Environmental Modelling & Software 59 (2014) 1–9

Contents lists available at ScienceDirect

Environmental Modelling & Software

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

Conditional bivariate probability function for source identification

Iratxe Uria-Tellaetxe [a], David C. Carslaw [b,*]

[a] Department of Chemical and Environmental Engineering, University of the Basque Country, School of Engineering of Bilbao, Alameda Urquijo, 48013 Bilbao, Spain
[b] King's College London, Environmental Research Group, Franklin Wilkins Building, 150 Stamford Street, London SE1 9NH, UK

ARTICLE INFO

Article history:
Received 1 November 2013
Received in revised form
24 April 2014
Accepted 2 May 2014
Available online

Keywords:
Receptor model
Source identification
Air quality data
Dispersion model
Air pollution
Openair

ABSTRACT

In this paper a new receptor modelling method is developed to identify and characterise emission sources. The method is an extension of the commonly used conditional probability function (CPF). The CPF approach is extended to the bivariate case to produce a conditional bivariate probability function (CBPF) plot using wind speed as a third variable plotted on the radial axis. The bivariate case provides more information on the type of sources being identified by providing important dispersion characteristic information. By considering intervals of concentration, considerably more source information can be revealed that is absent in the basic CPF or CBPF. We demonstrate the application of the approach by considering an area of high source complexity, where many new sources can be identified and characterised compared with currently used techniques. Dispersion model simulations are undertaken to verify the approach. The technique has been made available through the openair R package.

Software availability

The methods described in this work are available as part of software called openair. The openair software is freely available as an R package. Details on installing R and optional packages including openair can be found at R Core Team (2014) and http://www.r-project.org. R will run on Microsoft Windows, linux and Apple Mac computers. No special hardware is required to run openair other than a standard desktop computer. Some large data sets or complex analyses may require a 64-bit platform. Ref: R Core Team (2014). R: A language and environment for statistical computing. RFoundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

1. Introduction

1.1. Background

Identifying local and distant emission sources through receptor modelling is an important area in the management of air pollution. Receptor modelling techniques are diverse and have been applied to a very wide range of situations. Among the more important aspects of receptor modelling is the ability to identify and characterise emission sources, which would perhaps be difficult or impossible by other means. While air quality models can be used together with emission inventories to provide such information, in practice this is difficult. It is difficult for many reasons including incomplete information of the sources and the difficulty in modelling boundary layer processes. For this reason the analysis of ambient air quality data remains a central approach used for understanding emission sources.

A commonly used method for identifying sources is the Conditional Probability Function (CPF). The CPF is a simple but effective technique for providing directional information concerning major sources (Ashbaugh et al., 1985; Vedantham et al., 2013). The CPF calculates the probability that in a particular wind sector the concentration of a species is greater than some specified value. The value specified is usually expressed as a high percentile of the species of interest e.g. the 75th or 90th percentile. It is also possible to extract and filter source information data through conditional analysis as described by Malby et al. (2013). As Malby et al. (2013) show filtering air pollution data by wind speed, direction and time of day can help isolate specific source types for further analysis e.g. the calculation of long term trends.

Bae et al. (2011) used a CPF technique to help identify the directionality of sources contributing to observed pollutant concentrations at a rural site in New York State. The species considered included hourly averaged $PM_{2.5}$ mass, Organic Mass (OM)

* Corresponding author.
E-mail address: david.carslaw@kcl.ac.uk (D.C. Carslaw).

http://dx.doi.org/10.1016/j.envsoft.2014.05.002
1364-8152/

from Organic Carbon (OC), optical Elemental Carbon (optical EC), SO_2, CO, NO_y and O_3 for the period of December 2004 to December 2008. In addition, Bae et al. (2011) also considered seasonal variations of these species. Bae et al. (2011) enhanced the basic CPF technique by coupling the method with back trajectory calculations to provide more information on mid to long distance sources.

More sophisticated approaches have also been used to identify dominant sources using non-parametric statistical analysis. Henry et al. (2009) developed a non-parametric wind regression approach to identify and quantify the impact of possible source regions of pollutants as defined by wind direction sectors. Using this approach Henry et al. (2009) were able to quantify the source contribution of different emission sources and demonstrate that some large sources such as a steel mill made only minor contributions to concentrations of SO_2.

Kim and Hopke (2004) compared the CPF approach with the non-parametric regression approach for fine particle concentrations ($PM_{2.5}$) in the USA. They found that CPF and non-parametric regression methods both worked well in identifying known local point sources. However, the CPF approach was easier to calculate compared with the non-parametric regression approach. The results from Kim and Hopke (2004) in both cases provided broad, dominant source directions such as the Port of Seattle or in the case of sea salt, the Atlantic Ocean. An advantage of the non-parametric regression approach is that it is also able to provide uncertainties in the source direction for major sources.

Most of the techniques previously described are focused on identifying and may be quantifying dominant sources affecting a receptor. However, many or most receptors are affected by a much larger number of sources — but they can be difficult to identify. These other sources could include major point sources that are too far from the receptor to be detected clearly or local minor sources that are similarly difficult to detect. There are however compelling reasons why it is useful to detect such sources at a receptor. While there may only be a minor contribution at a specific receptor, it may well be the case that at other locations (perhaps where no measurements are made), the contribution could be much larger and should therefore be investigated. It is also useful to know the extent to which sources have an influence, as this can provide a more complete picture of how sites are affected by a wide range of sources. For example, if it can be shown that a major point source can be detected much further from its location than previously thought, such information is helpful for demonstrating this to be the case. There may also be occasions where isolating particular source types is useful e.g. thermal power plants. Furthermore, there may also be opportunities for enhanced model evaluation by being able to evaluate models over a much larger spatial area.

