Atmospheric Research 138 (2014) 30–40

Contents lists available at ScienceDirect

Atmospheric Research

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

The application of hierarchical cluster analysis and non-negative matrix factorization to European atmospheric monitoring site classification

Christopher S. Malley [a,b,*], Christine F. Braban [a], Mathew R. Heal [b]

[a] NERC, Centre for Ecology & Hydrology, Bush Estate, Penicuik EH26 0QB, UK
[b] School of Chemistry, University of Edinburgh, West Mains Road, Edinburgh EH9 3JJ, UK

ARTICLE INFO

Article history:
Received 25 July 2013
Received in revised form 22 October 2013
Accepted 28 October 2013

Keywords:
Ozone
EMEP monitoring sites
Cluster analysis
Non-negative matrix factorisation

ABSTRACT

The effective classification of atmospheric monitoring sites within a network allows conclusions from measurements to be extrapolated beyond the confines of the site itself and applied to larger areas or populations. This is especially important for the European EMEP 'supersites' because these are relatively few in number yet are subject to much investment in composition monitoring capability. Here, the representativeness of the two UK EMEP supersites, Auchencorth and Harwell, was evaluated using the hierarchical cluster analysis (HCA) of all available EMEP monitoring sites based on measured ozone concentration datasets for the period 1991–2010. A novel feature was to apply non-negative matrix factorization (NMF) to order the sites within the HCA dendrograms according to the relative anthropogenic influence on ozone. The ordered dendrograms enabled UK sites to be placed more precisely in a European context. For 2007–2010, all 19 UK EMEP sites were assigned to two of the site classification clusters, with 17 of the sites grouping closely with each other in each cluster. Auchencorth clustered with the sites characterised by less modification of hemispheric background ozone levels, whilst Harwell grouped with the sites showing a more polluted regime. A similar grouping of sites occurred between 1991 and 2010, with relatively closer clustering of Polluted UK sites compared with Remote UK sites due to the larger, transboundary spatial domain for which the Remote UK sites are representative. This tight clustering of the majority of the other UK ozone monitoring sites with either one of the supersites, shows that UK background ozone conditions are well represented by Auchencorth and Harwell, and gives confidence that more extensive chemical climatologies developed for the two supersites will have wider geographical relevance.

1. Introduction

The European Monitoring and Evaluation Programme (EMEP, www.emep.int) provides governments with scientific information to inform policy regarding the long-range, transboundary transport of air pollution (Torseth et al., 2012). The programme has three core strands: collation of atmospheric emissions inventories; modelling of atmospheric transport and deposition; and measurement of atmospheric composition at locations where the impact of local pollutant emission sources should be low. The EMEP guidance (EMEP/CCC-Report 1/95) outlines methods intended to ensure that air sampled at a monitoring site is representative of air not directly affected by local emission sources. These include: 50 km from major pollution sources (towns, power plants, etc.), 2 km from the application of manure, and consideration of meteorological and topographical features. EMEP Level I sites are designed to capture basic atmospheric composition, whilst Level II and III sites (often referred to as EMEP supersites) measure a wider range of atmospheric constituents at higher time resolution than Level I (see Torseth et al. (2012)).

Monitoring sites in a network are usually classified into different groups that internally share similar chemical climatologies, i.e. similar atmospheric composition, drivers of that

* Corresponding author at: School of Chemistry, University of Edinburgh, West Mains Road, Edinburgh EH9 3JJ, UK. Tel.: +44 7578 725402.
E-mail address: C.Malley@sms.ed.ac.uk (C.S. Malley).

0169-8095/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.atmosres.2013.10.019

composition, and impacts due to that composition. A balance is required which captures the major variations in composition and drivers across the network but in as few groups as possible so as to retain the ability to generalise. Various grouping methodologies have been applied (Joly and Peuch, 2012). These range from the relatively subjective use of metadata (Spangl et al., 2007), traditionally used in monitoring networks, to more objective techniques such as rankings based on statistical indicators (Kovac-Andric et al., 2010), linear discriminant analysis (Joly and Peuch, 2012), principal component analysis (Lau et al., 2009), and non-hierarchical (Ignaccolo et al., 2008) and hierarchical cluster analyses (Flemming et al., 2005; Henne et al., 2010; Tarasova et al., 2007). The latter is a multivariate approach that encompasses many separation/agglomeration techniques which aims to identify natural groupings, or clusters, amongst objects in a dataset through minimisation of the within-cluster variance and maximisation of the between-cluster variance (Kaufman and Rousseeuw, 1990). Clustering methods require user-defined parameters which may impact the objectivity of the analysis. For example, a method for calculating 'distance' between individual members needs to be specified (Dabboor et al., 2013), as must a method for calculating the separation between different groups of members (Mangiameli et al., 1996). Nevertheless, as cited above, clustering techniques have been widely applied to grouping atmospheric monitoring sites.

