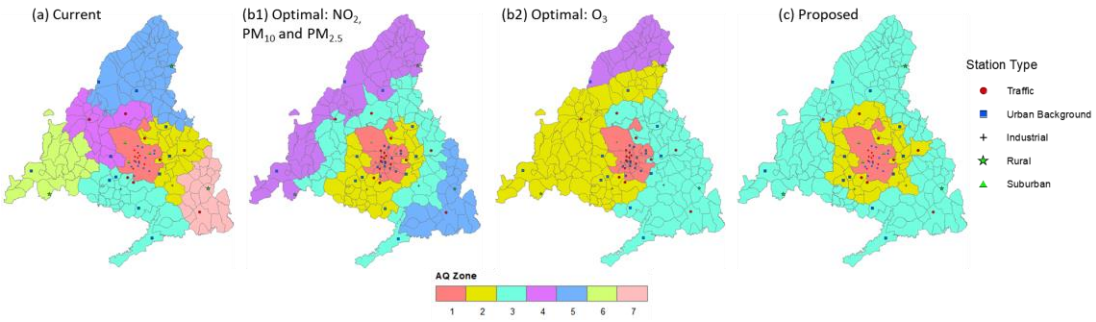

Figure 1. The current zoning of the Madrid region (a), the optimal obtained by the analysis (b): (b1) shows the zoning for the three pollutants and (b2) is for ozone, and the proposed by the Madrid Greater Region air quality service (c). The distribution and type of air quality monitoring stations is included in all cases.

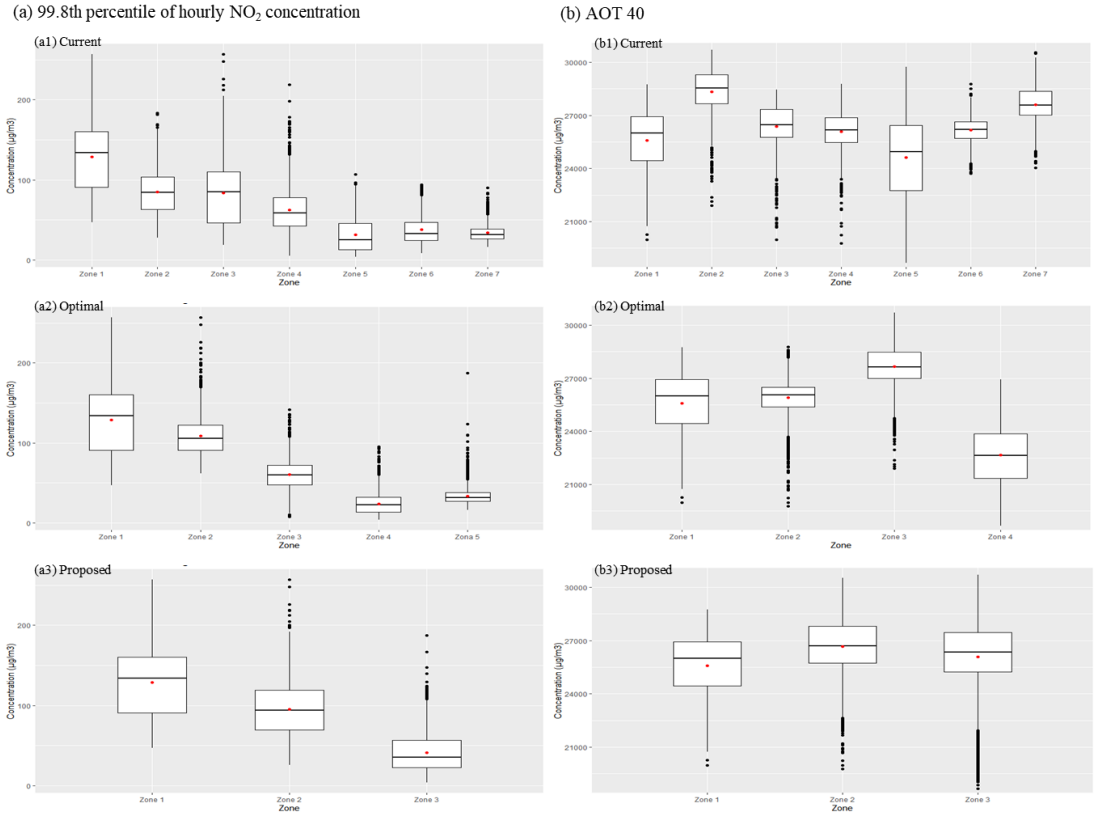

Figure 2. Boxplot graphs generated by each zone for 99.8th percentile of hourly NO_2 concentration (a) and AOT 40 (b): (a1) and (b1) present the current zoning, (a2) and (b2) are for the optimal and (a3) and (b3) for the proposed by the Madrid Greater Region.

After that, two statistical tests are performed: the Kruskal-Wallis test and the Dunn test. These tests are employed to find significant differences between the objects, the groups o *clusters* in this study. The Kruskal-Wallis test (Kruskal and Wallis, 1952) compares them with a null hypothesis which is the objects come from the same population as a non-parametric test, on the other hand, the Dunn test (Dunn, 1964) compares them in pair as *post-doc*, whose null hypothesis is two groups are from the same population. According to the results of the Dunn test, we can observe some non-different zones in the current zoning.

