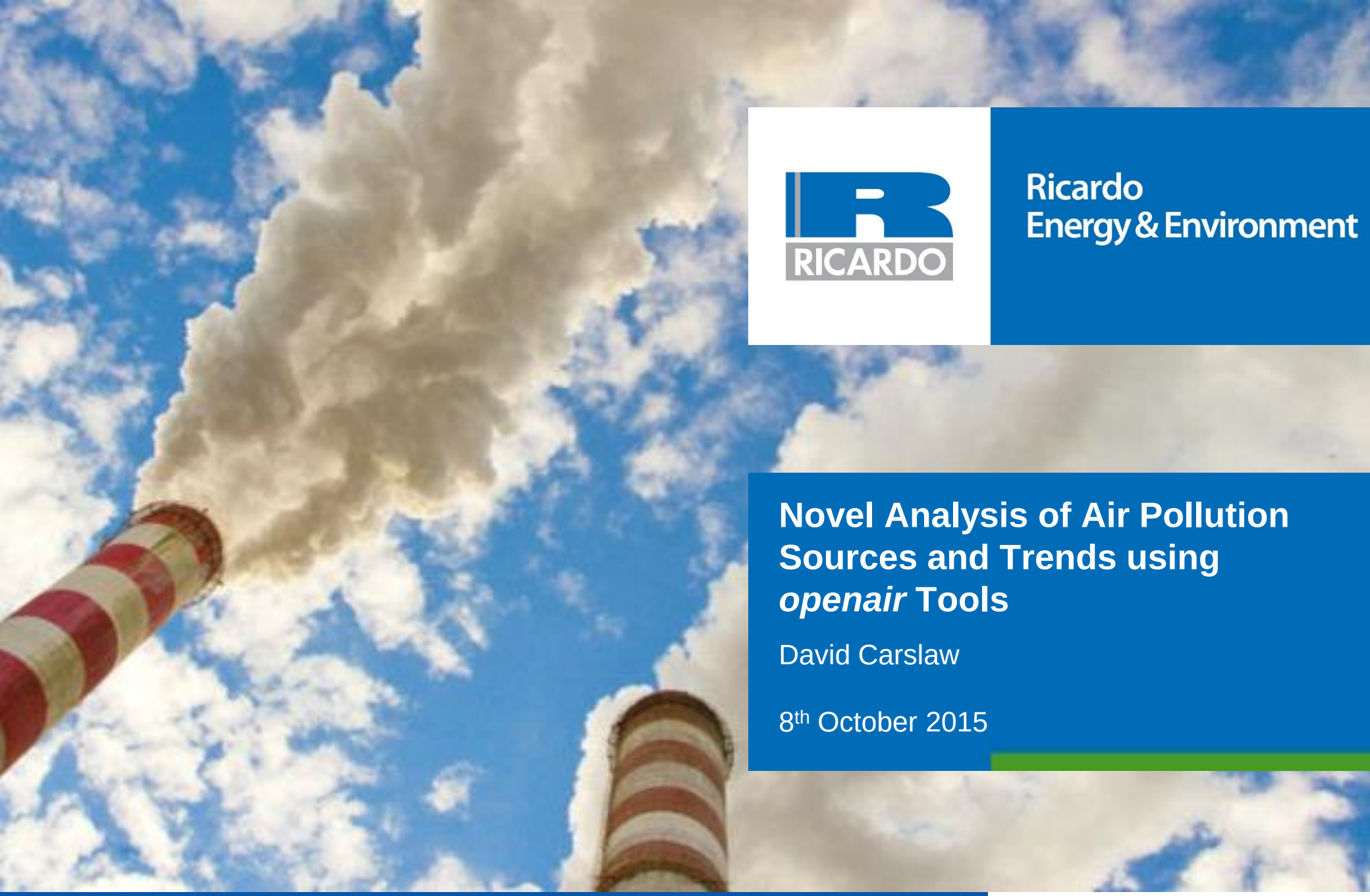


Ricardo
Energy & Environment

# Novel Analysis of Air Pollution Sources and Trends using *openair* Tools

David Carslaw

8th October 2015

# Briefly…

- What is **openair** and why was it developed?
- What can **openair** do?
  - Some examples of recent developments
  - Trajectory modelling
  - Future developments

# The importance of measurements

- High quality measurements are the cornerstone of effective air quality management
- There are 100s of continuous monitoring sites across the UK which represent a significant capital and on-going investment (£10Ms)

But…

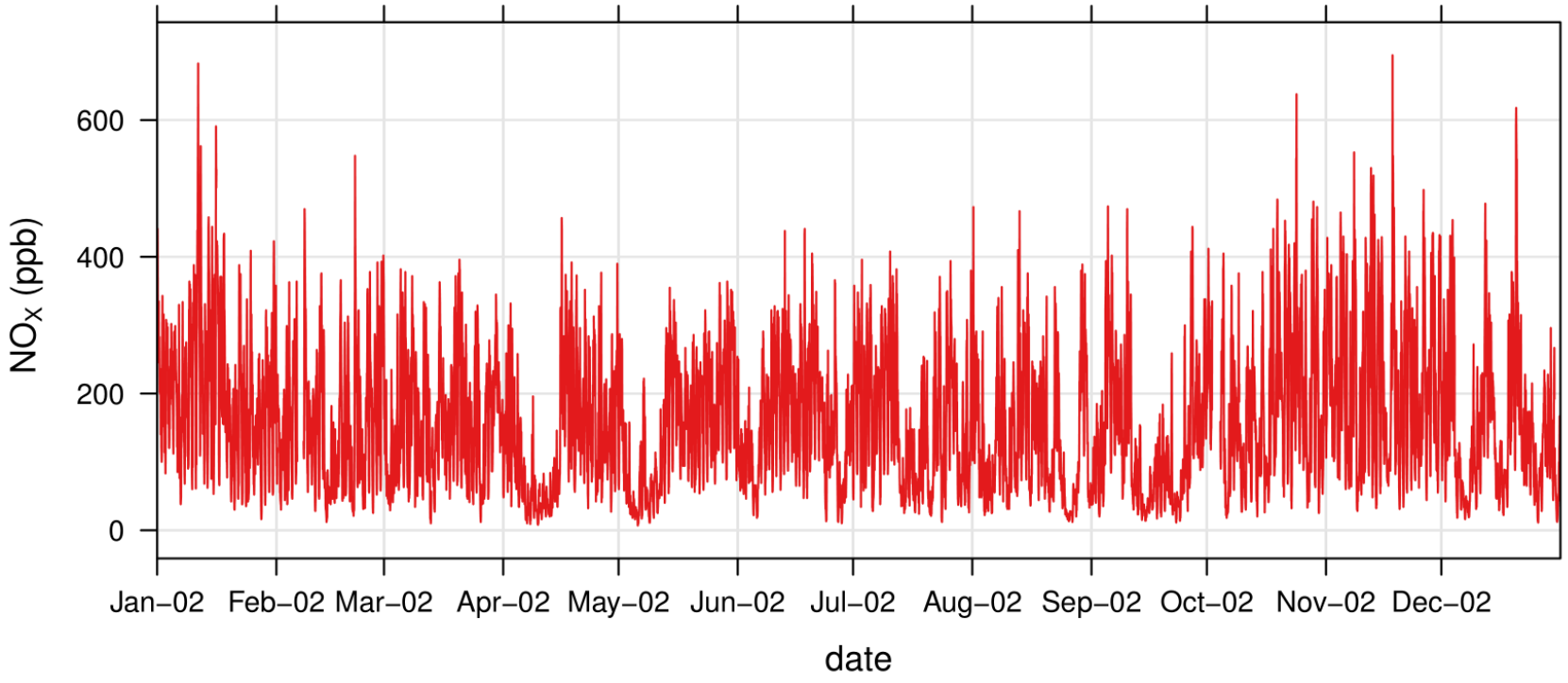
- If we only compare annual means for compliance reasons we are massively under-exploiting the value of these measurements

