Editorial
Modern tools for air quality data analysis – *openair*

The amount of air quality data collected around the world is enormous and continues to grow. This data is expensive to collect and manage, requiring specialist equipment, trained staff and a commitment over many years. Despite the large investment in collecting data, the analysis of data is more ad-hoc, fragmented and limited in ambition. Often, these data are analysed in very basic ways such as to compare measured concentrations with air quality standards and other guidelines. While such analysis is useful from a compliance perspective, it is also a wasted opportunity to gain more insight into the underlying characteristics of air pollution. An improved understanding of the causes of air pollution ultimately leads to improved air quality management.

One of the barriers to the more insightful analysis of air pollution data has been the lack of readily available tools for data analysis. Resources available for enhanced data analysis are often distributed across many potentially expensive types of specialist software e.g. GIS and other proprietary software. Globally, access to such specialist software and the knowledge required to use it is uneven. It is with this background that the *openair* project was developed. The project was funded by the UK Natural Environment Research Council (NERC) over three years with the central aim of providing a set innovative data analysis tools for the air quality community.

openair was developed with the aim of maximising use across a wide range of users including academia, the public and private sectors. In order to fulfil this and other aims, the software was developed using the open-source software called R. R is often called statistical analysis software but it is perhaps better to describe it as an interactive data analysis system: it is a programming language specifically designed for data analysis. R is ideally suited for the development of *openair* as it is free, open-source and works on Windows, Mac OS X and linux, thus maximising the potential for widespread use. While some knowledge of R is helpful for users of *openair*, it is not essential; efforts have been made to make *openair* analyses straightforward.

So what can *openair* do to help air quality researchers and practitioners? First, it provides basic, but powerful tools, commonly used by those interested in air quality data analysis. These tools include wind and pollution roses for example. A key functionality of *openair* is to allow users to view their data in many different ways while making it as simple as possible. For example, while a simple pollution rose is useful, it is trivial in

openair to produce a series of pollution roses that split the data by season, year, day of the week, hour of the day – or by any other variable in the data set of interest. This flexibility, which is at the heart of R and *openair*, immediately allows the user to follow different lines of enquiry in a fast and efficient way; building up a more complete picture about the characteristics of air pollution.

The plot below shows an example of pollution roses for NO_x by day of the week. The size of the segments gives information on the amount of time the wind was from a particular direction and the colour shows the concentration interval of NO_x. The prevalence of south-westerly winds and the reduced frequency of high NO_x concentrations at weekends is easily apparent. The plot was produced with one simple line of code:

```
pollutionRose(mydata, pollutant="nox", type="weekday")
```
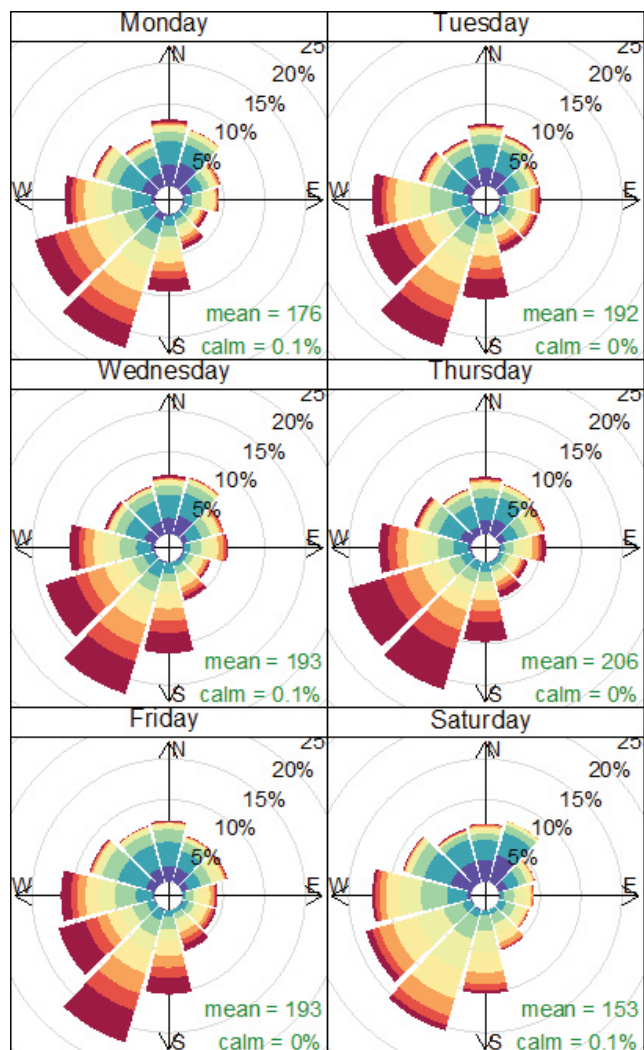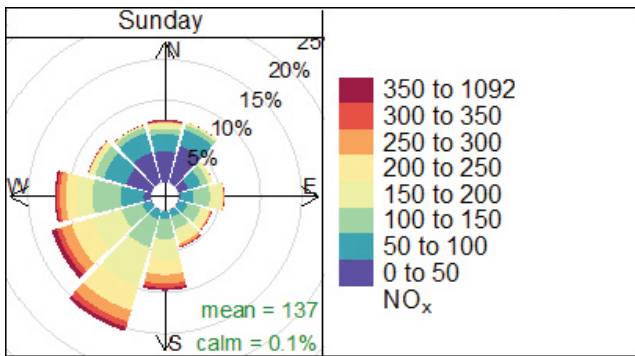
Frequency of counts by wind direction (%)

