Atmospheric Environment 47 (2012) 111–123

Contents lists available at SciVerse ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

Objective classification of air quality monitoring sites over Europe

Mathieu Joly [a,*], Vincent-Henri Peuch [b]

[a] Météo-France CNRM-GAME, 42 av. Coriolis, 31057 Toulouse Cedex 1, France
[b] ECMWF, Shinfield Park, Reading, United Kingdom

ARTICLE INFO

Article history:
Received 29 August 2011
Received in revised form
9 November 2011
Accepted 10 November 2011

Keywords:
Air quality
Classification
Monitoring
Representativeness
Pollutant
European AirBase

ABSTRACT

The observation sites that make up air quality monitoring networks can have very different characteristics (topography, climatology, distance to emission sources, etc), which are partially described in the meta-information provided with data sets. At the scale of Europe, the description of the sites depends on the institute(s) in charge of the air quality monitoring in each country, and is based on specific criteria that can be sometimes rather subjective. The purpose of this study is to build an objective, homogeneous, and pollutant-specific classification of European air quality monitoring sites, primarily for the purpose of model verification and chemical data assimilation.

Most studies that tackled this issue so far were based on limited data sets, and often took into account additional external data such as population density, emission estimates, or land cover maps. The present study demonstrates the feasibility of a classification only based on the past time series of measured pollutants. The underlying idea is that the true fingerprint of a given monitoring site lies within its past observation values. On each site to be categorized, eight indicators are defined to characterize each pollutant time series (O_3, NO_2, NO, SO_2, or PM_{10}) of the European AirBase and the French BDQA (Base de Données de Qualité de l'Air) reference sets of validated data over the period 2002–2009. A Linear Discriminant Analysis is used to best discriminate the rural and urban sites. After projection on the Fisher axis, ten classes are finally determined on the basis of fixed thresholds, for each molecule.

The method is validated by cross-validation and by direct comparison with the existing meta-data. The link between the classes obtained and the meta-data is strongest with NO, NO_2, and PM_{10}. Across Europe, the classification exhibits interesting large-scale features: some contrasts between different regions depend on the pollutant considered. Comparing the classes obtained for different pollutants at the same site reveals an interesting consistency between the separate classifications. The robustness of the method is finally assessed by comparing the classifications obtained for two distinct subsets of years. The robustness – and thus the skill of the objective classification – is satisfying for all of the species, and is highest with NO and NO_2.

1. Introduction

As a consequence of industrialization, urbanization, and fossil fuel use, air pollution has been rising in most parts of the world over the last decades (Vingarzan, 2004; Oltmans et al., 2006). Most developed countries have set up laws and developed air quality measurement networks to monitor pollutant concentrations, and issue warnings when acceptable levels are exceeded (Romano et al., 1999; ADEME, 2002; Lau et al., 2009). The design of air quality monitoring networks depends on different local constraints, such as financial resources, environmental priorities, or political decision-making. The pollutants to be monitored, and the scope and quality of the data collected are all subject to these constraints. Because air pollution is larger in urban and industrial areas, the monitoring effort is usually concentrated in and around the cities (e.g., Gramsch et al., 2006), where high emissions may lead to concentrations above the threshold values. The main task of the local governments is indeed to assess the population exposure and the impact on health, and to determine compliance with national or international standards. However, "background" air quality is also measured in countryside areas, as far as possible from the main emission sources, which is essential for evaluating large-scale variability and trends, as well as evaluating air quality models. This is for example the approach followed in the framework of EMEP (European Monitoring and Evaluation Programme). Overall, this results in quite heterogeneous networks – especially in Europe – both in terms of spatial distribution (high density of sites in and near the cities) and of spatial representativeness, i.e. the scale of the

* Corresponding author. Tel.: +33 561 07 98 32; fax: +33 561 07 96 10.
E-mail address: mathieu.joly@meteo.fr (M. Joly).

1352-2310/$ – see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.atmosenv.2011.11.025

area that the measurement is supposed to be representative of (Spangl et al., 2007).

Observations in street canyons and city centers are spatially less representative than observations in rural background areas. According to Spangl et al. (2007), the assessment of representativeness is equivalent to the delimitation of areas where air pollution has similar characteristics. Classifying monitoring sites and assessing representativeness are thus related tasks. In most air quality data sets, measurements are accompanied by a detailed description of the area in which it is done. Such meta-information is precious, since it provides a basis for a first estimate of the representativeness, based on a more or less semi-quantitative assessment of some parameters influencing the pollution level like emissions, population distribution, land use, and the topographic configuration. However, such classifications – which are most of the time the only one available – are not universal and rely on the data monitoring operators.

In Europe, there are presently three classifications of air quality monitoring. The first derives from the Council decision 97/100/EC called "Eol" (ADEME, 2002). A second comes from the work done by the European Topic Centre on Air and Climate Change (ETC-ACC) on behalf of the European Environment Agency within the framework of the EUROAIRNET project (Ibid.). The third classification derives from the European directives relating to air quality (especially directives 96/62/EC, 99/30/EC, 2000/69/EC), and from the new "ozone directive" 2002/3/CE (Ibid.). These different European classifications are not standardized; in particular the number of classes is not always the same. The primary ordering key can also be different: it corresponds to the *nature of the sources* in the "Eol" classification and to *exposure* in the "ozone directive". Besides, some countries have developed their own national classification rules in compliance with these general requirements (e.g., France and Great-Britain). This contributes to the inhomogeneity of meta-information at the scale of Europe.

Another shortcoming of the current meta-data is that it is not related to the different pollutants. The specification of a major emission source can therefore be quite ambiguous for the data user (Spangl et al., 2007). The Eol "type of station" refers to the "station" and does not take into account that the contributions of certain sources may differ largely for different pollutants. For example, industrial sources may contribute to some pollutants but not to others. Current classifications, that are not pollutant-specific, may thus obscure the impact of some pollutant sources (e.g., a contribution to SO_2 from industry at a traffic station). Beyond the emissions, the other factors (chemistry, dispersion, and transport) influencing air pollution levels are also pollutant-specific.

The purpose of this study is to build an objective classification that is homogeneous at the scale of Europe and specific to each pollutant. The classification should be stable over the considered period, and any new site should be easily classified *a posteriori,* provided that enough data is available. A number of previous studies had similar objectives: they are listed in Table 1. An important difference between the different approaches is the data employed. Some studies use both air quality data (measured or modeled) and additional data of some parameters influencing air quality (emissions, building structure, land use, topography, etc) or the receptors (human population, ecosystems, etc). Besides, most studies rely on very small air quality data sets or rather short periods, compared to the amount of air quality monitoring sites across Europe. Finally, some studies (Tarasova et al., 2007; Henne et al., 2010; Kovač-Andrić et al., 2010) rely on the EMEP network, which is specially designed to avoid influences and contamination from local sources, in order to assess long-range trans-boundary air pollution transport.

In the present paper, we have chosen to implement a classification based on the measurement data itself, using all the data available in the AirBase data set for Europe, and the French data set named BDQA hereafter (Base de Données de Qualité de l'Air, i.e. Air Quality Data Base), which is more complete. We deal with near-surface concentrations, which means that the vertical distribution of the pollutants is not taken into account. For each of the measured pollutants, the goal is to group time series that are homogeneous from the point of view of their statistical properties. In the framework of the MACC (Monitoring Atmospheric Composition and Climate) project, this objective classification is proposed for model verification and chemical data assimilation. MACC (http://www.gmes-atmosphere.eu/) is the current pre-operational atmospheric service of the European GMES program, for which an ambitious ensemble of regional air quality multimodel forecasts has been developed (Hollingsworth et al., 2008; Huijnen et al., 2010).

Section 2 details the statistical processing of the hourly time series: the data sets employed, the time-filtering, and the computation of eight indicators. Section 3 describes the behavior of the indicators, their transformation, and some preliminary statistical results. Section 4 details the classification procedure, the cross-validation, a description of the results, and a robustness assessment. Finally, Section 5 discusses the results and concludes the study.

Table 1
Overview of the data used in the literature to classify Air Quality (AQ) monitoring sites.