And…

- No dedicated analysis tools to do more with data

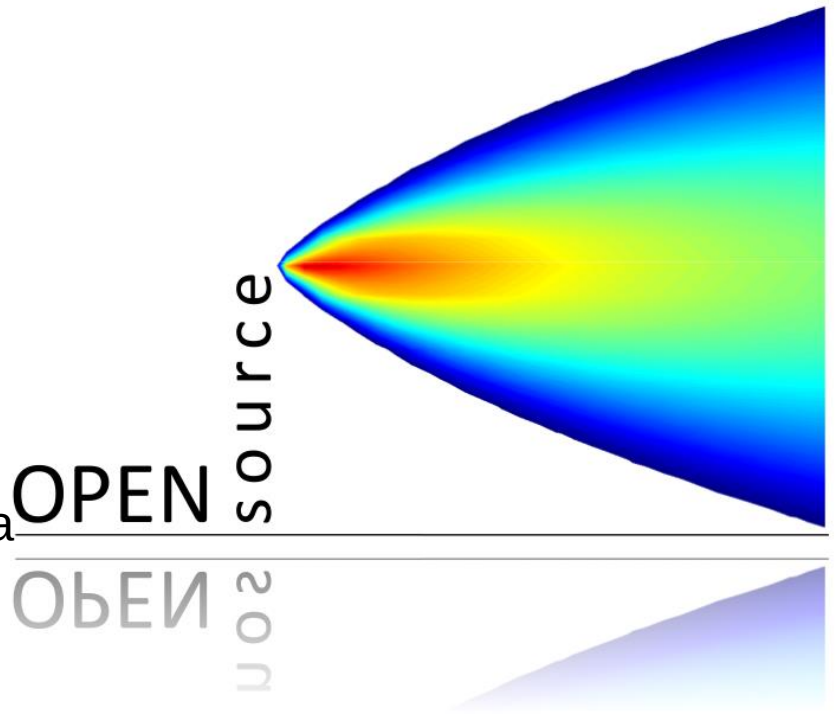
# The grand challenge…

How to extract meaning from this…?



# The openair project

- Started in 2008 with 3-year NERC funding with additional support from Defra
- **Aim: to make innovative open-source data analysis tools freely available to the air quality community**
  - *Sub aim*: As much as possible, no programming knowledge required by users
- Use software called R
  - Often thought of as statistical software
  - But it is really a programming language ***specifically designed for data analysis***
  - Usage and capabilities continue to grow at a rapid rate (> 7,000 'packages')
- **openair** is one of these R packages dedicated to the analysis of air quality data



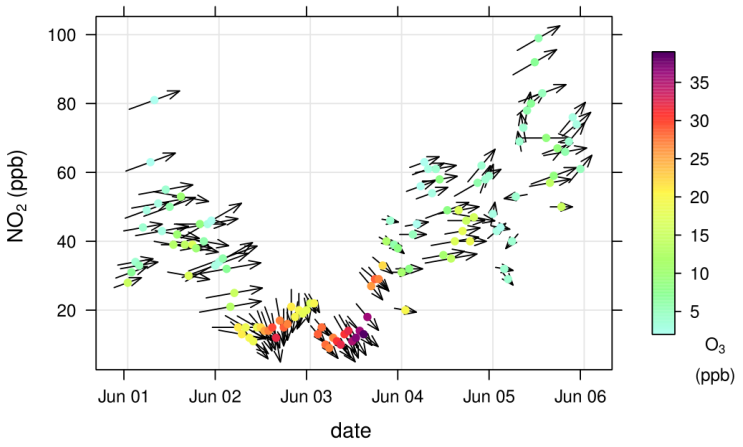
# R – software that can do many things…

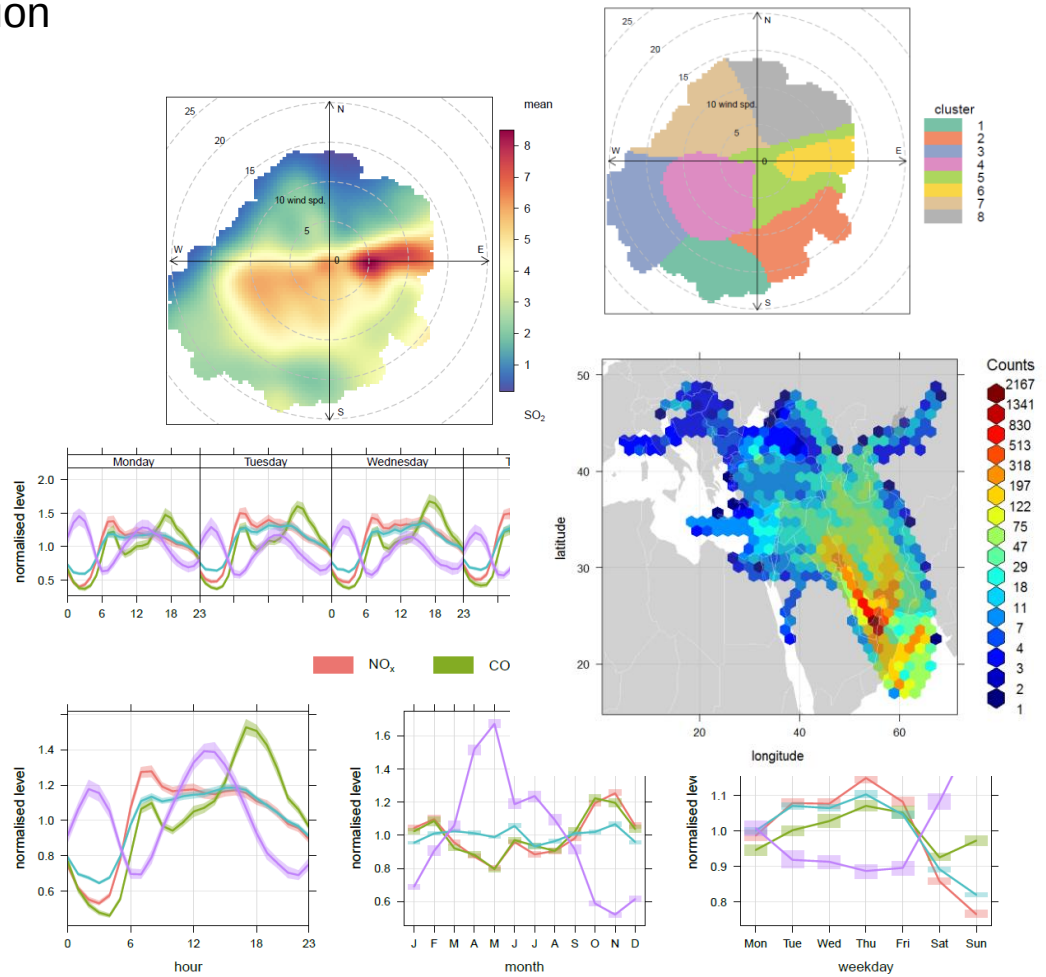
- Excellent for access to many data formats
  - csv, txt, Excel, binary files, databases (e.g. SQL Server, MySQL, Postgres…), XML, JSON, web scraping, NetCDF, …
- Did I mention > 7000 packages?
  - Almost endless possibilities
  - Excellent code sharing on *Github* (think Facebook for computer code)
- Growing capability for reproducible reporting / research and interactive web-based 'rich content' document