In this paper a new technique is developed that increases the potential to both detect and characterise source contributions at receptor locations. The new method combines a conditional probability function with bivariate polar plots. The former is useful for source detection and the latter for additional source characterisation. The approach is further enhanced by considering the full distribution of concentrations rather than concentrations exceeding a particular threshold. The method is described and then applied to an area of high source complexity that is affected by both near-field and distant sources. Model simulations are performed to show that similar findings can be gained through the analysis of model predictions.

2. Method

2.1. Bivariate CPF methodology

The ordinary CPF (Ashbaugh et al., 1985) estimates the probability that the measured concentration exceeds a set threshold criterion for a given wind sector. CPF is mathematically defined as:

$$\mathrm{CPF}_{\Delta\theta} = \frac{m_{\Delta\theta}|_{C\geq x}}{n_{\Delta\theta}} \tag{1}$$

Where $m_{\Delta\theta}$ is the number of samples in the wind sector θ having concentration C is greater than or equal to a threshold value x, and $n_{\Delta\theta}$ is the total number of samples from wind sector $\Delta\theta$. Thus, CPF indicates the potential for a source region to contribute to high air pollution concentrations. Conventionally, x represents a high percentile of concentration e.g. the 75th or 90th.

The conditional bivariate probability function (CBPF) couples ordinary CPF with wind speed as a third variable, allocating the observed pollutant concentration to cells defined by ranges of wind direction and wind speed rather than to only wind direction sectors. It can be defined as:

$$\mathrm{CBPF}_{\Delta\theta,\Delta u} = \frac{m_{\Delta\theta,\Delta u}\big|_{C\geq x}}{n_{\Delta\theta,\Delta u}} \tag{2}$$

Where $m_{\Delta\theta,\Delta u}$ is the number of samples in the wind sector $\Delta\theta$ with wind speed interval Δu having concentration C greater than a threshold value x, $n_{\Delta\theta,\Delta u}$ is the total number of samples in that wind direction-speed interval. The extension to the bivariate case provides more information on the nature of the sources because different source types can have different wind speed dependencies. The use of a third variable can therefore provide more information on the type of source in question. It should be noted that the third variable plotted on the radial axis does not need to be wind speed. The key issue is that the third variable allows some sort of discrimination between sources types due to the way they disperse. For example, Carslaw and Beevers (2013) show that temperature can be a useful radial variable.

Bivariate polar plots show how a concentration of a species varies jointly with wind speed and wind direction in polar coordinates. The plots have proved to be useful in a range of settings e.g. to characterise airport sources and dispersion characteristics street canyons (Carslaw et al., 2006; Tomlin et al., 2009; Carslaw and Ropkins, 2012). Wind direction together with wind speed can be highly effective at discriminating different emission sources. By using polar coordinates the plots provide a useful graphical technique which can provide directional information on sources as well as the wind speed dependence of concentrations.

Briefly, bivariate polar plots are constructed in the following way. First, wind speed, wind direction and concentration data are partitioned into wind speed-direction bins and the mean concentration calculated for each bin. The wind components $u = \bar{u}.\sin(2\pi/\theta)$, $v = \bar{u}.\cos(2\pi/\theta)$, where $\bar{u}$ is the mean wind speed and θ is the mean wind direction in degrees with 90° as being from the east, and concentration (C) provide a surface. The concentration surface produced by u, v and C is modelled using a Generalized Additive Model (GAM) (Wood, 2006). GAMs are a useful modelling framework with respect to air pollution prediction because often the relationships between variables are non-linear and variable interactions are important, both of which issues can be addressed in a GAM framework. The surface is fitted according to Equation 3:

$$\sqrt{C_i} = \beta_0 + s(u_i, v_i) + \varepsilon_i \tag{3}$$

where C_i is the ith pollutant concentration, β_0 is the overall mean of the response, $s(u_i,v_i)$ is the isotropic smooth function of ith value of covariate u and v, and ε_i is the ith residual. A penalized regression spline is used to model the surface as described by Wood (2003). Note that C_i is square-root transformed as the transformation generally produces better model diagnostics e.g. normally distributed residuals. Moreover the smooth function used is *isotropic* because u and v are on the same scales. The isotropic smooth avoids the potential difficulty of smoothing two variables on different scales e.g. wind speed and direction, which introduces further complexities. When fitting the GAM, wind speed-direction bins with few data points are down-weighted such that those with 1, 2 and 3 points have weights 0.25, 0.50 and 0.75, respectively, whereas for sample sizes >3 are given a weighting of one. This approach therefore gives less weighting to wind speed-direction intervals (and therefore conditional probability estimates) that contain very few data points.

The CBPF can be extended further to consider intervals of concentration rather than only values greater than some threshold. In this case the CBPF for concentration intervals is defined as:

$$\mathrm{CBPF}_{\Delta\theta,\Delta u}(i) = \frac{m_{\Delta\theta,\Delta u}\big|_{y\geq C\geq x}}{n_{\Delta\theta,\Delta u}} \tag{4}$$

Where $m_{\Delta\theta,\Delta u}$ is the number of samples in the wind sector $\Delta\theta$ with wind speed interval Δu having a concentration C between the intervals x and y, $n_{\Delta\theta,\Delta u}$ is the total number of samples in that wind direction-speed interval. The extension to considering intervals of concentration is important because it extends the basic CPF methodology (that only considers concentrations greater than a specified value) to provide much more comprehensive information for source identification. The basic CPF method focuses on identifying the most important dominant sources i.e. the ones that make the greatest contribution to high concentration conditions. However, as it will be shown, the basic CPF method discards a large amount of useful