	Period considered	Data sets used for classification	Pollutants considered
Flemming et al., 2005	1995–2001	German AQ data	O_3, NO_2, SO_2, PM_{10}
Henne et al., 2010	2005	- 34 EMEP AQ sites - Population density - Land-cover map - Meteorological fields	O_3, NO_2, CO
Ignaccolo et al., 2008	2006	68 Italian Piemonte AQ sites	O_3, NO_2, PM_{10}
Kovač-Andrić et al., 2010	1997–2003 summers	12 EMEP AQ sites	O_3
Lau et al., 2009	2001–2005	14 Hong-Kong sites	NO_2, PM_{10}
Monjardino et al., 2009	1995–2002	- 51 Portugal AQ stations - Population density	O_3, NO_2, NO, CO, SO_2
Snel, 2004	1999, 2001 and 2002	Dutch AQ stations	NO, NO_2
Spangl et al., 2007	2002–2004	- Austrian AQ data + Netherlands for validation - Emission inventory - Land-cover map - Population density	O_3, NO_2, PM_{10}
Tarasova et al., 2007	1990–2004	114 EMEP AQ sites	O_3
This study	2002–2009	- AirBase European AQ data - BDQA French AQ data	O_3, NO_2, NO, SO_2, PM_{10}

2. Statistical processing of the hourly time series

For each site and each pollutant, time series of past measurements are treated independently from each other, which means that spatial relationships are not considered. Time is thus the only variable. Given the length of the time series (from 1 to 8 years of hourly data for each site), the inter-annual variability cannot be taken into account. We will focus on the shorter time-scales (from a couple of hours to a few days), or on some averaged features (diurnal and weekly cycles, or winter vs. summer differences). The goal is to extract from the hourly data all the information that can help to segregate the monitoring sites.

2.1. European air quality observations

For this study, we have chosen two data sets of validated and officially reported data:

- The AirBase data set (Version 5 released in April 2011) is managed by the ETC/ACC on behalf of the European Environment Agency. It contains air quality monitoring data and information submitted by the participating countries throughout Europe. For the pollutants and the period chosen, 4956 sites are available in 35 countries.
- For France, the data set called "Base de Données Qualité de l'Air" (BDQA) was maintained until 2010 by the "Agence de l'Environnement et de la Maîtrise de l'Energie" (ADEME). For the period considered, 1190 sites are available, among which 200 are not in AirBase.

The period considered is 2002–2009. The hypothesis is that over this period, no important change has occurred in the configuration of the sites. Eight years is indeed a compromise between enough data for the statistical analysis, and a period hopefully short enough to avoid possible drifts or discontinuities in the measurements (e.g., urbanization of some rural areas). Besides, sites with less than 8760 hourly values (the equivalent of one full year of data) have been discarded.

Five pollutants are considered: O_3, NO_2, NO, SO_2 and PM_{10} (Particulate Matter, size < 10 μm). Two pollutants could not be considered: $PM_{2.5}$ because they are insufficiently represented within the AirBase data set; and CO, whose main drawback is the lack of sufficient data in rural areas (more than half of the CO measurements are located in "traffic" sites).

For the sites that are in common in AirBase and BDQA data sets, time series are identical. However, the meta-information differs. Table 2 compares the BDQA classification (ADEME, 2002) with the corresponding "station type of area" provided by AirBase. Strangely enough, some sites considered as "urban" in the French BDQA data sets are located in a "rural" area according to AirBase V5. Overall, the classification is different for 32% of the sites. This shows the limits of the meta-data, and the potential benefits of an automated procedure, at least for verification and harmonization at the European scale.

Table 2
Comparison between AirBase and BDQA meta-data for the French sites that are in both data sets. Only AirBase 'background' sites are considered.

		BDQA		
		Urban	Suburban	Rural
AirBase	Urban	228	21	
V5	Suburban	125	109	6
	Rural	15	25	66

2.2. The diurnal cycle

Because of the diurnal character of human activities and emissions, as well as of the solar radiation, pollutant levels are in general dependent upon the hour of the day. The diurnal cycle is thus at the heart of the time evolution of anthropogenic pollutants in the lowest troposphere.

The diurnal cycle of the concentrations is an average feature of the hourly time series. It evolves during the year, and is modulated by the inter-annual variability. To take into account this time-dependence, it has been chosen to compute the diurnal cycle over a 31-day sliding window. For each day, the diurnal cycle is an average that takes into account the 15 preceding and 15 following days. Note that for the hourly average to be computed, the ratio of missing values has to be lower than 20%, otherwise the value is set to missing.

Fig. 1a illustrates the sliding diurnal cycle computed for the NO_2 measurements of a polluted site located close to Paris "Boulevard Périphérique". Over the considered period (September 2006), the monthly averaged diurnal cycle is not evolving much. It is characterized by two diurnal peaks of high concentrations at the rush hours, the evening peak being the largest. Unsurprisingly, lowest values are observed by night, when the traffic is lower.

For each day for which a monthly averaged diurnal cycle could be obtained, two values are stored: the daily maximum and the daily amplitude (daily maximum *minus* daily minimum). For each month of the year, those values are averaged (if at least 20 values are available, otherwise the average is set to missing), which yields – for each time series – an annual cycle of the diurnal cycle maximum and amplitude. Two other indicators are also computed: the annual mean (if the 12 monthly values are available, otherwise the site is discarded) and the "summer (April to September) *minus* winter (October to March) difference". In total, four indicators are kept, in order to characterize the diurnal cycle of the pollutants.

2.3. The high frequency variability

If we subtract the sliding diurnal cycle from the raw time series, there remains variability in the range from a couple of hours to some days (curve s-c in Fig. 1b). The lowest frequencies are directly tightened to the meteorological large-scale synoptic conditions. Fig. 2 shows that such large-scale fluctuations are synchronous over a same region, no matter the type of site (rural, urban, etc). Since our purpose is to distinguish the different types of sites, the interesting information is to be found in the high frequency variability. To filter out the low frequency from the time series, a FIR filter has been chosen, with a cut-off frequency of 3 days and a Kaiser window characterized by a 2-day transition width (Hamming, 1977). The FIR filter has been tuned for this study, and presents the advantage of having a constant group-delay throughout the frequency spectrum (half the filter order).

In order to filter out the very large-scale variability – that is common to all of the stations over a given region – without affecting the high frequency that we seek to quantify, a cut-off period at three days has been chosen. The desired indicator is then defined as the standard deviation of the filtered time series (curve s-c-l in Fig. 1b), in order to quantify the strength of high frequency anomalies (e.g., the influence of local plumes).

The issue of missing values is a tricky one when dealing with filtering. Discarding the missing values before filtering inevitably leads to artificial "steps" in the data. In our case, it has been verified that the impact of the discarded missing values on the high frequency signal (as we have defined it) is not discernible in the standard deviation of the filtered time series (not shown).

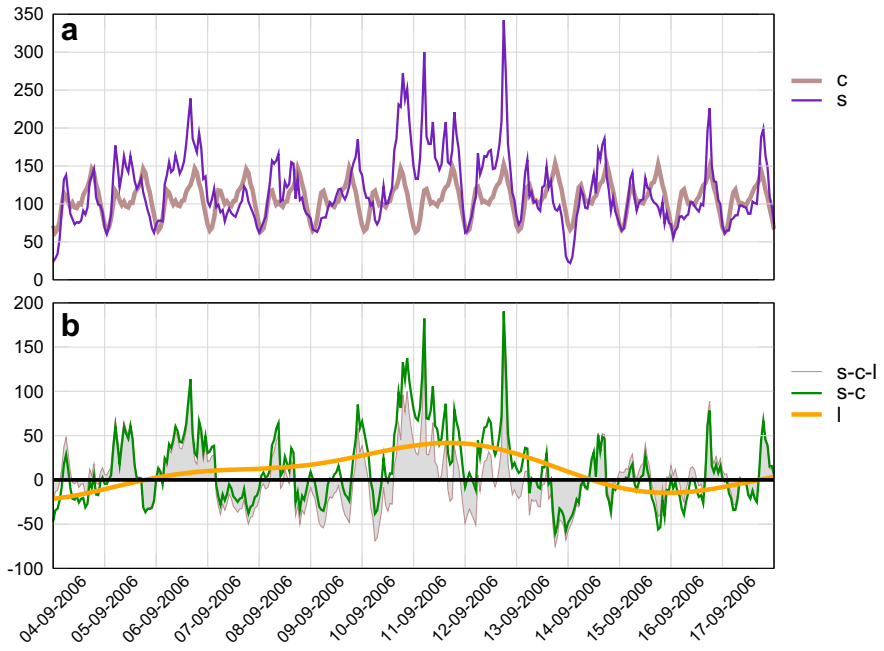

Fig. 1. Illustration of the filtering procedure for NO_2 hourly concentrations near Paris, along the "Boulevard Périphérique" (AirBase station FR04053). The first panel shows the row series (s) and the diurnal cycle (c) averaged over a sliding 31-day window centred on the considered day. The second panel shows the series filtered from the diurnal cycle (s-c), the signal filtered from the frequencies higher than 3 days (s-c-l), and the low frequency residual (l).

2.4. The "weekend" effect

Because pollutant concentrations at the surface strongly depend on anthropogenic emissions, the measurements have different characteristics on weekdays and during weekends (Cleveland et al., 1974; Beirle et al., 2003; among others). Such a weekly cycle – often called "weekend effect" in the literature – is expected to be strongest in polluted areas, and to be small or negligible in areas with less anthropogenic pressure.

For this study, the "weekend effect" is accounted for by extracting from each time series the couples of Sunday and Monday dates for which the 48 values of hourly data are available. Three values are then calculated: the daily mean, the daily maximum, and the daily standard deviation. If at least 20 couples of Sunday and Monday data are available, three indicators result from the ratio of the average Sunday values divided by the average Monday values.