# Overview of openair

## Many capabilities

- Source detection and characterisation
- Robust trend analysis
- Local and regional cluster analysis
- Back trajectory analysis
- Model evaluation
- Training possibilities
- Widely used*



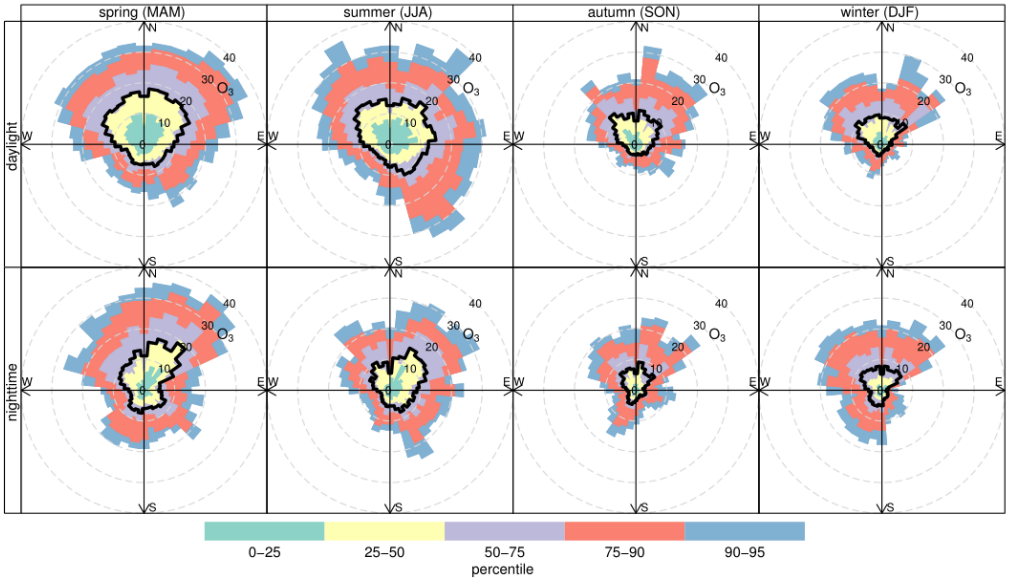




*Downloaded >28,000 times via RStudio, 1,500 to 2,000 times a month, top 7% of all R packages

# Conditioning plots – central theme

- Almost all **openair** functions have a ‘type’ option that allows data to be partitioned in flexible ways
- Built-in types include
  - "year" splits data by year
  - "month" splits variables by month of the year
  - "season" splits variables by season.
  - "weekday" splits variables by day of the week
  - "weekend" splits variables by Saturday, Sunday, weekday
  - "daylight" splits variables by nighttime/daytime.
  - "wd" if wind direction (wd) – will split the data up into 8 sectors: N, NE, E, SE, S, SW, W, NW.

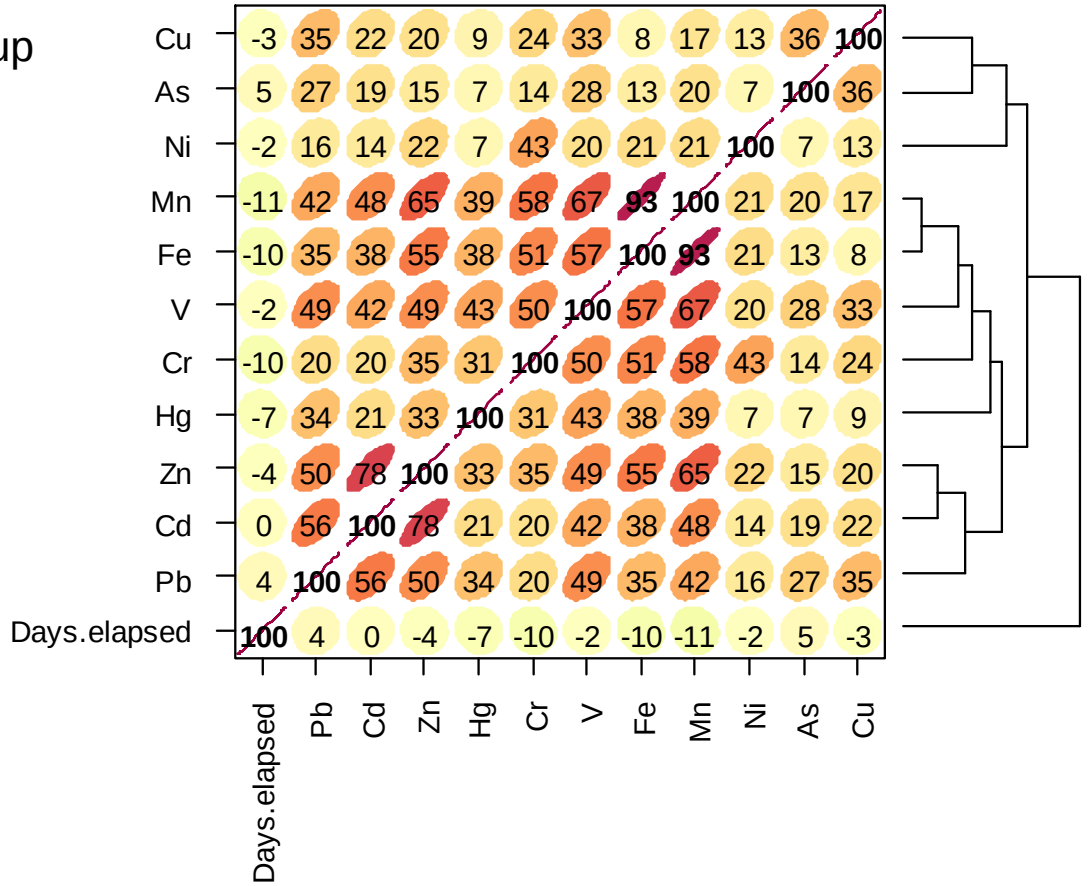
```
percentileRose(mydata, type = c("season", "daylight"), pollutant = "o3",
               col = "Set3", mean.col = "black")
```



