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
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 "EoI" (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 "EoI" 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 EoI "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}