To conclude this section, Fig. 3 illustrates the procedure applied to each time series. For each pollutant, each site, and provided that

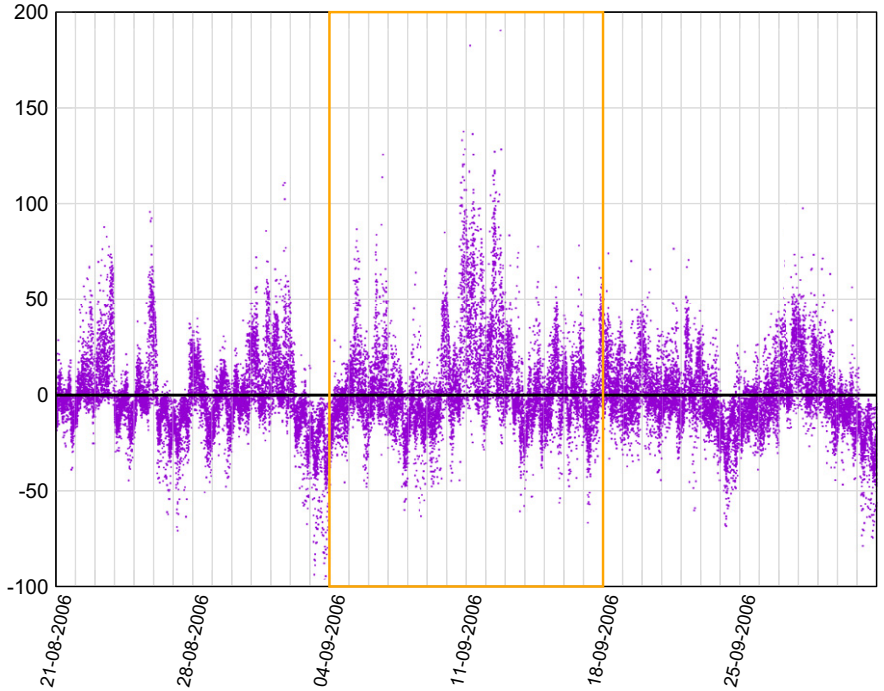

Fig. 2. Superposition of the hourly NO_2 concentrations measured over the Paris area (all sites confounded). The orange box highlights the period considered in Fig. 1.

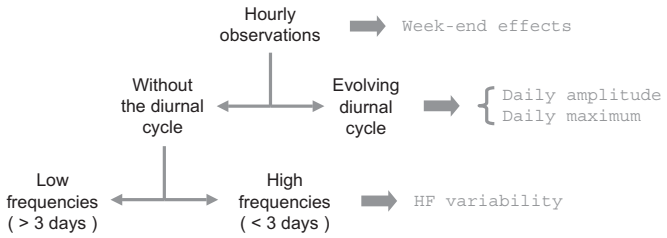

Fig. 3. Schematic illustration of the time series treatment for the definition of the 8 indicators.

the amount of data is sufficient, eight indicators are derived from each time series, and are expected to cover the main features of the measured pollutant behavior.

3. Statistics on the obtained indicators

Now that each hourly time series is characterized by eight variables, let us scrutinize the obtained values before going further into the classification procedure.

3.1. Behavior of the raw values

In order to describe the distribution of the values obtained for each indicator and each pollutant, Table 3 simply gives the traditional first and last quartiles of the distributions obtained. Some indicators – especially the winter *versus* summer ratios and the so-called "weekend effect" – bring us to some comments:

- Nitrogen oxides (NO_2 and NO): the maximum of the diurnal cycle is significantly lower in summer (almost two times lower for the NO) than in winter. This is related both to the lower emissions (domestic heating in winter and no excessive use of air conditioning in Europe) and to meteorological conditions (stronger convection and thicker boundary layer on average). The "weekend effect" is also significant, and is associated with lower concentrations on Sundays (about three quarters for NO_2 and half of Monday values for NO), in part due to the lower traffic emissions. Note that those values do not take into account the type of station and include both rural (*a priori* less affected by the weekend effect) and urban sites, which potentially reduces the overall signal.
- Sulfur dioxide: conclusions are about the same as for nitrogen oxides, with lower concentrations in summer, and lower concentrations during weekends. According to Bigi and Harrison (2010), such a behavior of SO_2 concentrations might be explained in part by a decrease in emissions from power plants due to a smaller energy demand (i.e. an annual cycle and a weekly cycle of energy demand).
- Particulate matter (PM_{10}): in most of the sites, the maximum value of the diurnal cycle is significantly lower in summer. This has however no discernible impact on the amplitude of the diurnal cycle. The "weekend effect" is significant: Sunday values are about 10% lower than Monday values. Those results agree with Bigi and Harrison (2010), who focus on an urban site.
- Ozone (O_3): unsurprisingly, the diurnal cycle has a higher maximum in summer, and a higher amplitude due to photochemistry. The "weekend effect" is significant, especially for mean values: in half of the sites, ozone concentrations are increased by 5%–15% on Sundays relatively to Mondays. As described in the literature (Cleveland et al., 1974; Blanchard and Fairley, 2001; Jenkin et al., 2002; Atkinson-Palombo et al., 2006; Sadanaga et al., 2008; Stephens et al., 2008; Tonse et al., 2008; and many others), the main reason is the weaker ozone titration by NO, resulting in higher ozone concentrations on weekends.

Despite the fact that all types of observation sites are mixed in Table 3 (from the most remote rural to the most polluted traffic site), those preliminary findings are consistent with the literature, and show that the eight indicators yield a realistic picture of the main known features of the pollutants behavior in Europe.

3.2. Data transformation and outliers detection

Table 3 gives some insight on the range covered by the indicators, but is insufficient to depict the distributions. Some distributions are indeed strongly non-normal. For some indicators and some pollutants, the distribution is skewed to the right, and looks more like a lognormal distribution. Therefore, it has been decided to proceed to a log-transformation of some indicators before further analysis. This is the case for the weekend effect on the daily means of ozone, and for three of the indicators computed for NO_2, NO, SO_2 and PM_{10}: the diurnal cycle maximum, the diurnal cycle amplitude, and the high frequency standard deviation. It has been verified (not shown), that all of those indicators are better distributed after a logarithm transformation (on a decimal basis). As expected, the log-transformation reduces the highest values and spreads out the small values.

Along with the distribution shape, data outliers can also strongly influence the estimation of the group centers and covariances in multivariate analysis (Reimann et al., 2008). Outliers have thus to be removed prior to the analysis, which is a tricky task as far as multivariate data is considered. Since we deal with a large data set, an automated procedure was needed, and we adopted the simplest approach, by discarding one per cent of the data at the upper and lower ends of the distributions (for each pollutant and each indicator). Such a procedure cannot distinguish between the extreme values of the distribution and the spurious outliers of the data. However, for our classification purpose, extreme values are not at stake and can be discarded without further care.

Table 3
For each pollutant, first and last quartiles of the distribution of the 8 indicators. "W/S" stands for Winter/Summer ratio.

	NO_2 2499 sites		NO 1993 sites		SO_2 1708 sites		PM_{10} 1647 sites		O_3 1942 sites	
	Q25	Q75	Q25	Q75	Q25	Q75	Q25	Q75	Q25	Q75
Diurnal cycle Maximum ($\mu g\ m^{-3}$)	26.3	53.1	13.7	53.3	4.7	12.7	27.5	43.3	63.9	78.2
Diurnal cycle Maximum W/S	0.68	0.878	0.431	0.7	0.647	1.02	0.784	1.01	1.48	1.99
Diurnal cycle Amplitude ($\mu g\ m^{-3}$)	14.6	30.9	10.9	45.3	2.23	7.8	10.2	21.7	32.2	48.7
Diurnal cycle Amplitude W/S	0.732	1.03	0.431	0.73	0.668	1.18	0.779	1.18	1.76	2.62
Weekend Effect on the daily Mean	0.697	0.806	0.428	0.66	0.865	0.982	0.867	0.966	1.04	1.14
Weekend Effect on the daily Maximum	0.714	0.824	0.359	0.564	0.824	0.978	0.836	0.964	1.01	1.07
Weekend Effect on the Standard Deviation	0.671	0.817	0.337	0.521	0.772	0.968	0.798	0.971	0.945	1.03
High Frequency Standard Deviation ($\mu g\ m^{-3}$)	8.44	14.3	8.89	25.8	2.21	7.94	9.75	16.5	14.2	16.8

The different variables cover different ranges of values. Before to proceed to multivariate statistical analyses (e.g., Principal Component Analysis, PCA) all variables have been rescaled. Here, the indicators have been first centered relatively to the median, and then divided by the inter-quartile distance. The median and the quartiles are more robust than the mean and the standard deviation when dealing with distributions with substantially different shapes, and a sizable proportion of strong values (Reimann et al., 2008).

