HARMO22

22nd International Conference on
Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes
10-14 June 2024, Pärnu, Estonia

RETRIEVAL OF AIR QUALITY ANNUAL STATISTICS FROM A LIMITED NUMBER OF PROFILES

Tereza Pikousová[1], Kryštof Eben[2], Ondřej Vlček[1], Jaroslav Resler[2] and William Patiño[1]

[1] Czech Hydrometeorological Institue, Na Šabatce 2050/17, 143 00 Prague 12, Czech Republic
[2] Institute of Computer Science of the Czech Academy of Sciences, Pod Vodárenskou věží 271/2, 182 00 Prague 8, Czech Republic

Abstract: Annual air quality statistics in an urban environment are usually computed from data collected from a sparsely distributed network of monitoring stations. Complex numeric models may provide air quality simulations at the street level, these simulations, however, are not computationally feasible for large time periods like a year. We propose a method for identification of a limited number of "typical" days, which, if simulated in microscale, guarantee a reasonable coverage of different scenarios during the year. Annual statistics then can then be estimated on street level from simulated fields. The identification method is based on k-medoids clustering. We also develop a means for validating the approach so as to add confidence in estimates derived in this manner.

Key words: *air quality, annual statistics, urban environment, k-medoids, clustering*

INTRODUCTION

According to European legislation on air quality, national authorities are obliged to report a number of annual statistics of pollutant concentrations. These statistics are usually calculated from data given by a sparse monitoring network. Local differences in human exposure in the urban environment have to be quantified by other means, presumably using model simulations. In this way variability in reported values and differences in human exposure to pollutants would be assessed more precisely.

Recent advances in supercomputing capabilities have allowed the application of large-eddy simulation models like PALM (Maronga et al., 2020) for the assessment of urban air quality at a resolution of meters and over domains covering areas of units of square kilometers. Simulating selected episodes gives insight into local variability of air quality. The enormous computational requirements of complex models like PALM, however, make the simulation of a whole year, needed for computation of annual statistics, impossible. As a consequence, different methods start to emerge in the literature on how to reconstruct annual statistics from a restricted number of "typical" scenarios. No widely adopted methodology exists so far. The aim of this research is to develop methods for identification of "typical" daily profiles of key pollutant concentrations. Performing a set of microscale simulation for typical days using e.g. PALM and using an appropriate retrieval method would bring a better notion of local differences of air quality in the urban environment.

DATA

It has been documented recently (Radovic et al., 2023, Resler et al., 2024) that the driving initial and boundary conditions for any PALM simulation have a decisive influence on its result and precision. Thus, for selecting typical scenarios we have to use the data on a much larger domain than the microscale simulation domain. Therefore we use the data provided by the urban air quality monitoring network as a basis for our computations. In this case study we use data for the territory governed by Prague Municipality (the capital of the Czech Republic). If we want to claim legitimacy of extension of the results of a limited

number of microscale simulations to the entire year, a minimal requirement is that the standard reported statistics are estimated with reasonable accuracy.

The data consists from hourly measurements of key pollutant concentrations from Prague air quality stations together with meteorological data from the professional stations located in Prague. The examined year is 2023.

The starting pool of profiles is the set of 365 daily courses of NO_2, O_3, PM_{10}, and components of wind U and V. These daily profiles, concatenated for each day from many different stations into one long vector, enter a clustering method described below. In order to maintain balance in comparing differences in various components, the data underwent standardization, and further exclusion of days with missing data was done. After that preprocessing, we ended up with 306 days.