# Multispecies correlations + clustering

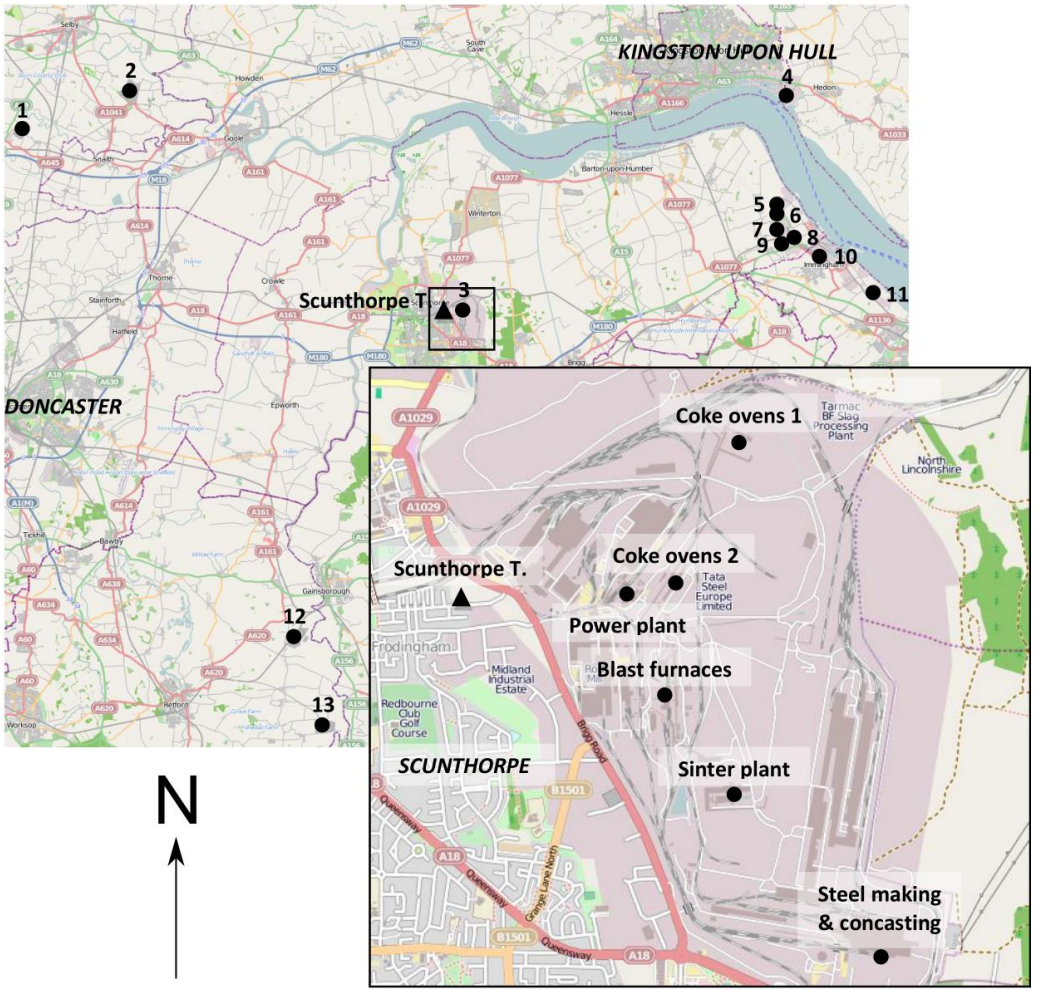
- Can be hard to get a feel for what's going on with multiple measurements
- Can look at correlations
- Apply hierarchical clustering to group things that are most similar to one another

```
corPlot(metals, dendrogram = TRUE)
```



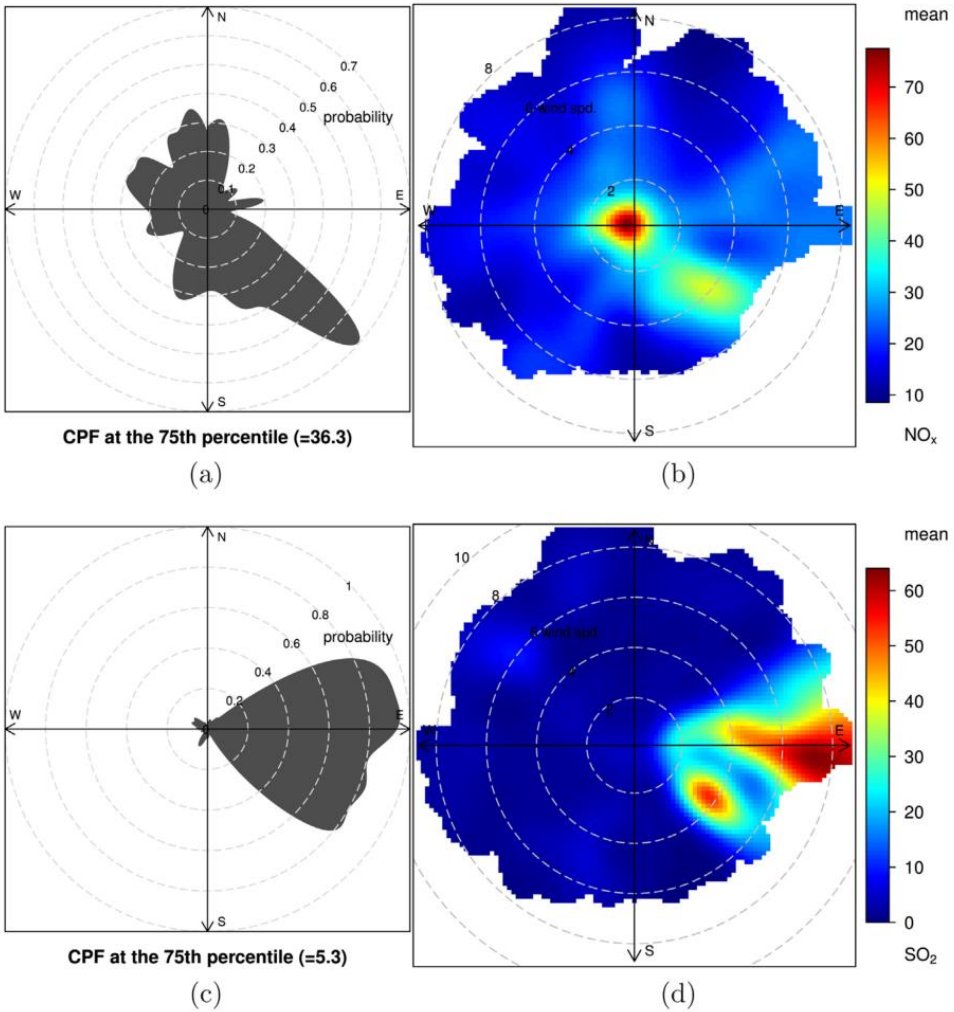
# Source characterisation with bivariate polar plots

- Scunthorpe is an interesting example
- Very complex local sources of multiple pollutants from integrated steelworks
- Major but much more distant sources e.g. Drax power station (#2 on the map, ~35 km from measurement site)
- How to un-pick the contributions made by these very different sources?
- Examples of what can be done with 3 simple variables: concentration, wind speed and wind direction