3.3. Use of the meta-data

AirBase and BDQA data sets provide precious meta-information that qualifies the sites in a specific and thus rather subjective way. The AirBase meta-data describes both the station type (traffic, industrial, or background) and the station area (urban, suburban, or rural), whereas the French BDQA meta-data has a unique and merged typology (traffic, industrial, urban, suburban, or rural).

This meta-data is essential for the further analysis. For the sake of clarity, it has been decided to simplify AirBase meta-data, as is done in the BDQA data set. In the following, the classes R, S and U refer to AirBase "background" sites that are respectively "rural", "suburban" or "urban". The class T refers to the stations that are "traffic" in an "urban" or "suburban" area. Finally, the class "O" regroups the stations without meta-data, the stations with extreme values of the indicators (cf. the previous paragraph on the outliers), and the types that are not straightforward to categorize, such as industrial sites (that can be located in rural, suburban, or urban areas), or the traffic sites that are in a rural environment.

Table 4 shows the number of sites for which data were sufficient to compute the 8 indicators over the period 2002–2009. Note that depending on the pollutant, the number of rural sites varies quite a lot relatively to the number of traffic or urban sites. The disequilibrium is particularly strong for NO_2, NO, and PM_{10}, that have two or three times more traffic sites than rural ones.

3.4. Preliminary analysis

In order to go deeper into an analysis of the data, a graphical representation could be helpful. However, given the number of variables, it is not straightforward to represent the full data set in a synthetic way. The number of dimensions has thus to be reduced, for example using a Principal Component Analysis. Here, PCA is used as an illustration, and as a first step towards further analysis.

In an ideal world, we would expect the first Principal Component (PC) to gather most of the total variance, and to reflect the separation between the different types of sites (from the most rural to the most polluted). This is however not the case. Fig. 4 shows that the first PC accounts for 39%–49% of the total variance depending on the pollutant. The second and third PCs also explain significant amounts of the total variance (17%–33%).

Table 4
For each meta-data group, number of sites for which the analysis is computed.

	AirBase site	AirBase area	BDQA	O_3	NO_2	NO	PM_{10}	SO_2
T	Traffic	Urban/Suburban	Traffic	312	719	568	442	337
U	Background	Urban	Urban	611	748	552	532	518
S	Background	Suburban	Suburban	374	349	302	213	217
R	Background	Rural	Rural	422	294	235	165	215
O	Others	Others	Others	223	389	336	295	421
Total				1942	2499	1993	1647	1708

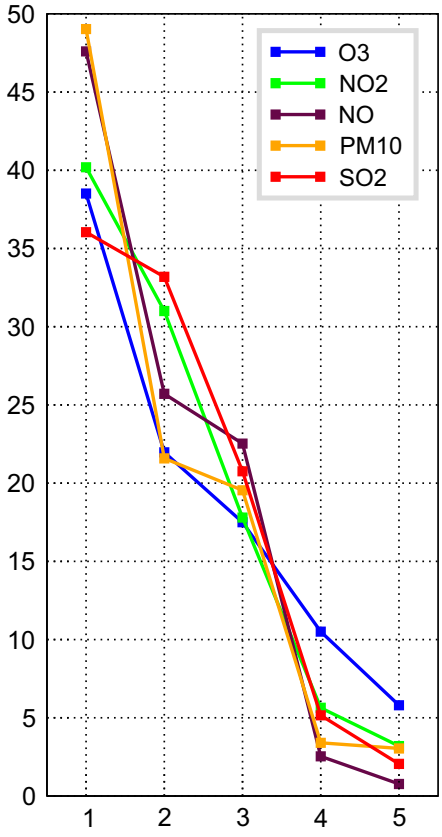

Fig. 4. Percentage of the total variance explained by the first components of a Principal Component Analysis of the data.

Fig. 5 displays the projection of the data on the first two principal components. Some interesting comments can be put forward:

- Apparently, the data – as described by the 8 variables – is not made of well-separated groups. The R, S, U and T groups based on the meta-data partly overlap each other in Fig. 5.
- The R and T groups based on the meta-data are clearly at the margins of the ensemble. On the contrary, U and S sites constitute the core of the ensemble. Obviously, with the defined indicators, it will be hard to distinguish between those U and S types. This suggests that *from the point of view of the past measurements* the urban/suburban distinction found in the meta-data is not necessarily justified.
- The intra-group variance can be quite different depending on the pollutant. E.g., for the nitrogen monoxide, rural stations are scattered, with a wide range of values, whereas traffic stations form a dense group of points.
- The groups based on the meta-data are best separated with NO_2 and NO data, with rural sites being clearly apart from the other sub-groups based on the meta-data.
- The first principal component is not necessarily the direction that best discriminates between the rural and the more polluted sites (e.g. for the SO_2). This shows that the criterion that underlies a PCA (maximizing the explained variance) might not be – in our case – the optimal criterion to synthesize the information and classify the sites.

Because there is a continuum regarding the statistical properties of the sites, clustering procedures can only be rather an arbitrary grouping of sites. The goal of the following classification is

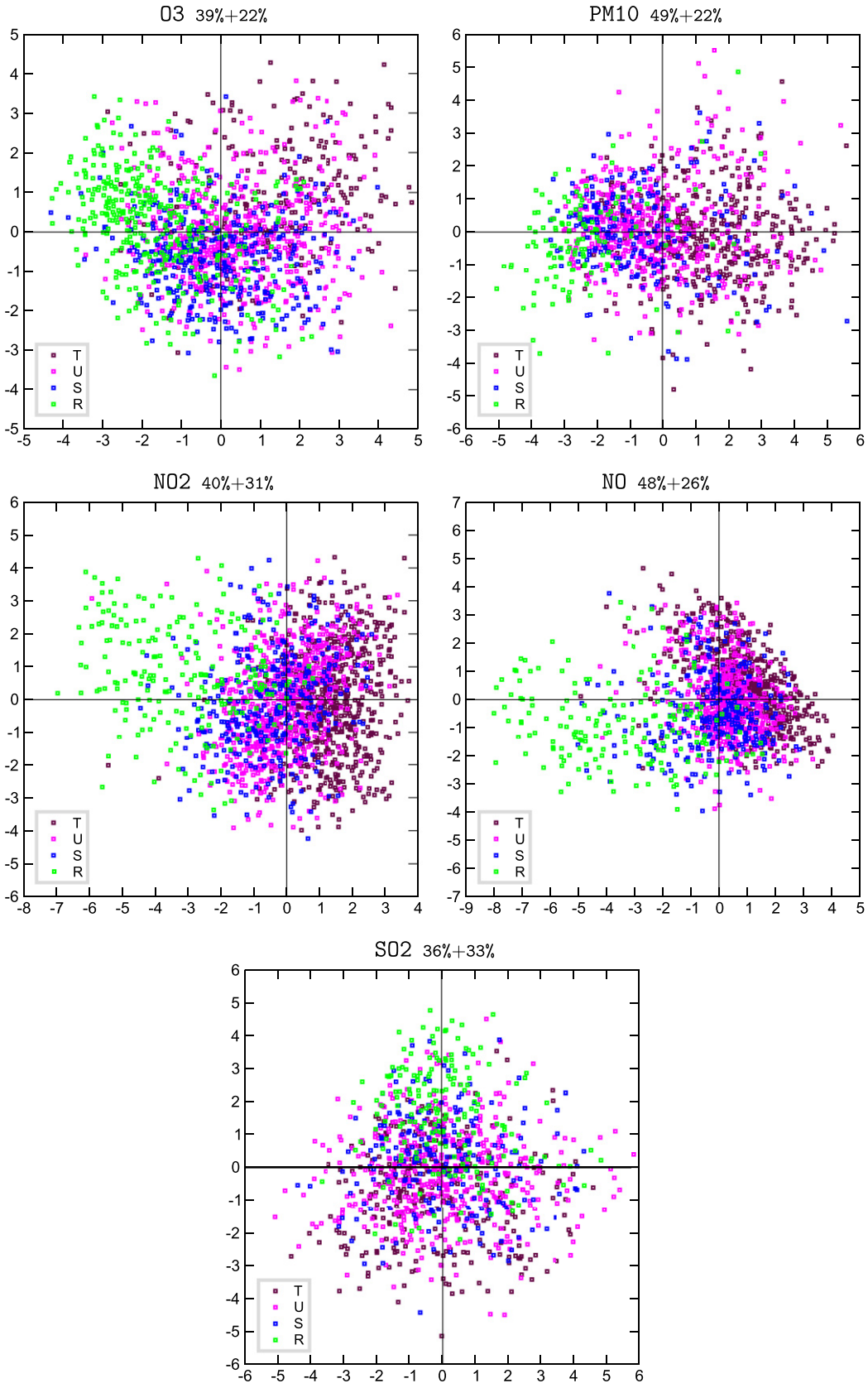

Fig. 5. Projection of the data on the two first components of a PCA, and percentage of the total variance explained by the two first components.

therefore to stratify the sites according to the behavior of the defined indicators.