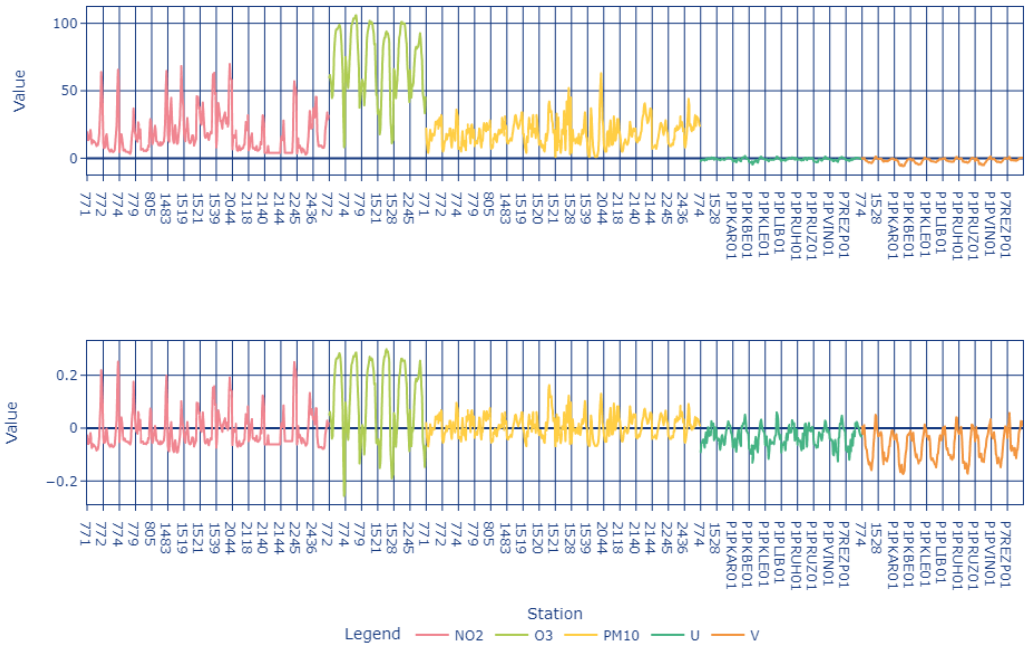

Figure 1. Concatenated daily courses of variables on stations for 4th July 2023. Raw data profile (top), standardized and weighted profile (bottom). The codes on the x-axis are national station identificators. For more information on stations see https://www.chmi.cz/files/portal/docs/poboc/OS/stanice/ShowStations_CZ.html.

METHOD

The basic principle of retrieving the annual statistics in a local scale can be described using the following approach:

1. Select a restricted number of "typical days" according to the measurements
2. Classify all the days of the whole year into categories represented by these "typical" days
3. Reconstruct fields of the annual concentration profile by replacing each day with the corresponding "typical day" according to the previous classification and compute the required statistics (no further processing of typical day values was applied).

The last step can be performed with the concentration fields simulated for "typical" days with a microscale model, thus giving an estimate of local variability of exposure. Instead of this, we in this study perform the retrieval for monitoring stations themselves, for the purpose of validation of the method. Such a validation must be carried out before any local simulation is done and we check the results against true statistics observed at the monitoring stations entering the process. This parallels the standard reporting process as if only typical days were available.

It turns out that our task leads to a standard problem of cluster analysis. Denote x a feature vector described in section DATA and (m_i , i = 1,...,k) the set of profiles selected as typical. Also, denote $d(x,m)$ a measure of distance or dissimilarity between feature vectors x and m. Then if C_i is the i-th cluster, the error of the yearly time series arising from replacing each day by the corresponding typical day may be quantified with equation (1).

$$TD = \sum_{i=1}^{k} \sum_{x \in C_i} d(x, m_i) \tag{1}$$

Minimisation of (1) can be identified as one of the well-known tasks of cluster analysis, namely selecting the "medoids" from a set of feature vectors. The cost function (1) is sometimes called *total distance*.

Now, if we can preselect the number of clusters, we come to the problem of k-medoids. The method dates back to the eighties (Kaufman and Rousseeuw, 1987) and since then a number of algorithms have been designed to solve this (NP-hard) problem. The most common one is the PAM algorithm (Kaufman and Rousseeuw, 2009), which we use in this study. We selected the standard Euclidean metric as the distance in equation (1). Since our data is moderate in size, we could compute sets of medoids for all possible values k = 1, ..., 306. Selecting an optimal number of clusters is a tradeoff between the accuracy of the retrieval and the computational burden of the prospective microscale simulation. This choice has to be made on an expert basis and we may supply different criteria to support this process.

Since our target is to retrieve a number of different statistics, we can take into account the accuracy of each of them and possibly form an aggregate criterion. These statistics are annual means of key compounds together with additional statistics describing extremes.

