The open-source air pollution project

openair

David Carslaw and Karl Ropkins

Towards Smarter Air Quality Analysis
1 October 2009

Outline

Outline

Opportunities and barriers

Analysis of measurement and model output data

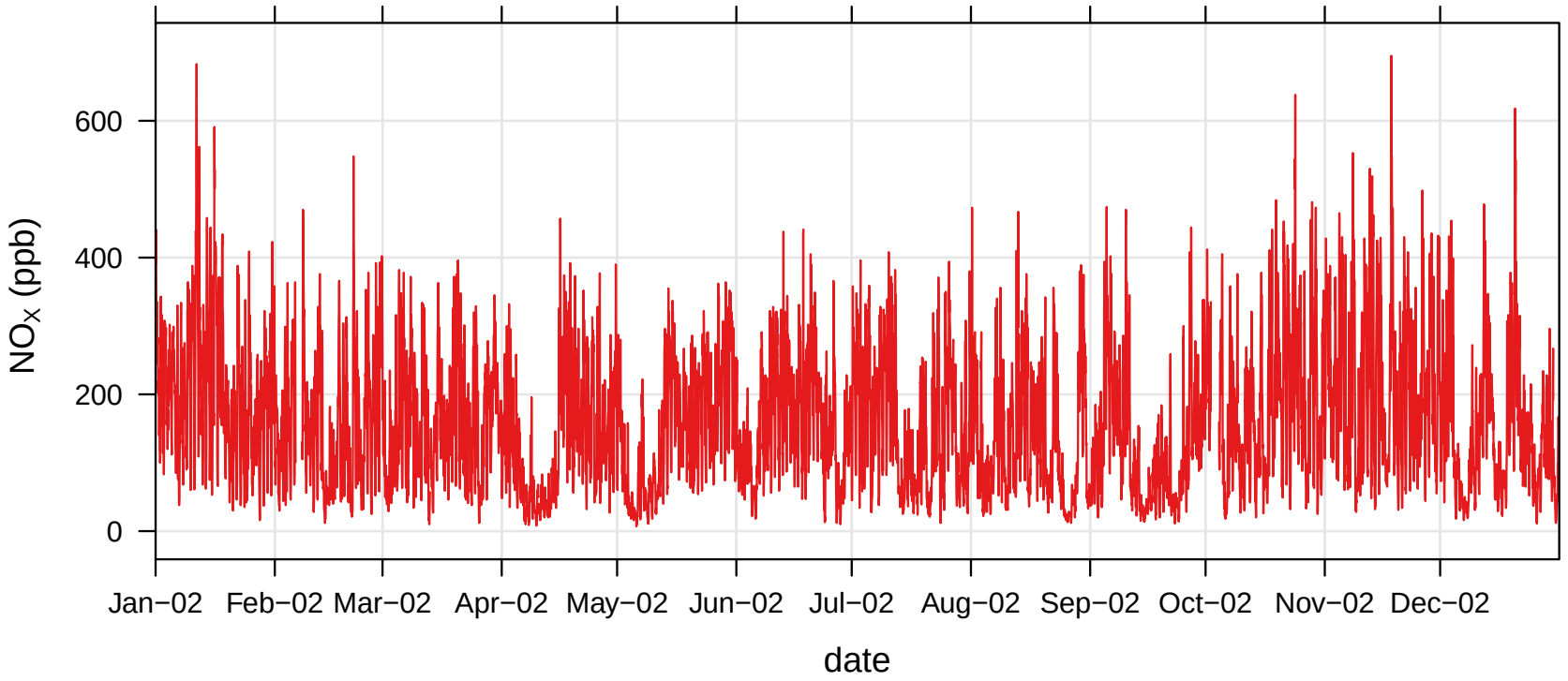
Opportunities

- The analysis of air quality data can provide important insights into air pollution
- Huge amount of data available — but under used
- Insightful analysis provides *evidence* and can reveal unexpected behaviours

Barriers

- No consistent set of tools available to carry out analysis
- Tools can be spread across many different software applications
- Many useful approaches are simply unavailable
- Lack of time, money or ideas about what can be done

The challenge. . .