4. The classification procedure

Now that each pollutant time series is characterized by 8 variables that gather the key information, the goal in this section is to build – for each pollutant – an objective classification of the observation sites.

4.1. The Linear Discriminant Analysis

A number of classification methods are available today in the literature. The challenge is to find a method that makes the best use of all the available data, eventually including meta-data. Cluster analysis is not recommended, since there are apparently no clear clusters in the data (Fig. 5). Instead of trying to discriminate fully separate groups, we can try to "stratify" the sites on the basis of the 8 indicators, with the constraint to keep sufficiently consistent with the meta-data groups.

A way forward is to combine linearly the 8 indicators (some of these being maybe redundant or even irrelevant to the classification), in order to provide an index of the "rurality" or "pollutedness" of the sites. This is a typical issue of dimensionality reduction, for which different statistical tools are possible. PCA is the most common, but it has been shown in the previous section that the criterion of maximum variance cannot apply fully in our case. Since we want the classification to be as consistent as possible with the meta-data, while having its own coherence based on the past measurements, a statistical tool that seems relevant is the Linear (or Fisher) Discriminant Analysis (LDA).

The LDA (Fisher, 1936) uses an *a priori* knowledge about existing group memberships (here, the meta-data). This knowledge is used to develop a function that will result in an optimal discrimination of the groups. Here, we need to separate the rural sites (R), that are the most representative of the large scale (and can be therefore used for model validation or data assimilation), from the most polluted (T and U that have low spatial representativeness). The group S has been discarded at this stage of the analysis, because it overlaps the group U (Fig. 5) and may not facilitate the discrimination between rural and polluted sites.

4.2. Cross-validation, classification, and comparison with the meta-data

The LDA seeks a linear combination of the variables that has a maximal ratio of the separation of the class means to the within-class variance (Reimann et al., 2008). The obtained discriminant function makes it then possible to classify any new time series as belonging to one of the groups based on the 8 indicators alone. With the existing data, this makes possible to test the accuracy of the statistical model by cross-validation. The idea is to repeatedly split the data set into training and test subsets, and repeat the LDA many times. As the "actual" membership of training set members is known, a table can be prepared showing the number of times observations are correctly classified or misclassified. Table 5 reveals that the LDA is particularly efficient with NO_2 and NO data. The error ratio is greater for the other pollutants, especially for SO_2 (reaching 15%).

Since R sites are much less numerous than U + T sites, it is interesting to focus on the R group misclassifications. In the case of PM_{10} and SO_2, more than half of the R sites are misclassified. This supports the findings of the PCA (Fig. 5): there is no clear separation between the meta-data groups. For most of the pollutants here considered, a classification in a limited number of clusters seems therefore inappropriate.

Table 5
Cross-validation of the LDA results. The "Error Ratio" is the number of misclassifications divided by the total number of sites.

Observed → / Predicted ↓	O_3		NO_2		NO		PM_{10}		SO_2	
	R	U + T	R	U + T	R	U + T	R	U + T	R	U + T
R	248	23	164	19	128	13	63	27	78	36
U + T	99	799	50	1353	50	1039	63	874	104	746
Error Ratio	10%		4%		5%		9%		15%	

The LDA yields a linear combination of the 8 indicators that best discriminates the rural stations from the others. Therefore, in order to go further than simply reallocating to the R and U + T groups (as in the cross-validation above), it is tempting to forget about the groups based on the meta-data, and to use the projection on the Fisher axis as an index to define new classes. Using the nine percentiles from 10% to 90% as fixed thresholds, ten classes have been defined (this arbitrary number of classes is justified in Section 4.3). Note that with this simple method, any new time series can be (i) characterized by the 8 indicators, (ii) projected on the Fisher axis, and (iii) classified using the fixed thresholds.

Fig. 6 shows the distribution of the obtained classes for each sub-group of the meta-data. This yields a more refined picture of the LDA than in Table 5, and it is still consistent with the results of the PCA and of the cross-validation. Let us describe Fig. 6 for each sub-group of the meta-data:

- Group R: with the pollutants NO_2, NO, and PM_{10}, respectively 95%, 94%, and 89% of the R group falls into classes 1 to 3. For ozone, 63% of the R group falls into classes 1–3, 17% into classes 4–6, and 20% into classes 7–10, which is consistent with the fact that some stations qualified as "rural" in the meta-data are in fact located not far downwind of emission sources, and thus affected by ozone pollution through transport. This is another reason why a classification based on the past measurements may be in certain cases more objective than based on purely local information. Finally, concerning SO_2, 77% of R sites fall into classes 1–3, and 15% into classes 4–5.
- Groups S and U: for O_3 and SO_2, most of the S and U sites are shared among the classes 3 to 9. In the case of SO_2, a significant proportion of S sites fall into classes 1 and 2, maybe in connection with some recent changes in SO_2 emissions that have not been taken into account in the meta-data yet. With the other pollutants, the behavior of the S and U sites is slightly different: U sites are concentrated in classes 4–8, whereas S sites mostly fall into classes 2–6. This confirms that a full separation of S and U groups is not possible statistically based on our 8 variables solely. However, the projection on the Fisher axis helps stratifying urban and suburban measurements in a rather continuous manner.
- Group T: with NO_2, NO, and PM_{10}, respectively 74%, 67%, and 61% of the traffic sites fall into classes 8–10, whereas this is the case for 54% and 47% of the sites with O_3 and SO_2. For those two species, the T and U groups overlap each other. Note that for the LDA computation, the U and T sites have been grouped (cf. §4.1), so the Fisher axis was not constructed specifically to separate the most urban and the traffic sites.

Another way to analyze the results is to scrutinize Figs. 7 and 8. Given the size of the domain, only large-scale structures can be commented.

- Ozone (O_3): over the north-eastern part of Europe (especially the north of Germany, Czechia, Poland, Lithuania, and the countries around the Baltic), quite a lot of S, U, or even T sites

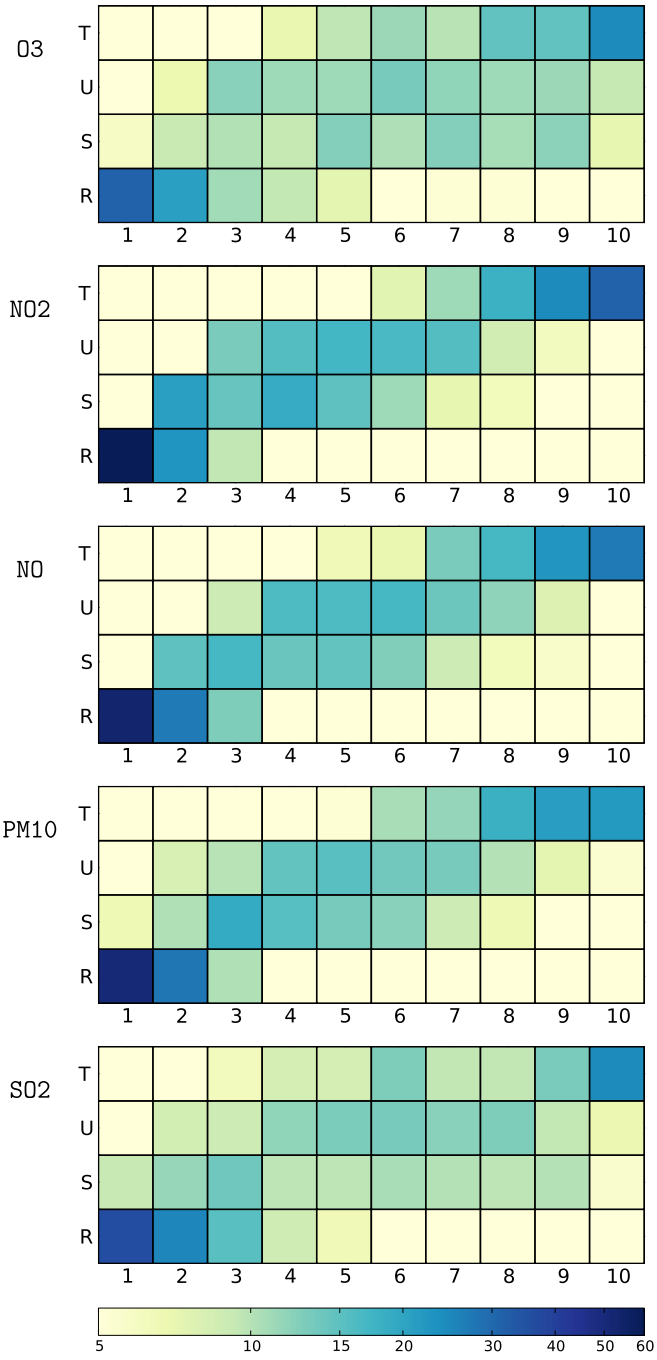

Fig. 6. Comparison between the obtained classes and the groups based on the meta-data. For each group, the colour corresponds to the percentage of stations falling in each class (logarithmic colour scale).