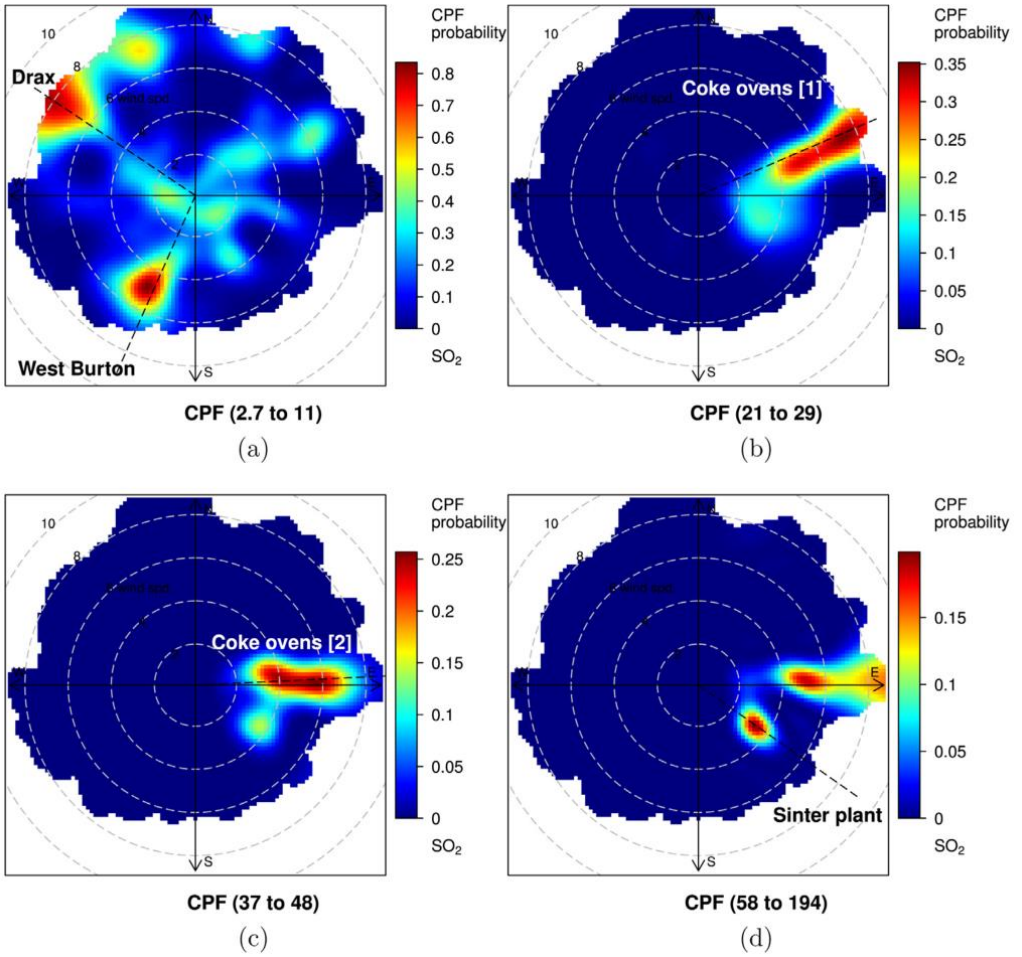
# Source characterisation with bivariate polar plots

- Example at Scunthorpe Town
  - Steelworks to the east
- Just considering concentration by wind direction only tells us about the **direction** of major sources
- Considering the joint wind speed-direction variation says much more about the **nature** of the emission sources



# New developments with bivariate polar plots

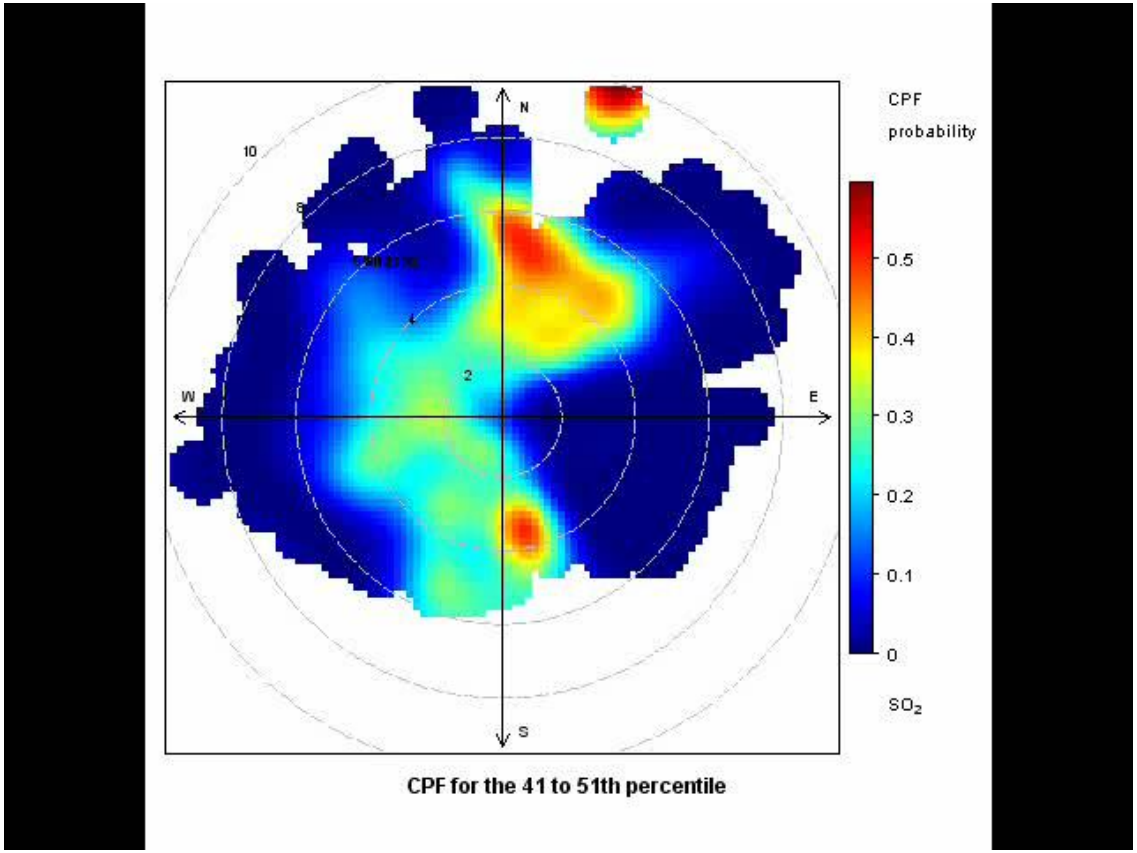
- Consider intervals of concentration and look at the probability of finding such concentrations for specific wind speed-directions
- Can 'scan' full range of concentration intervals
- Sources tend to occupy distinct ranges in concentration
- Can reveal many otherwise 'hidden' sources




Uria-Tellaetxe, I. and D. C. Carslaw (2014). Source identification using a conditional bivariate probability function. *Environmental Modelling & Software*, Vol. 59, 1-9.