There are little differences between zones 6 and 7 of the parameters of NO_2, and zones 4 and 7 of the 90,4th percentile of daily concentration of PM_{10} are similar. Also, zones 1 and 4, and zones 5 and 6 of the 93.2th percentile of eight-hourly O_3 concentration, and zones 4 and 6 of the AOT 40 are considered identical. On the other hand, all zones are significantly different in the proposed zoning, while the two zones found very similar in the optimal alternative. This result is due to the similarity among center values (mean or median) although the IQR value of zone 1 is much greater.

Besides, the distribution of the Sum of Squared Error (SSE) between BSS and WSS for each zoning is analyzed. The total SSE (TSS) is the sum of distances between an object and the mean value of the entire dataset and consists of between SSE (BSS) that is the inter-cluster variance and within SSE (WSS) that is the intra-cluster variance. Therefore, BSS is close to TSS if the groups are well-separated. As shown in Table 1, the optimal zoning obtains the best separation for all pollutants according to this.

Table 1. SSE values of the three zonings

Pollutant	Parameter	Current zoning		Optimal Zoning		Proposed Zoning	
		k	BSS/TSS (%)	k	BSS/TSS (%)	k	BSS/TSS (%)
NO_2	Annual Mean	7	56.7	5	60.3	3	45.5
	Hourly Percentile	7	71.3	5	88.3	3	65.0
PM_{10}	Annual Mean	7	48.7	5	58.9	3	33.6
	Daily Percentile	7	57.5	5	76.3	3	40.8
$PM_{2.5}$	Annual Mean	7	44.0	5	53.1	3	29.0
O_3	8-hourly Percentile	7	45.0	4	68.8	3	7.9
	AOT 40	7	48.3	4	78.1	3	4.8

Additionally, it was checked that the optimal zoning complies with the minimum coverage and distribution of air quality monitors currently available in the region. In general, redundancy increases along with coverage. The current zoning has poor coverage and redundancy, while the optimal has better coverage and redundancy, but less than the proposed one. The proposed zoning shows less redundancy than the optimal zoning, mainly because it consider a smaller number of groups.

CONCLUSION

A novel methodology is used to revise the current air quality zoning of Madrid and to propose an alternative one that reflects a better homogeneity in the AQ zones from the AQ point of view. We use a combination of a Chemical-transport model and k-mean clustering analysis. This methodology is found quantitative and replicable. However, it requires a precise AQ simulation, and the results could be changed depending on the model validation.

We found that a single classification cannnot define homogeneous AQ areas for all pollutants since the dynamics of O_3 considerably differs from that of the rest of the species of interest (NO_2, PM_{10} and $PM_{2.5}$). The zoning assessment indicates that the optimal zoning obtained with the new methodology shows the best statistical result and the most homogeneous zones from the AQ point of view and improves the current zoning of the Madrid region.

REFERENCES

Borge, R. *et al.* (2008) 'A comprehensive sensitivity analysis of the WRF model for air quality applications over the Iberian Peninsula', *Atmospheric Environment*, 42(37), pp. 8560–8574. doi: https://doi.org/10.1016/j.atmosenv.2008.08.032.

Borge, R. *et al.* (2014) 'Emission inventories and modeling requirements for the development of air quality plans. Application to Madrid (Spain)', *Science of the Total Environment*. Elsevier B.V., 466–467, pp. 809–819. doi: 10.1016/j.scitotenv.2013.07.093.

Borge, R. *et al.* (2018) 'Application of a short term air quality action plan in Madrid (Spain) under a high-pollution episode - Part II: Assessment from multi-scale modelling', *Science of The Total Environment*. Elsevier, 635, pp. 1574–1584. doi: 10.1016/J.SCITOTENV.2018.04.323.

Byun, D. and Schere, K. L. (2006) 'Review of the governing equations, computational algorithms, and other components of the Models-3 Community Multiscale Air Quality (CMAQ) modeling system'.

Dunn, O. J. (1964) 'Multiple comparisons using rank sums', *Technometrics*, 6(3), pp. 241–252.

Hartigan, J. A. and Wong, M. A. (1979) 'Algorithm AS 136: A k-means clustering algorithm', *Journal of*

the Royal Statistical Society. Series C (Applied Statistics), 28(1), pp. 100–108.
Kaufman, L. and Rousseeuw, P. J. (2009) *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.
Kruskal, W. H. and Wallis, W. A. (1952) 'Use of ranks in one-criterion variance analysis', *Journal of the American statistical Association*, 47(260), pp. 583–621.
Lletı, R. *et al.* (2004) 'Selecting variables for k-means cluster analysis by using a genetic algorithm that optimises the silhouettes', *Analytica Chimica Acta*, 515(1), pp. 87–100.
MacQueen, J. (1967) 'Some methods for classification and analysis of multivariate observations', in. Oakland, CA, USA, pp. 281–297.
Skamarock, W. C. and Klemp, J. B. (2008) 'A time-split nonhydrostatic atmospheric model for weather research and forecasting applications', *Journal of Computational Physics*. Academic Press Inc., 227(7), pp. 3465–3485. doi: 10.1016/j.jcp.2007.01.037.
Tibshirani, R., Walther, G. and Hastie, T. (2001) 'Estimating the number of clusters in a data set via the gap statistic', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2), pp. 411–423.
University of North Carolina at Chapel Hill (UNC) (2015) 'SMOKE's v365 user's manual'. Chapel Hill, NC: University of North Carolina. Available at: https://www.cmascenter.org/smoke/documentation/3.6.5/manual_smokev365.pdf.