(i.e., quite polluted according to the meta-data) fall into the lowest classes 1–3. On the contrary, over the Netherlands, the south of Germany, the north of France, and the north of Italy, many R sites fall into the highest classes, probably due to the transport of polluted air-masses in those densely populated countries. A rural site (according to the meta-data) in the Netherlands and a rural site in Scandinavia can thus exhibit very different statistical behaviors, that are distinguished by the objective classification.

- Nitrogen dioxide and monoxide (NO_2 and NO): in eastern Germany (especially the region of Leipzig), Czechia, Slovakia, and Poland, a lot of S, U, and T sites fall in the lowest classes, as for ozone. This is also true for France, where the classification tends to reduce the "polluted" character of urban sites, compared to the meta-data. The objective classification seems to act as a "recalibration" of the meta-data of some European countries.
- Particulate matter (PM_{10}): there is an overall good agreement with the meta-data. The regions with the most marked discrepancies with the meta-data are the north of Germany and Finland, where most of the S, U, and T sites fall in the classes 1–3. This is also the case for several sites in France and England. As for NO_2 and NO, the objective classification tends to moderate the polluted character of the meta-data U and T groups.
- Sulphur dioxide (SO_2): as for PM_{10}, there is an overall good agreement between the classification and the meta-data, except for some regions of Germany. The 10 classes of the classification yield a more contrasted picture than the 5 types derived from the meta-data. The "extreme" classes (1–3 and 8–10) tend to be more represented than with the meta-data (R and T types), which is a very useful information regarding high or respectively low representativeness.

Overall, our statistical classification contributes to some large-scale reorganization of the sites characteristics, while keeping a general broad agreement with meta-data based information. This agrees with the *a priori* statement that the meta-data is largely subjective and inhomogeneous. As required, the classification is pollutant-specific, and some differences between the classes obtained for the different pollutants can already been seen right from Figs. 7 and 8. For example, the Netherlands is characterized by high classes for ozone, but not so much for the other pollutants. On the contrary, Poland exhibits low classes for ozone, but higher classes as far as PM_{10} and SO_2 are concerned.

4.3. Robustness of the classification

The approach chosen in this study is that the classification is conducted for each species *independently* in order to account for the different configurations of the measurement sites, and especially the influence from emission sources that can be quite different depending on the pollutant. Nevertheless, on average and at the large-scale, we expect the classifications obtained for the different pollutants to be "sufficiently" consistent.

Fig. 9 compares the classes obtained for the stations that measure two (or more) pollutants. The consistency between the classes depends on the pollutants compared:

- With NO and NO_2, the agreement between the classes is particularly good. For 74% of the sites, differences are lower or equal to 1, and for 96% of the sites differences are lower or equal to 3. Since NO and NO_2 are linked to the same emission sources, such a consistency is quite satisfactory.
- The coherence of the nitrogen oxides (NO_2 and NO) with the particulate matter PM_{10} – probably dominated by primary emissions – is also quite good, with more than 89% of the sites that have differences inferior or equal to 3.
- For the remaining comparisons (curves of Fig. 9) that involve either ozone or SO_2, differences are slightly more marked. For ozone, this may be due to the influence of the long-range transport and of the photochemistry, while for SO_2 this may be due to the emissions sources, that can be different from the NO_x or PM_{10} (e.g., industrial sites). However, it should be noted that for all of the pollutants, differences are equal or inferior to 3 for at least 76% of the sites.

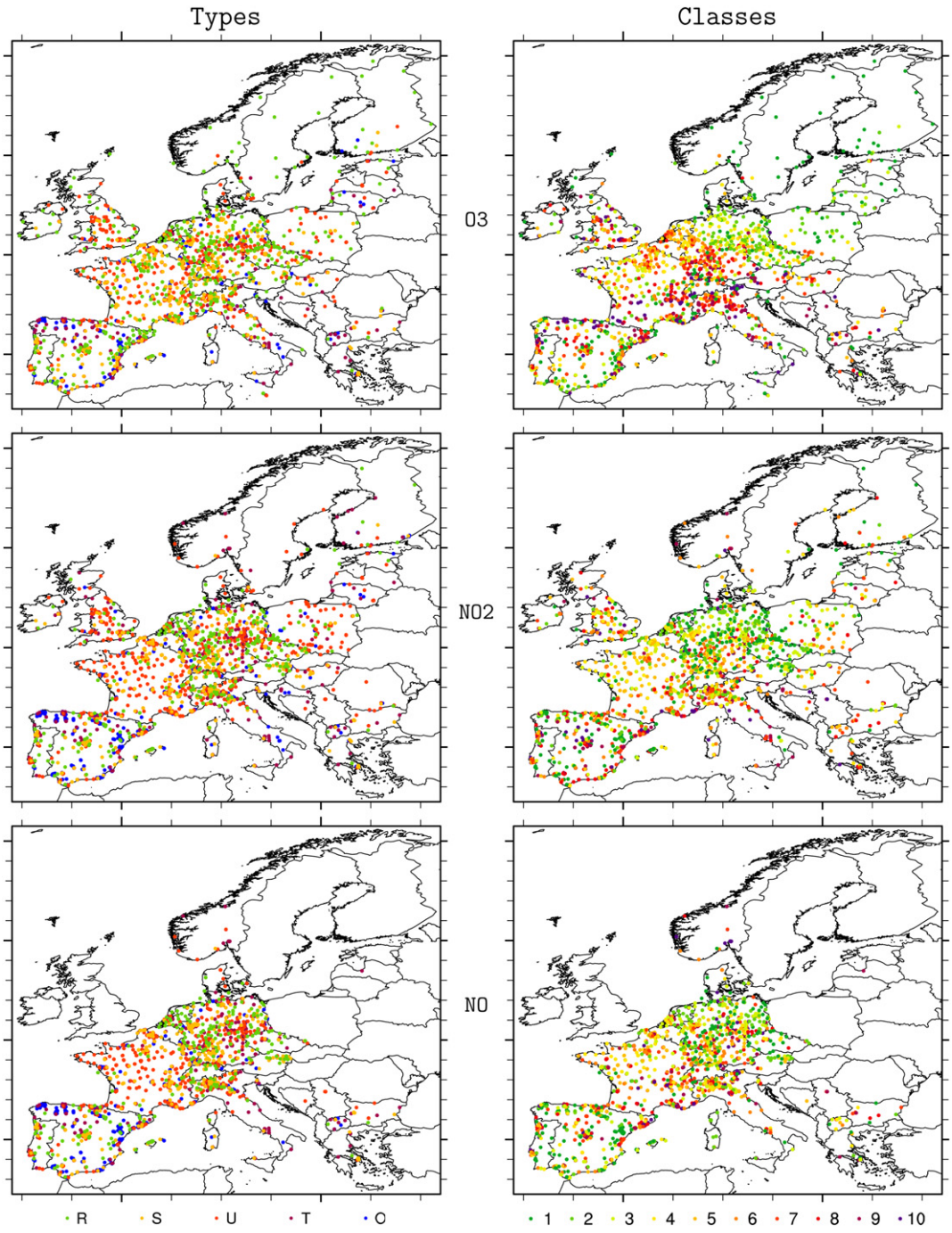

Fig. 7. Comparison between the station types based on the meta-data and the classes of the objective classification.

The good consistency between the classes obtained for the different species is a satisfactory result, but it is not sufficient to be in a position to conclude that the classification procedure is robust. A better benchmark requires an independent data set. In the absence of such data at the European scale, it has been chosen to proceed to the classification of the same sites, but for two distinct groups of years: the years 2003, 2005, 2007, and 2009 (that include the summer 2003 heat-wave), and the years 2002, 2004, 2006, and 2008. For the stations that have enough data in both subsets, the classes have been computed and are compared on Fig. 10. For NO_2 and NO, there is a remarkable agreement between the classes computed for both periods. Differences are equal or lower than 1 for more than 96% of the sites. The robustness is substantially weaker but still very similar for the three other pollutants (O_3, PM_{10}, and SO_2). For more than 80% of the sites the differences between the classes computed for the two distinct subsets of years are lower or equal to 2.