Each of these statistics is observed on several monitoring stations and errors of retrieval form a set of replications (this set cannot be viewed as a random sample but computing descriptive statistics bring the information desired). For a fixed k and a statistic S observed on $n(S)$ stations with true values ($S_1^*, \dots, S_{n(S)}^*$) and retrieved estimates $S_1, \dots, S_{n(S)}$), the error is measured with the relative root mean square error (RRMSE)

$$\mathrm{RRMSE}(S) = \frac{\left[\frac{1}{n(S)} \sum_{j=1}^{n(S)} (S_j - S_j^*)^2\right]^{1/2}}{\frac{1}{n(S)} \sum_{j=1}^{n(S)} S_j^*} \tag{2}$$

which guaranties some extent of comparability of these quantities for different statistics. We may track each of these RRMSEs separately for increasing k or we can sum all such values to form an aggregate overall criterion.

All calculations (including the PAM algorithm) were performed in python (Schubert and Lenssen, 2022).

RESULTS

The annual statistics for compounds NO_2, PM_{10} and O_3 were estimated for Prague stations in the year 2023 with the method described above. **Figure 2** (left) shows the sum of RRMSEs over all statistics as a function of k. These statistics comprise annual means and further statistics providing limits of the pollutants examined: 19th highest hourly value of NO_2, 36th highest daily mean of PM_{10} and 26th highest daily maximum of 8-hour rolling mean of O_3.

The only input of the k-medoids clustering algorithm is the distance matrix, formed by the distances $d(x_i, x_j)$ of two feature vectors x_i, x_j (cf.(1)), in our case computed from standardized and weighted data. If we order the feature vectors by cluster affiliation, we may compare heatmaps of distance matrices for different number of clusters. These heatmaps reflect the similarity within clusters (diagonal blocks) and dissimilarity between the clusters. **Figure 2** (middle and right) shows two such heatmaps, one for 5 clusters and the

other one for 15 clusters (the number of 15 days correspond to number of days feasible to be simulated by PALM using computational resources authors have currently available).

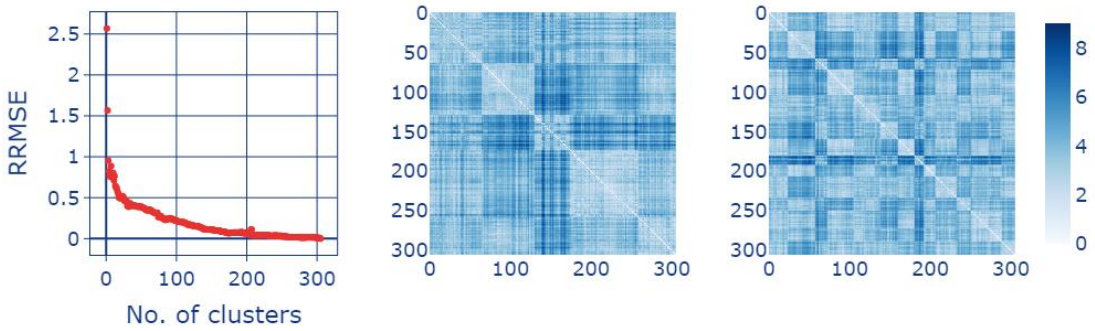

Figure 2. Aggregated RRMSE as a function of number of clusters; PAM k-medoids computed for all possible values of k (left). Distance matrix ordered by clusters affiliation for k = 5 (middle) and k = 15 (right).

After checking the medoids and homogeneity of clusters from the meteorological point of view, the estimates of annual statistics can be computed. As an example we show in **Figure 3** the results of the retrieval for annual mean concentration of O_3 and NO_2 based on 15 medoids, whereas **Figure 4** displays results for annual 19th maximal value of NO_2 and 36th daily mean of PM_{10}, based on 15 medoids.

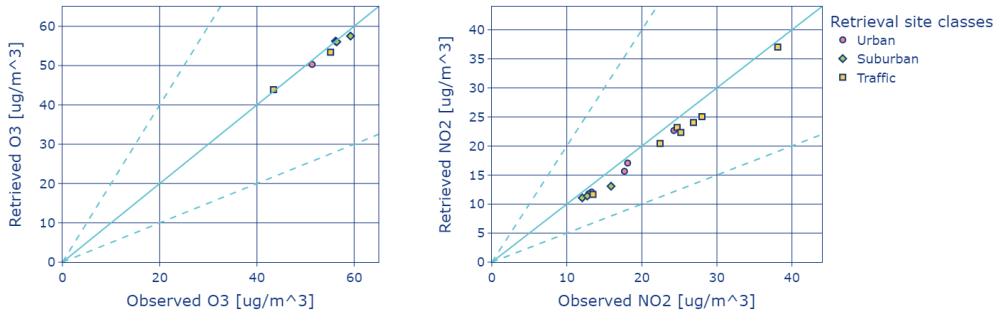

Figure 3. Retrieved and observed annual means of O_3 (left) and NO_2 (right) for traffic, urban and suburban stations in Prague in 2023 (15 medoids were used)

