Atmospheric Environment 145 (2016) 128–134

Contents lists available at ScienceDirect

Atmospheric Environment

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

Source apportionment advances using polar plots of bivariate correlation and regression statistics

Stuart K. Grange [a, *], Alastair C. Lewis [a, b], David C. Carslaw [a, c]

[a] Wolfson Atmospheric Chemistry Laboratory, University of York, York, YO10 5DD, United Kingdom
[b] National Centre for Atmospheric Science, University of York, Heslington, York, YO10 5DD, United Kingdom
[c] Ricardo Energy & Environment, Harwell, Oxfordshire, OX11 0QR, United Kingdom

HIGHLIGHTS

- Bivariate polar plots are a common method for exploring pollutant sources.
- Polar plots were enhanced with the addition of pair-wise statistics.
- Usage examples of the enhanced polar plots are given for two London monitoring sites.
- Processes were illuminated that were not detected by other plotting methods.
- Potential future applications and extensions are discussed for bivariate polar plots.

ARTICLE INFO

Article history:
Received 20 June 2016
Received in revised form
5 September 2016
Accepted 10 September 2016
Available online 14 September 2016

Keywords:
Air quality
Relationships
Robust regression
Particulate matter
Black carbon

ABSTRACT

This paper outlines the development of enhanced bivariate polar plots that allow the concentrations of two pollutants to be compared using pair-wise statistics for exploring the sources of atmospheric pollutants. The new method combines bivariate polar plots, which provide source characteristic information, with pair-wise statistics that provide information on how two pollutants are related to one another. The pair-wise statistics implemented include weighted Pearson correlation and slope from two linear regression methods. The development uses a Gaussian kernel to locally weight the statistical calculations on a wind speed-direction surface together with variable-scaling. Example applications of the enhanced polar plots are presented by using routine air quality data for two monitoring sites in London, United Kingdom for a single year (2013). The London examples demonstrate that the combination of bivariate polar plots, correlation, and regression techniques can offer considerable insight into air pollution source characteristics, which would be missed if only scatter plots and mean polar plots were used for analysis. Specifically, using correlation and slopes as pair-wise statistics, long-range transport processes were isolated and black carbon (BC) contributions to $PM_{2.5}$ for a kerbside monitoring location were quantified. Wider applications and future advancements are also discussed.

1. Introduction

Determining how variables are related to one-another is a key component of data analysis and statistics. Within the atmospheric sciences, exploring the relationships between chemical constituents and meteorological parameters is extremely common and the techniques for comparing, correlating, and determining relationships are very diverse. Analysis involving the correlation of two pollutants can often be insightful because it can lead to the identification of emission source characteristics, as can investigation into ratios or slopes from regression analysis between two pollutants (Statheropoulos et al., 1998). Within atmospheric disciplines, data analysis can also benefit from being able to integrate wind behaviour (Elminir, 2005). The use of wind speed and direction can be informative because it often leads to the suggestion of source locations and source characteristics, such as height of emission above the surface (Henry et al., 2002; Westmoreland et al., 2007).

Exploration of relationships among variables can be achieved with many different methods that can range from the simple to numerically complex. However, a technique that is used very

* Corresponding author.
E-mail addresses: skg511@york.ac.uk (S.K. Grange), david.carslaw@york.ac.uk (D.C. Carslaw).

http://dx.doi.org/10.1016/j.atmosenv.2016.09.016
1352-2310/© 2016 Published by Elsevier Ltd.

widely is the simple *x-y* scatter plot (Bentley, 2004). Scatter plots are useful because they allow for the visualisation of variables and model fitting can be evaluated quickly and simply with visual feedback. Regression techniques, most commonly ordinary least-squared regression, are often employed to formally quantify how *x* and *y* are related. The use of least-squared regression is however technically questionable in many cases, and despite a large collection of alternative techniques available, its use remains a persistent feature of air quality data analysis. The use of simple scatter plots is usually carried out with entire datasets or with simple or superficial filtering and therefore have potential to hide some discrete relationships which are present in the global data if they do not conform to the mean rate of change (Cade and Noon, 2003).

Slopes from regression models relating two pollutants to one-another are often used in applications that use monitoring data such as emission inventories and pollutant models. When measurements are not available, slopes for the unknown pollutants are often substituted from the literature, short-term monitoring, or data collected at a near-by location. However, the use of simple and static ratios is likely to be deficient in many situations because they can be expected to be highly dependent on source (Manoli et al., 2002). To differentiate sources in air quality data, techniques other than simple scatter plots often need to be used.