We have chosen to stratify the results of the LDA into ten classes, which is arbitrary. Fig. 10 shows that such a precision is justified in the case of the NO and NO_2 classifications. However, with O_3, PM_{10}, and SO_2, such a precision is not needed. For those species, the users of the classification will probably choose to group some of the classes. One way to do so might be to group classes two by two, which would lead to 5 classes. Another way might also be to rely on

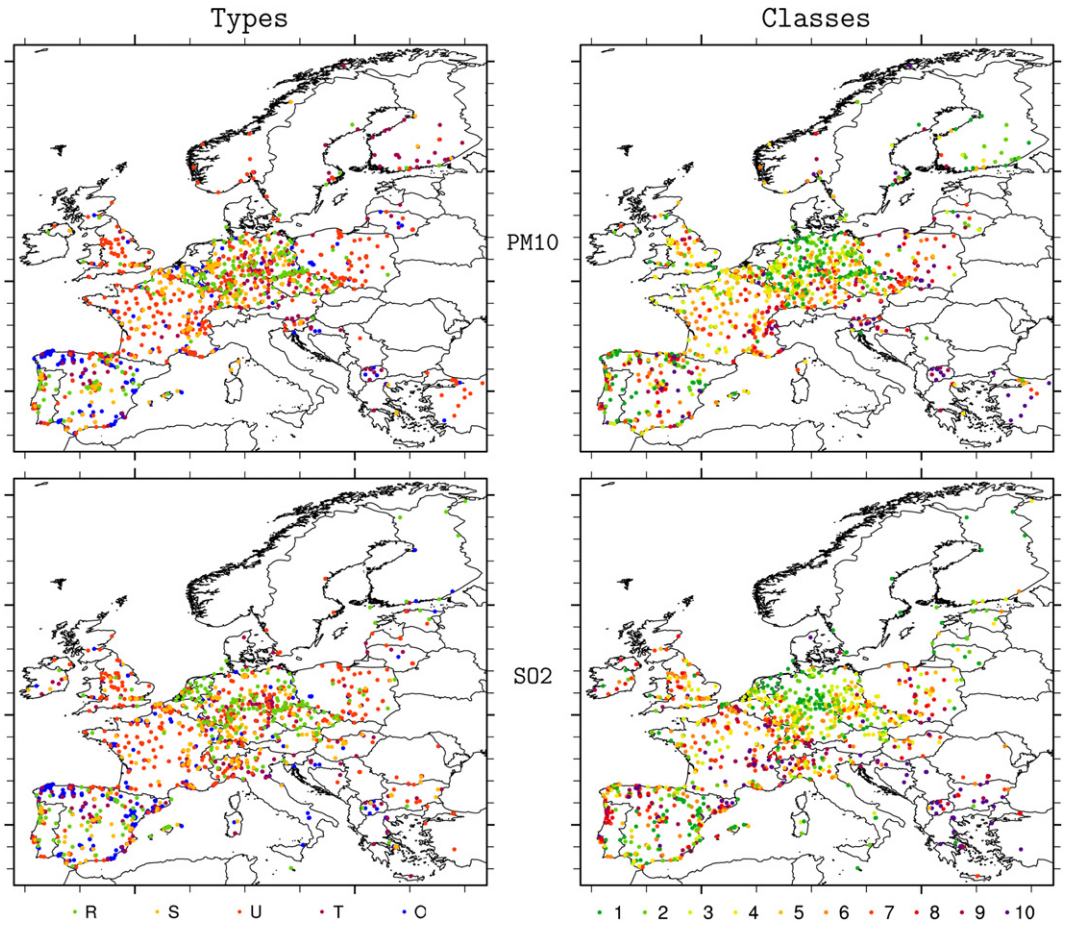

Fig. 8. Comparison between the station types based on the meta-data and the classes of the objective classification.

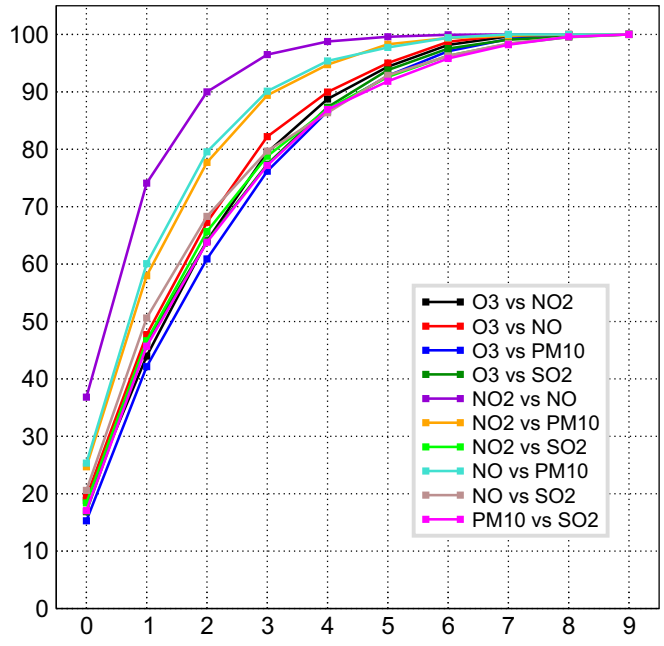

Fig. 9. Cumulative frequency diagram of the differences between the obtained classes, when two pollutants are measured at the same site.

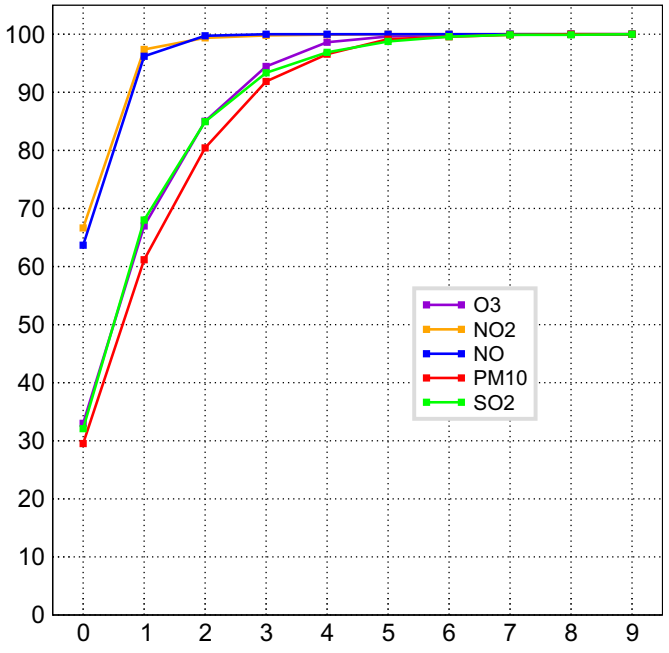

Fig. 10. Cumulative frequency diagram of the differences between the classes obtained for the two subsets of years: 2003–2005–2007–2009 and 2002–2004–2006–2008.

the results shown on Fig. 6, in order to define some "super-classes", that provide best consistency with the meta-data (e.g., 1–3, 4–7, and 8–10), if the meta-data is thought to be reliable enough.

5. Discussion and conclusion

Air quality at the surface is strongly heterogeneous, due to variable surface fluxes (emissions, surface deposition, etc), meteorological conditions, and diverse local configurations (city buildings, valleys, etc). Depending on the purpose of the measurements, monitoring sites have been located in very diverse environments, from highly populated areas (to estimate exposure), to remote areas (to assess background conditions and contribution of long-range transport of pollutants). In each region, the sites are categorized on the basis of criteria that are largely subjective, which results in inhomogeneous meta-information at the scale of Europe. In the framework of the European MACC project (Monitoring Atmospheric Composition and Climate), there was a need to classify measurement sites with a robust approach concerning representativeness, in order to perform European-wide verification and data assimilation activities.

Most papers that had tackled this issue so far were based on limited data sets or periods (see Table 1). Besides, those studies often took into account additional data such as population density, land cover, or emission estimates… As discussed in Spangl et al. (2007), the inclusion of too many parameters in the classification process might lead to an over-categorization of the sites, with too many sub-groups for practical interpretation and use for model verification or in data assimilation systems. We have shown here that it is possible to classify air quality sites, based only on their past time series. The advantage is that such a classification can be easily updated whenever required, and that any new site can also be classified *a posteriori* without re-running the Linear Discriminant Analysis (provided it has a sufficiently long record of measurements).

The proposed classification is pollutant-specific, which means that different classes can be obtained for the different pollutants monitored at a same site. From a technical point of view, this is of course the simplest approach since the monitoring sites do not measure the same list of pollutants. This is also motivated by the fact that emissions, chemistry, deposition and processes are pollutant-specific. The largest data set available for Europe has been used (AirBase) and complemented over France with national data (BDQA). The pollutants investigated over the 2002–2009 period are: O_3, NO_2, NO, SO_2, and PM_{10}.

Time series are firstly filtered in order to extract some essential statistical features such as the diurnal cycle, the so-called "weekend effect", and the high frequency variability (periods lower than 3 days). Eight indicators are defined, taking also into account the summer/winter differences in the characteristics. Because of the skewness in their distribution, some indicators have been log-transformed. To best benefit from the information held by the eight indicators, while being as much coherent with the meta-data as possible, a Linear Discriminant Analysis has been then computed for each pollutant with the scope to best separate rural and urban sites. After projection on the Fisher axis, ten classes have been determined on the basis of fixed percentile thresholds.