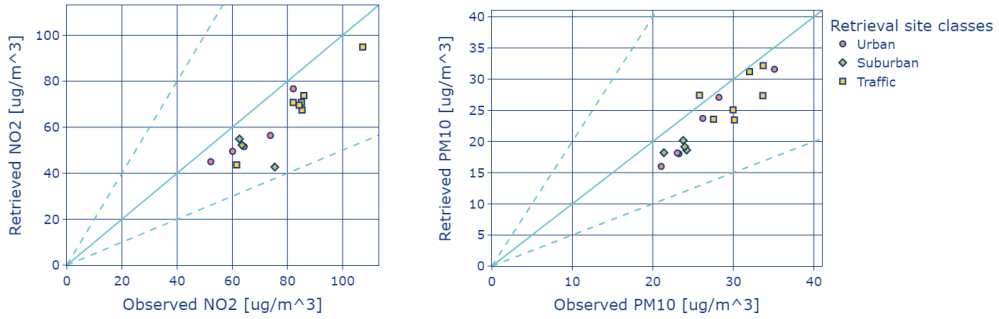

Figure 4. Retrieved and observed annual 19th maximal value of NO_2 (left) and 36th daily mean of PM_{10} (right) for traffic, urban and suburban stations in Prague in 2023 (15 medoids were used)

DISCUSSION AND CONCLUSION

Estimation of annual statistics from the data observed during typical days has been performed in the most straightforward way as stated in sec. METHOD. In some cases this may be sufficient as documented by

the retrieved annual means on monitoring stations, which may be recovered surprisingly well even from 15 daily courses (**Figure 3**). This is not necessarily true for all compounds and statistics. From **Figure 4** it is seen that statistical correction or postprocessing of the results may be required to eliminate bias. This is expectable in case of statistics based on extremal values or order statistics. A deeper examination of the statistical properties of the method will be necessary.

On the other hand, **Figure 2** (left) suggests that after reaching a larger number of clusters, the precision may be sufficient and we could discard a large number of days which otherwise would enter the yearly simulation.

The method seems to be promising and we expect that it can bring a progress in quantifying local differences in human exposure to atmospheric pollutants in urban environment.

ACKNOWLEDGMENT
The contribution was financially supported by the Technology Agency of the Czech Republic through the program Environment for Life (project SS02030031 ARAMIS), and by the project TURBAN (TO0100021) funded by Norway Grants and the Technology Agency of the Czech Republic within the KAPPA Programme.

REFERENCES

Kaufman, L. and P. J. Rousseeuw, 1987: Clustering by means of Medoids. *Statistical data analysis based on the L1 – norm and related methods*, edited by Y. Dodge , North-Holland.

Kaufman, L. and P. J. Rousseeuw, 2009: Finding groups in data: an introduction to cluster analysis. John Wiley & Sons.

Maronga, B., S. Banzhaf, C. Burmeister, T. Esch, R. Forkel, D. Fröhlich, V. Fuka, K.F. Gehrke, J. Geletič, S. Giersch and others, 2020: Overview of the PALM model system 6.0. *Geoscientific Model Development*, **13**, 3, 1335-1372.

Radović J., M. Belda, J. Resler, K. Eben, M. Bureš, J. Geletič, P. Krč, H. Řezníček and V. Fuka, 2023: Challenges of constructing and selecting the "perfect" boundary conditions for the LES model PALM. *Geoscientific Model Development Discussions*, Göttingen, Germany, **2023**, 1-40, doi: 10.5194/gmd-17 2901-2024.

Resler, J., P. Bauerová, M. Belda, M. Bureš, K. Eben, V. Fuka, J. Geletič, R. Jareš, J. Karel, J. Keder, P. Krč, W. Patiño, J. Radović, H. Řezníček, M. Sühring, A. Šindelářová and O. Vlček.: Challenges of high-fidelity air quality modeling in urban environments - PALM sensitivity study during stable conditions. Submitted to Geoscientific Model Development

Schubert, E. and L. Lenssen, 2022: Fast k-medoids Clustering in Rust and Python. *J. Open Source Softw*. https://doi.org/10.21105/joss.04183