How to extract meaning and useful information from this?

The openair project

Summary of project

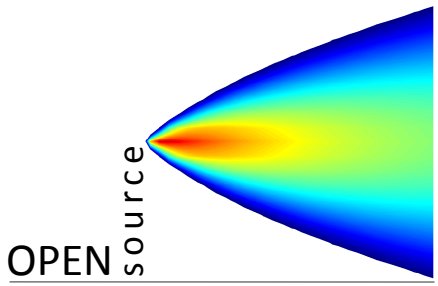
Key points

- 3-year NERC project to October 2011
  - With additional funding from Defra, AEA and several local authorities
- Develop and make available open-source data analysis tools to air quality community
  - Backed up with case studies to show usage and help with appropriate interpretation
- Use **R** statistical software as the platform
  - Highly capable software for "programming with data"
  - Develop a "package" of tools and progressively include advanced methods not widely available
  - Open-source, free and very well supported

Why open-source?

The benefits of an open-source approach

- Availability of the source code, the right to modify it and use it for any purpose
- Allow scrutiny of methods used
- Maximise user input through the ability to contribute and improve the code
- Builds trust — no 'black boxes' and analysis can easily be made reproducible
- Ideal for fully-engaged knowledge exchange (key to NERC funding)
- Through R there are excellent opportunities for international collaboration and dissemination



openair website

Central resource for the project

- Available at www.openair-project.org
- **openair** package – development version
- All documentation, data sets etc.
- News group and newsletters to keep up to date

Data analysis

How best to analyse data?

John Tukey sums it up:

> *"The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data."*

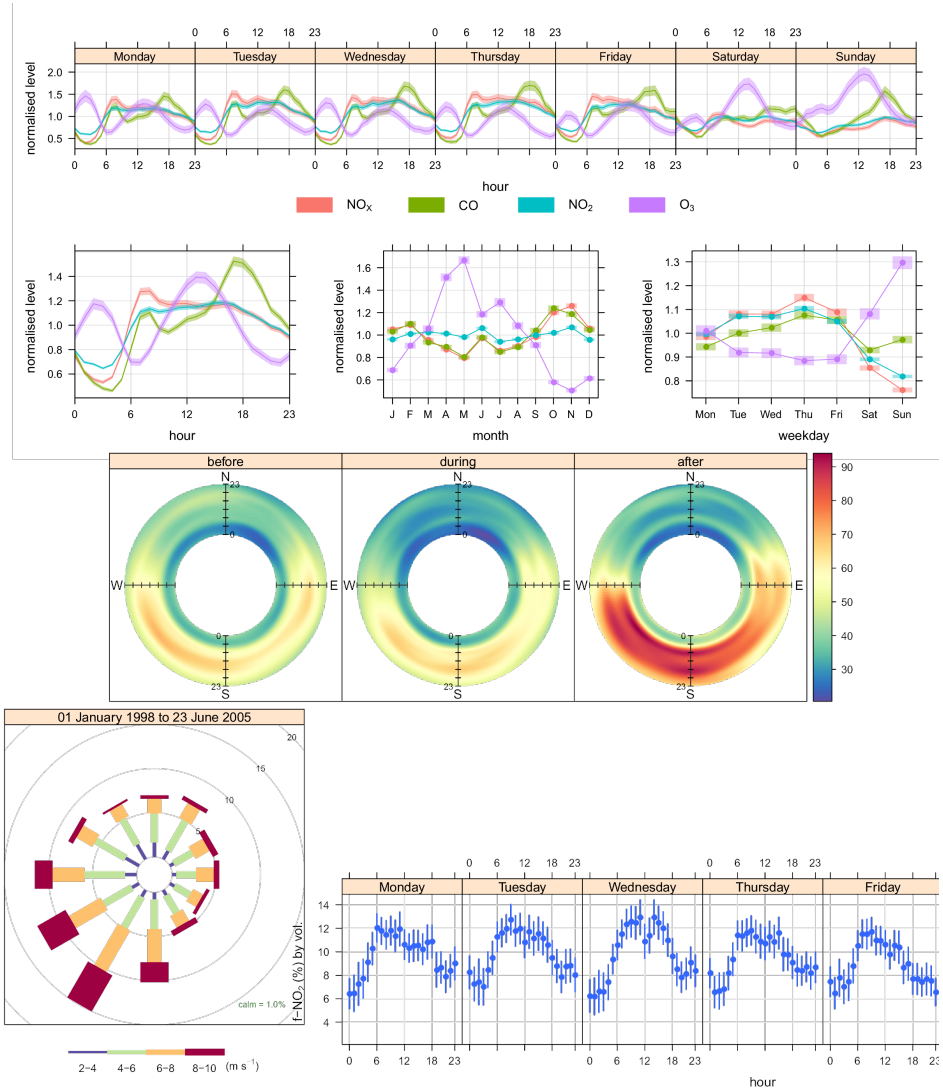
- Data analysis is most useful when built around specific questions (need an aching desire), however...
- Exploratory data analysis can be very insightful and is under-used — but time consuming
- Case studies can provide fresh thinking and new ideas about **what** can be done and **how** to draw inferences from data