# Polar plots – the movie!

- When scanning concentration intervals – can convert to an animation
- Makes it easier to see source signatures ‘emerge’ and then ‘disappear’



# Regional impacts and data analysis

- Many techniques such as polar plots are most informative at the local scale
- At the regional scale **back trajectories** can be very useful
- NOAA make available a global model called Hysplit that can generate back trajectories

- **openair** makes available pre-calculated 96-hour back trajectories at specified locations, run every 3-hours of each year
- Can be used in many different ways to inform air quality data analysis

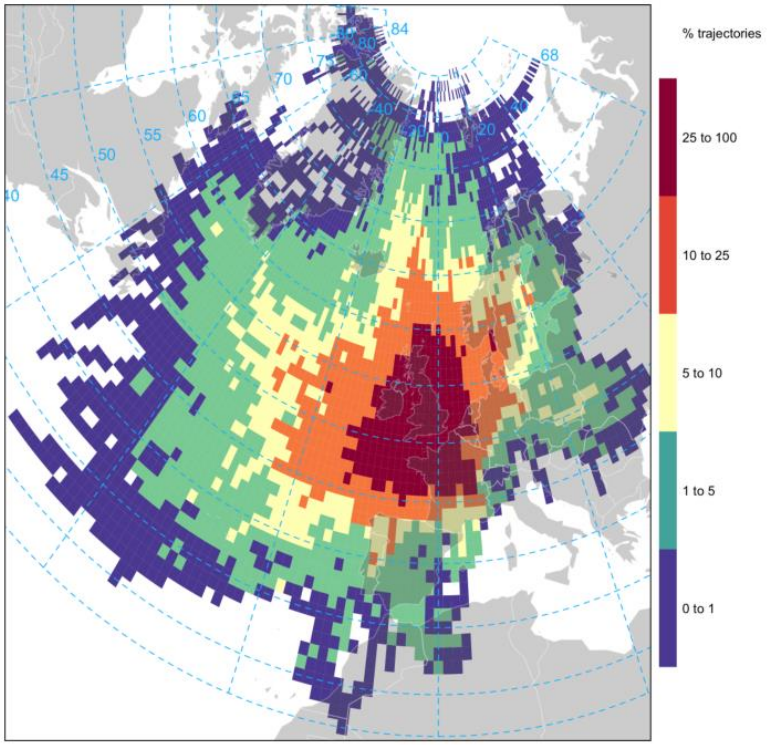
# Regional impacts and data analysis

- Easy to import data

```
traj <- importTraj(site = "london", year = 2010)
```

- And show trajectory frequencies…

```
trajLevel(traj, statistic = "frequency")
```

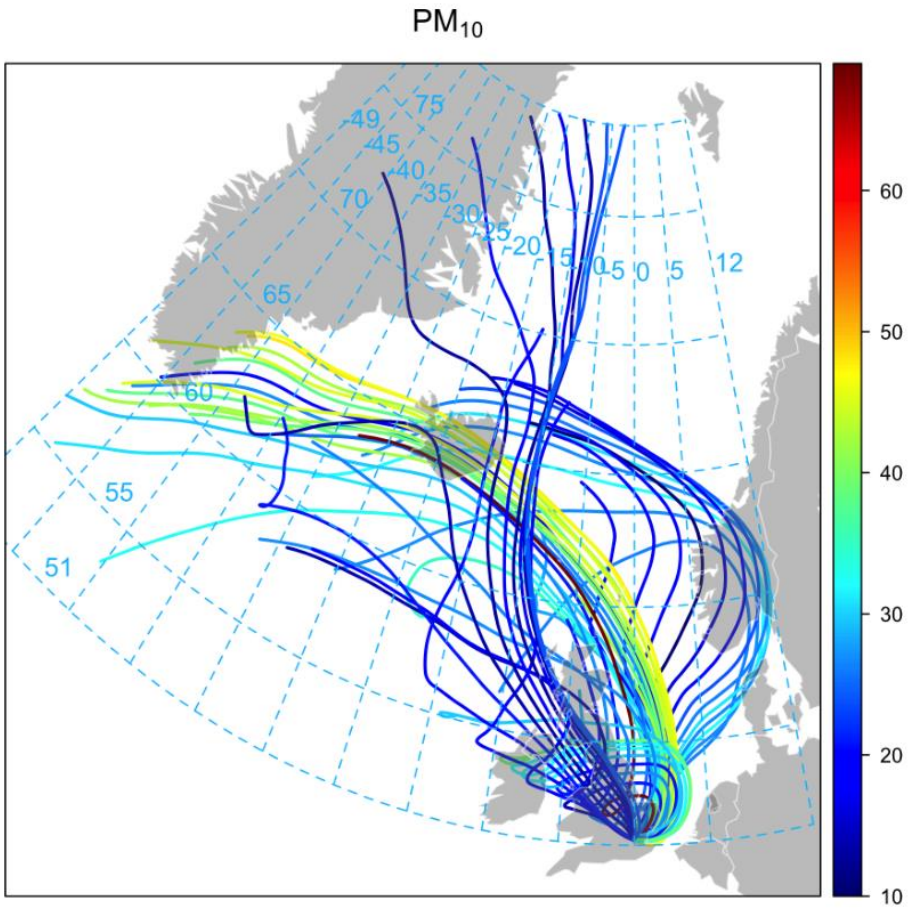



- Many different map projections available and can be used anywhere in the World
- All EMEP sites in Europe available soon (contributed by the University of Edinburgh, Chris Malley)

# Regional impacts and data analysis

- Straightforward to link back trajectories to pollutant concentrations
- Useful for analysing specific pollution episodes in detail
- Having access to more data e.g. sulphate/nitrate – or PM composition data in general can be very useful

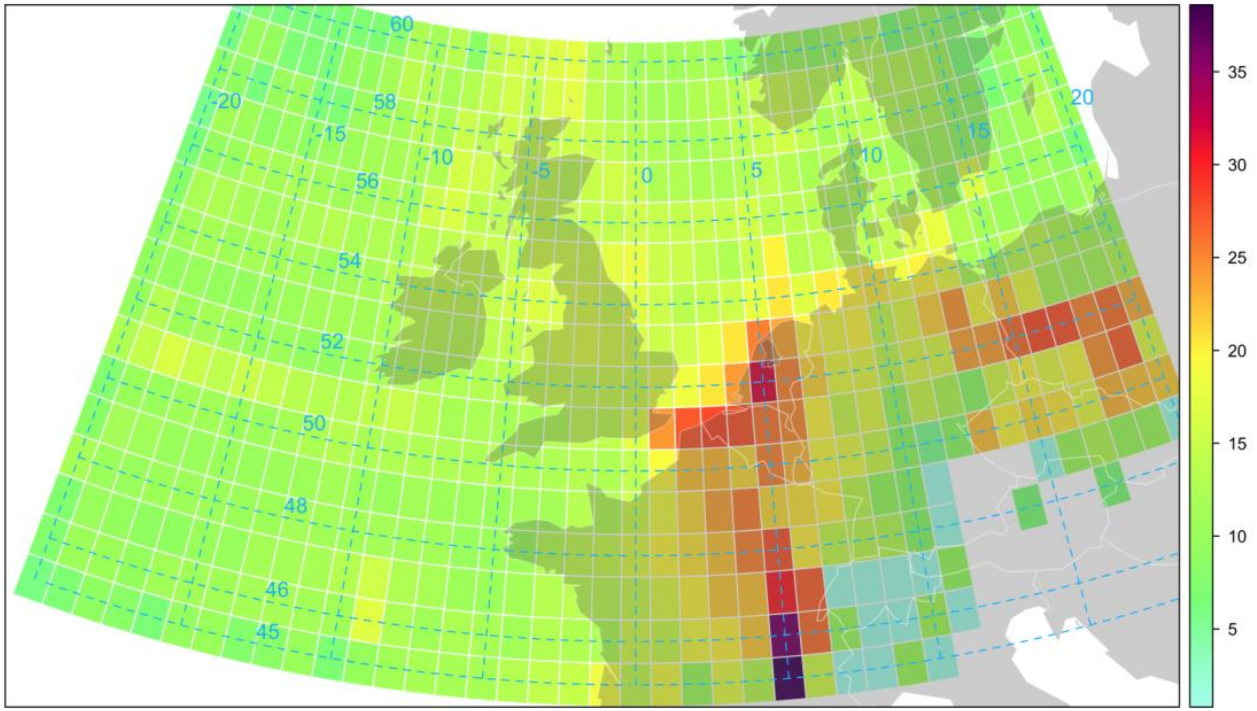
```
trajPlot(selectByDate(traj, start = "15/4/2010", end = "21/4/2010"),
         pollutant = "pm10", col = "jet", lwd =2)
```