The objective classification has been first validated by cross-validation. Best scores are obtained with NO_2 and NO data. The consistency with the available meta-data is lower with the remaining pollutants, especially SO_2, which is not surprising because the LDA is designed to distinguish rural and urban sites (that depend mostly on domestic and traffic emissions), while SO_2 concentrations are more linked with industrial emissions. Concerning ozone, the consistency with meta-data is rather good, despite the fact that ozone concentrations are strongly influenced by long-range transport (which is not accounted for by the meta-data). With the 8 defined variables, the separation between the so-called "suburban" and "urban" sites of the meta-data seems rather arbitrary, with a continuum of the "polluted" character. Across Europe, the objective classification has a visible impact on the large-scale distribution of the sites.

The objective classification has been then challenged by comparing the classes obtained for different pollutants at the same site, and by comparing the classes obtained for the same sites over two distinct periods. A broad agreement is found between the classes obtained for the different pollutants, which is not particularly surprising. Due to the differing lifetimes, a site can be very little polluted according the short-lived NO or NO_2, while under the influence of polluted plumes for the longer-lived pollutants like ozone.

Working with air quality data at the scale of Europe is far from straightforward, because measurement configurations are likely to represent very different environments. In other words, measurement sites often respond to certain local needs of air quality monitoring, but the use of the network aggregated over Europe (e.g., in AirBase data set) is not straightforward. In the framework of the MACC project, a central purpose of the present classification is to facilitate the validation of model outputs by grouping monitoring sites into classes that are homogeneous in their statistical characteristics. By construction, the classification proposed is pollutant-specific. The comparison with meta-information and the robustness of the results are quite satisfactory. Classes from different countries and at different latitudes can now be compared directly.

The practical use of this classification in the framework of the MACC project is to discard the sites with highest classes for use in verification or data assimilation. The underlying idea is that a classification based on the measurement characteristics is the best way to assess overall "representativeness". And yet, data assimilation procedures may be improved by selecting the monitoring sites that are representative of geographical areas related to the spatial resolution of the models. This has however to be tested and requires further work.

Beyond the issue of representativeness, we believe that any statistical use of air quality data, such as in model output statistics, mapping, trend analysis, network optimization, or automatic quality control, benefits from a stable classification methodology for the observation sites.

Acknowledgments

We acknowledge the European Environment Agency for the AirBase data set, and the French ADEME (Agence de l'environnement et de la maîtrise de l'énergie) that maintained the BDQA database over the considered period. Thanks are also due to Bertrand Michel (Université Pierre et Marie Curie) for his helpful comments on the statistical tools used in this study.

References

ADEME (Agence de l'Environnement et de la Maîtrise de l'Energie), 2002. Classification and criteria for setting up air-quality monitoring stations Technical Report. http://www2.ademe.fr.

Atkinson-Palombo, C.M., Miller, J.A., Balling, R.C., 2006. Quantifying the ozone "weekend effect" at various locations in Phoenix, Arizona. Atmospheric Environment 40, 7644–7658.

Beirle, S., Platt, U., Wenig, M., Wagner, T., 2003. Weekly cycle of NO_2 by GOME measurements: a signature of anthropogenic sources. Atmospheric Chemistry and Physics 3, 2225–2232.

Bigi, A., Harrison, R.M., 2010. Analysis of the air pollution climate at a central urban background site. Atmospheric Environment 44, 2004–2012.

Blanchard, C.L., Fairley, D., 2001. Spatial mapping of VOC and NO_x-limitation of ozone formation in central California. Atmospheric Environment 35, 3861–3873.

Cleveland, W.S., Graedel, T.E., Kleiner, B., Warner, J.L., 1974. Sunday and workday variations in photochemical air pollutants in New Jersey and New York. Science 186, 1037–1038.

Fisher, R.A., 1936. The use of multiple measurements in taxonomic problems. Annals of Eugenics 7, 179–188.

Flemming, J., Stern, R., Yamartino, R.J., 2005. A new air quality regime classification scheme for O_3, NO_2, SO_2 and PM_{10} observations sites. Atmospheric Environment 39, 6121–6129.

Gramsch, E., Cereceda-Balic, F., Oyola, P., von Baer, D., 2006. Examination of pollution trends in Santiago de Chile with cluster analysis of PM_{10} and ozone data. Atmospheric Environment 40, 5464–5475.

Hamming, R.W., 1977. Digital Filters. Prentice-Hall, ISBN 0132125714.

Henne, S., Brunner, D., Folini, D., Solberg, S., Klausen, J., Buchmann, B., 2010. Assessment of parameters describing representativeness of air quality in-situ measurement sites. Atmospheric Chemistry and Physics 10, 3561–3581.

Hollingsworth, A., Engelen, R.J., Textor, C., Benedetti, A., Boucher, O., Chevallier, F., Dethof, A., Elbern, H., Eskes, H., Flemming, J., Granier, C., Kaiser, J.W., Morcrette, J.-J., Rayner, P., Peuch, V.-H., Rouil, L., Schultz, M.G., Simmons, A.J., 2008. Toward a monitoring and forecasting system for atmospheric composition: the GEMS project. Bulletin of the American Meteorological Society 89, 1147–1164.

Huijnen, V., Eskes, H.J., Poupkou, A., Elbern, H., Boersma1, K.F., Foret, G., Sofiev, M., Valdebenito, A., Flemming, J., Stein, O., Gross, A., Robertson, L., D'Isidoro, M., Kioutsioukis, I., Friese, E., Amstrup, B., Bergstrom, R., Strunk, A., Vira, J., Zyryanov, D., Maurizi, A., Melas, D., Peuch, V.-H., Zerefos, C., 2010. Comparison of OMI NO_2 tropospheric columns with an ensemble of global and European regional air quality models. Atmospheric Chemistry and Physics 10, 3273–3296.

Ignaccolo, R., Ghigo, S., Giovenali, E., 2008. Analysis of air quality monitoring networks by functional clustering. EnvironMetrics 19, 672–686.

Jenkin, M.E., Trevor, J.D., Stedman, J.R., 2002. The origin and day-of-week dependence of photochemical ozone episodes in the UK. Atmospheric Environment 36, 999–1012.

Kovač-Andrić, E., Šorgo, G., Kezele, N., Cvitaš, T., Klasinc, L., 2010. Photochemical pollution indicators-an analysis of 12 European monitoring stations. Environmental Monitoring and Assessment 165, 577–583.

Lau, J., Hung, W.T., Cheung, C.S., 2009. Interpretation of air quality in relation to monitoring station's surroundings. Atmospheric Environment 43, 769–777.

Monjardino, J., Ferreira, F., Mesquita, S., Perez, A.T., Jardim, D., 2009. Air quality monitoring: establishing criteria for station classification. International Journal of Environment and Pollution 39, 321–332.

Oltmans, S.J., Lefohn, A.S., Harris, J.M., Galbally, I., Scheel, H.E., Bodeker, G., Brunke, E., Claude, H., Tarasick, D., Johnson, B.J., Simmonds, P., Shadwick, D., Anlauf, K., Hayden, K., Schmidlin, F., Fujimoto, T., Akagi, K., Meyer, C., Nichol, S., Davies, J., Redondas, A., Cuevas, E., 2006. Long-term changes in tropospheric ozone. Atmospheric Environment 40, 3156–3173.

Reimann, C., Filzmoser, P., Garrett, R., Dutter, R., 2008. Statistical Data Analysis Explained. Ed. Wiley.

Romano, D., Cirillo, M., Coppi, R., D'Urso, P., 1999. Optimal design of air quality networks detecting warning and alert conditions. Statistical Methods and Applications 8, 61–73.

Sadanaga, Y., Shibata, S., Hamana, M., Takenaka, N., Bandow, H., 2008. Weekday/weekend difference of ozone and its precursors in urban areas of Japan, focusing on nitrogen oxides and hydrocarbons. Atmospheric Environment 42, 4708–4723.

Snel, S., 2004. Improvement of Classifications for AirBase ETC/ACC Technical Report.

Spangl, W., Schneider, J., Moosmann, L., Nagl, C., 2007. Representativeness and Classification of Air Quality Monitoring Stations Umweltbundesamt report.

Stephens, S., Madronich, S., Wu, F., Olson, J.B., Ramos, R., Retama, A., Muñoz, R., 2008. Weekly patterns of México City's surface concentrations of CO, NO_x, PM_{10} and O_3 during 1986–2007. Atmospheric Chemistry and Physics 8, 5313–5325.

Tarasova, O.A., Brenninkmeijer, C.A.M., Jöckel, P., Zvyagintsev, A.M., Kuznetsov, G.I., 2007. A climatology of surface ozone in the extra tropics: cluster analysis of observations and model results. Atmospheric Chemistry and Physics Discussions 7, 12541–12572.

Tonse, S.R., Brown, N.J., Harley, R.A., Jin, L., 2008. A process-analysis based study of the ozone weekend effect. Atmospheric Environment 42, 7728–7736.

Vingarzan, R., 2004. A review of surface ozone background levels and trends. Atmospheric Environment 38, 3431–3442.