Outline

Brief examples of openair functions

Lots already available — highlight one

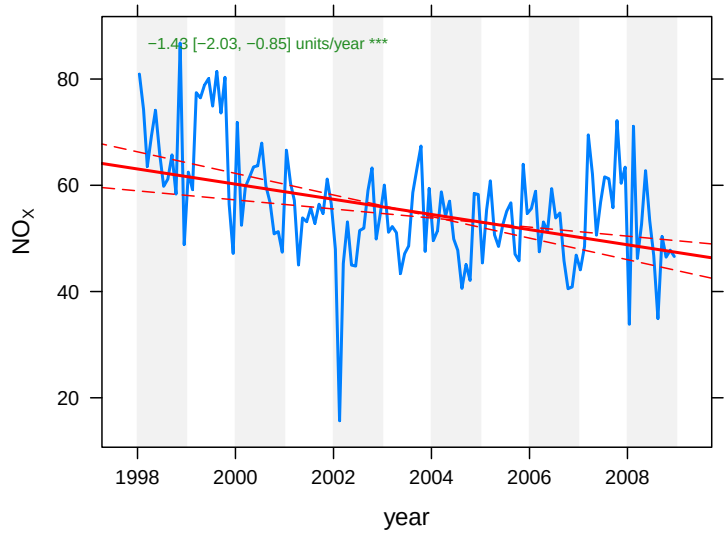
- Wide range available — covered fully in documentation
- Here we outline the approaches used
- Consider trends — in a flexible way
- Highlight the idea of **conditioning**



Mann-Kendall analysis of trends

Consider trends in NO_X concentrations at London Bloomsbury

- **Mann-Kendall** analysis often used for environmental time series
- Consider monthly trends with option to de-seasonalise the data
- Use bootstrap simulation techniques to estimate uncertainties and block bootstrap to deal with autocorrelation