A common method for source characterisation is the use of bivariate polar plots (Carslaw et al., 2006; Westmoreland et al., 2007; Carslaw and Beevers, 2013; Uria Tellaetxe and Carslaw, 2014). Polar plots are typically used to visualise and explore mean pollutant concentrations for single species based on wind speed and wind direction. In the atmospheric sciences, it is intuitive to plot wind direction (from 0 to 360 $^{\circ}$ clockwise from north) on the angular 'axis' and wind speed to be used for the radial scale. Aggregation functions other than the arithmetic mean can be used and different variables apart from wind speed can be used for the radial scale. For example, atmospheric temperature or stability are often useful variables to use. The main attribute for the choice of radial-axis variable is that it helps to differentiate between different source characteristics in some way due to different source types responding differently to values of the angular scale. Despite the range of potential options, wind speed is widely used to help discriminate different source types and is particularly useful when used together with wind direction and the concentration of a species (Harrison et al., 2001; Kassomenos et al., 2012).

This type of polar plot analysis has, in part, become wide-spread due to the open-source polarPlot function available in the *openair* R package (Carslaw and Ropkins, 2012; R Core Team, 2016). Other similar techniques such as non-parametric wind regression have also shown their ability to determine source locations for various pollutants by using polar plots (Henry et al., 2002, 2009; Donnelly et al., 2011).

1.1. Objectives

Combining correlation and regression techniques with those that provide information on source apportionment potentially offers considerably more insight into air pollution sources. The use of wind behaviour has the potential to evaluate correlation and slopes based on source locations and therefore different processes. It is common for emission inventories to use ratios for pollutants when they are not measured or when high quality data is lacking. It is hypothesised that the combination of correlation, regression, and polar plots could lead to significant additions to data analysis by understanding how different pollutants are related to one another depending on source.

In this paper, the combination of bivariate polar plots approaches with correlation and regression techniques is considered for comparing two pollutants. This combination of methods is then used to aid the interpretation of air quality data. The primary objectives of this paper are as follows. First, to develop methods to combine bivariate polar plot techniques with correlation and a range of linear regression approaches. Second, apply the methods to commonly available measurements of air pollutants to demonstrate the new insights made possible by these techniques. Third, to consider the wider potential uses of the approaches in air quality science. The software developed has been released with an open-source licence and can be found in the *polarplotr* R package (Carslaw and Grange, 2016).

2. Methods

2.1. Function development

2.1.1. Kernel weighting and scaling

The plotting mechanism for polar plots when using wind direction as the polar axis generally involves first aggregating a time-series into wind speed and direction intervals (or 'bins'). The specific intervals and numbers of the bins can be altered for a particular application, but all combinations of the two types of bins are summarised by an aggregation function such as the mean or maximum. In the *openair* `polarPlot` function, a smoothed surface is fitted to these binned summaries using a generalised additive model (GAM) to create a continuous surface which can be plotted with polar coordinates. Further details of the approach can be found in Carslaw and Beevers (2013) and Uria Tellaetxe and Carslaw (2014).

When applying a simple aggregation function, the number of observations in a time-series which compose a discrete wind speed and direction bin is not critical for the calculation or the visual presentation of the surface, except at the edges of the plot where there are (usually) few observations. However, when calculating correlations or relationships between two variables, it becomes important to consider the minimal number of observations which would create a valid summary. If there are too few observations for a particular bin and a statistic such as the correlation or slope is calculated between a pair of variables, it is likely that unreliable summaries will be generated due to large variations between neighbouring bins. To overcome this limitation, for each wind speed and direction bin, the entire time-series was evaluated but observations were *weighted* by their proximity to a wind speed and direction bin *i.e.*, wind speed or direction values further from the bin centre are weighted less than those closer to the centre of the bin. Like previous works such as Henry et al. (2002, 2009), a weighting kernel was used to create weighting variables.

The weighting kernel used was the Gaussian kernel (Equation (1)). The Gaussian kernel has infinite tails and therefore all input bins are given a non-zero weighting, but observations furthest from the bin being analysed have very small weights associated with them. The Gaussian kernel was used for weighting both wind speed and direction because it is considered more utilitarian than many other kernels such as the Epanechnikov kernel which have finite bounds and therefore at times, will give observations weights of zero which can cause ambiguity issues.

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} \quad (1)$$

To ensure the weighing variable was appropriate for the particular wind speed and direction application, the input wind speed and direction variables required scaling. The scaling process used was simple; the wind variables were multiplied by an integer to increase their bounds and therefore influence within the