# Regional impacts and data analysis

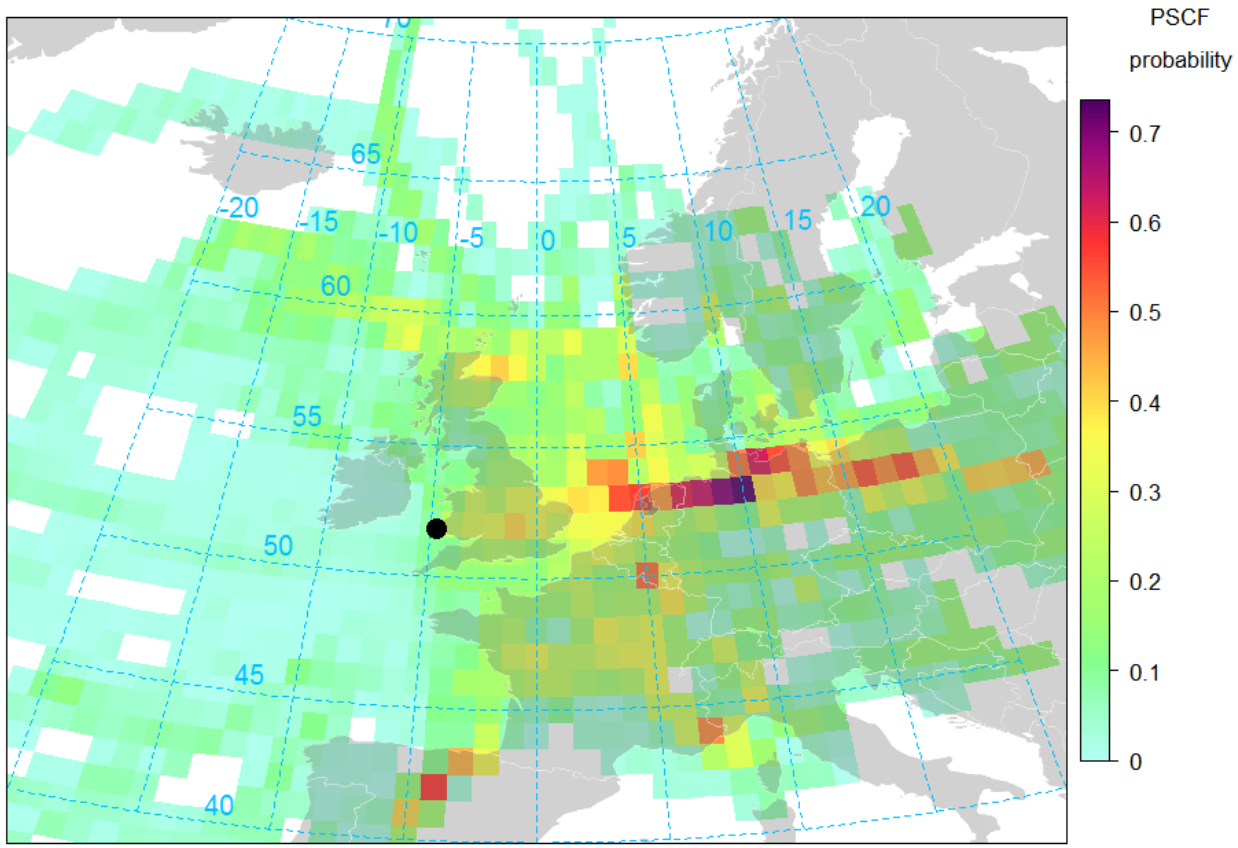
- Can estimate source emission locations
- Plot is the estimated important source regions for $PM_{2.5}$ at the North Kensington site

```
trajLevel(subset(traj,lon > -20 & lon < 20 & lat > 45 & lat < 60),
          pollutant = "pm2.5", statistic="cwt", col = "increment",
          border = "white")
```



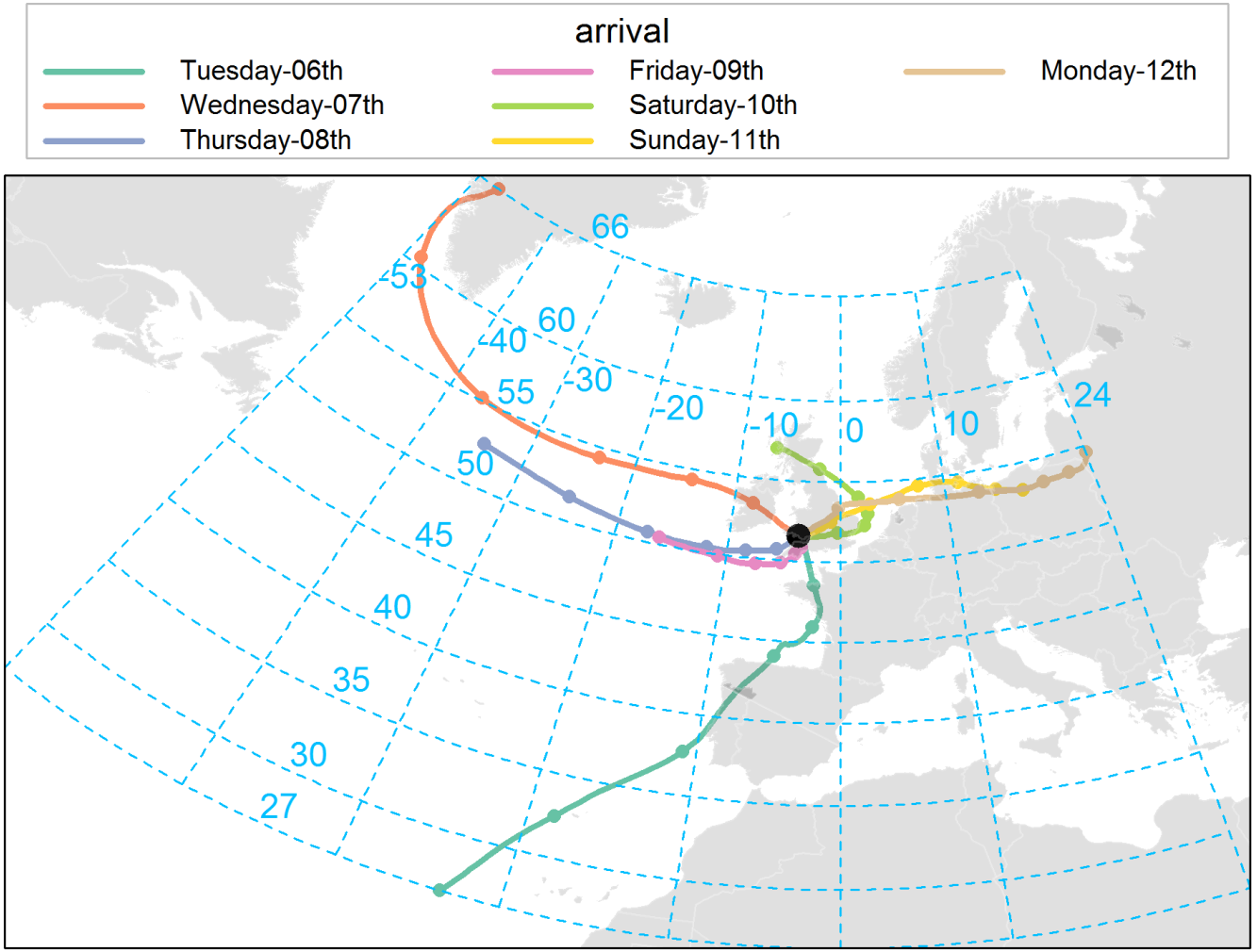
# Origin of highest $PM_{10}$ concentrations at Narberth (2014)