The aim of this study was to assess whether the locations of the two Level II UK supersites, at Auchencorth in south-east Scotland and Harwell in southern England, are representative of UK background conditions, even though they do not fully meet the EMEP criteria for non-locally influenced "background" sites (this is the case at other EMEP sites, as acknowledged in Torseth et al. (2012)): Auchencorth is located 17 km from Edinburgh, although prevailing winds mean that it is predominantly upwind from the city, whilst Harwell is 7 km from a 1360 MW natural gas power station. Effective site classification is particularly important for EMEP Level II 'supersites' because these are considerably few in number yet subject to much investment in composition monitoring capability.

In this work, sites across the EMEP domain were classified according to the annual and daily patterns in ground-level ozone concentrations. Ozone was chosen for two reasons. First, it is the most widely measured constituent across the EMEP network — between 2007 and 2010, 113 sites measured hourly ozone concentrations and 49 sites have continuous ozone time series since 1991. Second, measured ozone concentrations are a result of the combination of a wide variety of drivers which are also relevant to many aspects of atmospheric composition, including precursor emissions, photochemistry, deposition, meteorological and climatic conditions and long-range transport (AQEG, 2009; Royal Society, 2008). A major driver of temporal ozone variation is hemispheric background concentrations (AQEG, 2009). Regional and local-scale processes lead to modification of these values. Under suitable conditions, efficient photochemical processing of NO_x and volatile organic compounds (VOCs) lead to additional ozone formation and high ozone episodes, whilst local-scale depletion of ozone occurs due to reaction with NO, an effect which increases with higher NO concentrations (Jenkin, 2008).

Hierarchical clustering was applied to the monthly-diurnal ozone concentrations (average diurnal cycle for each month of the year) at each EMEP site over 4-year periods. Although hierarchical clustering has been applied previously to monitoring site classification (Tarasova et al., 2007), the novelty here was the subsequent application of non-negative matrix factorisation (NMF) (Lee and Seung, 2001) to order the sites across the dendrogram according to an extracted factor. In this case, the factor represented the extent of anthropogenic influence on ozone concentrations. Hierarchical cluster analysis was chosen in preference to non-hierarchical techniques as the robustness and suitability of the cluster assignment are more objectively investigated through the resulting dendrogram, particularly when this is combined with NMF. By using NMF, the ozone concentrations at the two UK EMEP supersites were placed more precisely in the European context. The analysis was carried out separately for five 4-year periods spanning 1991–2010 to assess the consistency of site representativeness over time.

2. Methodology

Data arrays of 4-year averaged monthly-diurnal ozone concentrations were calculated for each EMEP site, i.e. 288 ($=24\ \text{h} \times 12$ months) ozone concentrations per site where, for example, the ozone concentration for 'Jan-00.00' was the average of the 00.00–01.00 hourly ozone on all days in January in the 4-year period under consideration (1991–1994, 1995–1998, 1999–2002, 2003–2006 and 2007–2010) (measured data from http://ebas.nilu.no). Four-year averages of monthly-diurnal concentrations were considered a reasonable compromise between long enough to smooth out inter-annual variability and short enough to avoid incorporation of long term trends. The number of sites included in each time period and the number of countries within which these sites are located are summarised in Table 1. 154 sites contributed ozone data to at least one 4-year period, of which 49 contributed to every 4-year time period.

The choice of clustering parameters can impact the clustering result. In this study, the standard Euclidean distance between two n-dimensional data arrays was used and in this case $n = 288$ (Eq. (1)).

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (1)$$

In hierarchical clustering, each object (here each site's monthly-diurnal ozone concentrations) initially constitutes its own cluster. The two nearest clusters are then combined

Table 1
Number of sites used in cluster analysis for each four year period. The increasing number of countries with sites indicates the increasing geographical coverage across Europe with time. 49 sites are common to all time periods.

Time period	No. of sites	No. of countries
1991–1994	76	14
1995–1998	100	20
1999–2002	117	27
2003–2006	117	27
2007–2010	113	26