openair also includes some more sophisticated analysis capabilities. The use of bivariate polar plots for example have been extensively used throughout the world to better understand emissions source and dispersion characteristics. Using the in-built capabilities of R it is possible to access sophisticated data fitting techniques such as the use of Generalized Additive Models that allow smooth surfaces to be modelled. Additionally, cluster analysis of the patterns observed on these plots can be carried out to allow the user to refine and further analyse interesting 'features'. Other capabilities include the easy access to NOAA Hysplit back trajectory analysis, allowing users to determine air mass origins, carry out cluster analysis and other analyses in *openair*. More recently, powerful and flexible functions have been added for the evaluation of air quality models; again allowing users to easily investigate model performance split for example by day of the week, season and daylight/nighttime.

The plot below shows where 96-hour back trajectories during 2013 spend most of their time based on a receptor located in London. The plot is again easy to produce. The first line imports the back trajectory data from the King's College web server and the second line plots the frequency of time air masses spend over different geographic regions.

```
traj <- importTraj(site="london", year=2013)
trajLevel(traj, col="increment")
```
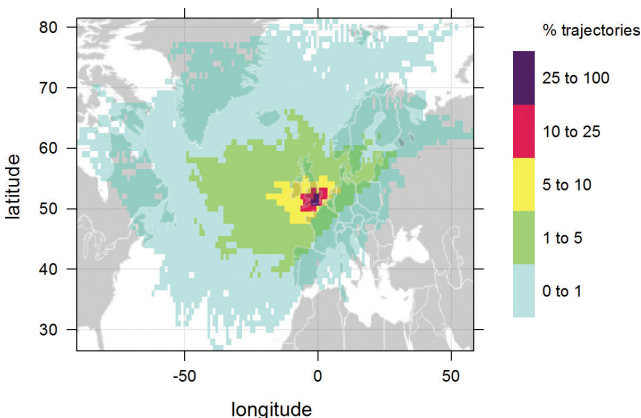

The number and types of analysis that can be conducted by *openair* is potentially vast and continues to grow, develop and be refined. There is an increasing number of journal publications that use *openair* for data analysis, which provides users with useful information on how air quality researchers use the software to develop new insights. One very recent example considers wood smoke and other particulate sources in London (Crilley *et al.*, 2014). However, *openair* is also extensively used across the world by consultancies and public bodies, showing that is has wide reach. More information about the software and project can be found at the *openair* website (www.openair-project.org). A considerable amount of time has been used to develop a comprehensive user manual (available on the website), to provide users with fully reproducible examples of how to use *openair*. In the longer term it is hoped that *openair* will provide fast and simple access to a much broader range of data sources, as well as continuing to develop improved methods of data analysis for air quality researchers and practitioners.

L. R. Crilley, W. J. Bloss, J. Yin, D. C. S. Beddows, R. M. Harrison, J. D. Allan, D. E. Young, M. Flynn, P. Williams, P. Zotter, A. S. H. Prevot, M. R. Heal, J. F. Barlow, C. H. Halios, J. D. Lee, S. Szidat, and C. Mohr (2014). Sources and contributions of wood smoke during winter in London. Atmospheric Chemistry and Physics Discussions. 14, 27459–27530. http://www.atmos-chem-phys-discuss.net/14/27459/2014/acpd-14-27459-2014.pdf

Dr David Carslaw
King's College London