- Associate back trajectories with concentration
- Lots of trajectories over many different paths can give indication of most important **source** origins



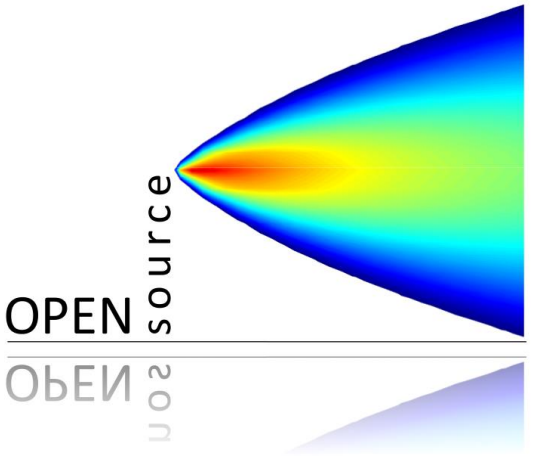
# Where does the air come from?

- Models are available such as Hysplit (from NOAA) that allow back trajectories to be calculated
- Openair has for a long time allowed the analysis of back trajectories
- Developments in progress provide **forecast** back trajectories
- Run daily to give past few and next few days
- Example based on Port Talbot



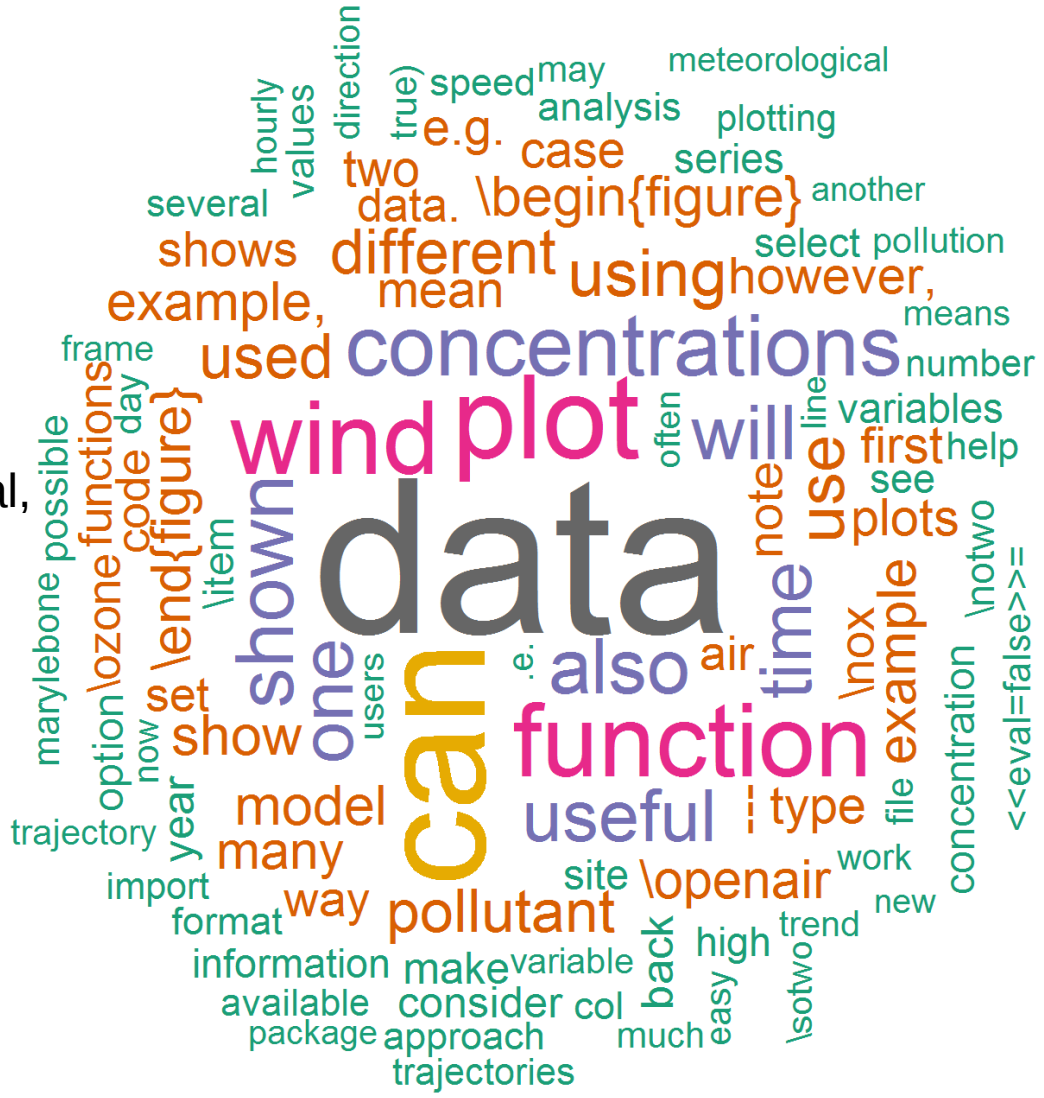
# Future directions

- Make much more data available
  - Better integration with European measurements + interest from the US EPA
- Better web integration
- Plotting on maps more effectively
- Removing the influence of meteorology
  - Understanding air pollution would be much easier if we had the same weather every day!
- Pervasive sensors will become increasingly important
  - What new analysis opportunities will these bring?
- Measurements ‘on the move’ e.g. personal exposure
  - Going beyond plotting colours on a map
- Training courses



# Thank you for your attention

Word cloud of the **openair** manual, generated in R of course…!



David Carslaw

david.carslaw@ricardo.com