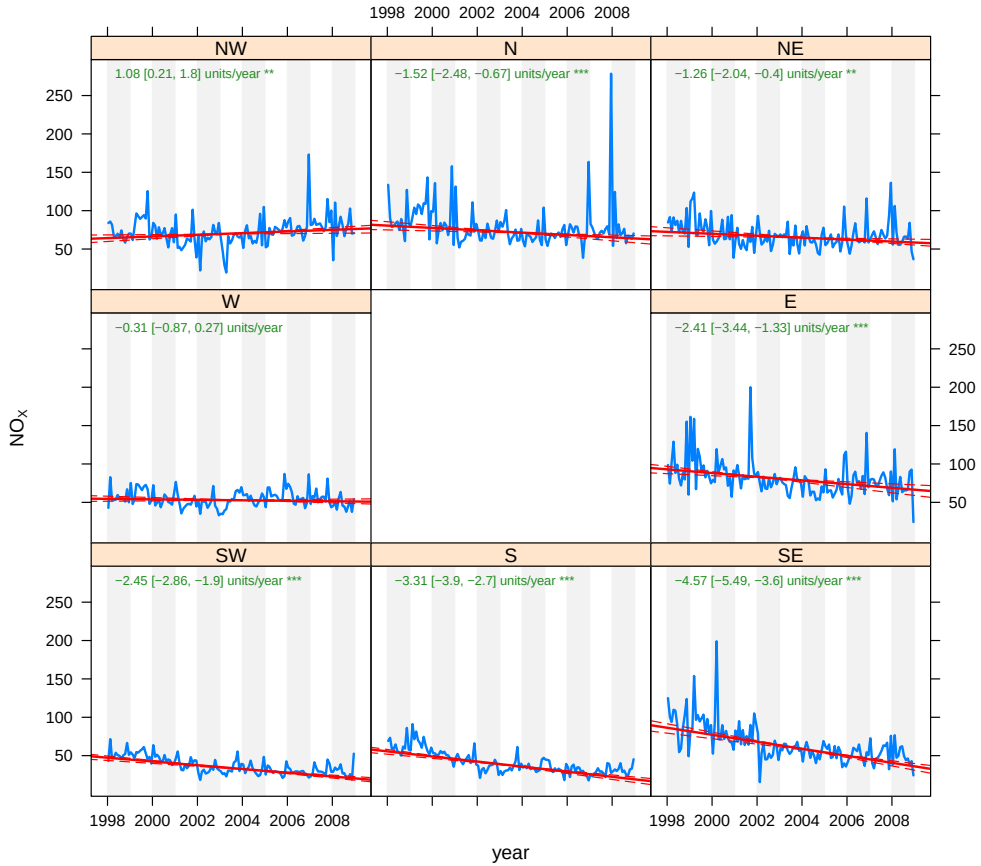
Example

Read some data in and plot the trend

```
mydata = import("d:/data/bloomsbury.csv")
MannKendall(mydata, pollutant = "nox", deseason = TRUE)
```

Mann-Kendall analysis of trends

Trends by wind sector

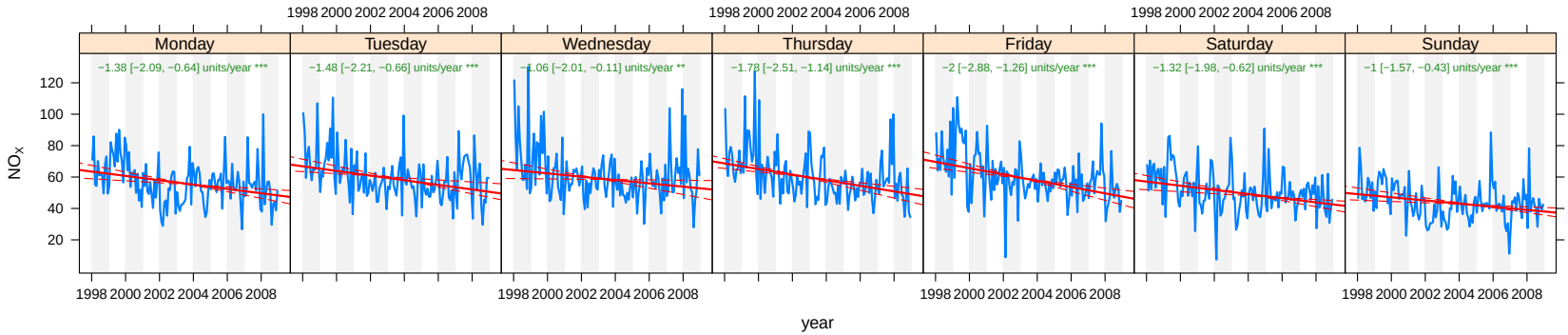


Example

MannKendall(mydata, pollutant = "nox", deseason = TRUE, type = "wd")

Mann-Kendall analysis of trends

Trends by day of the week

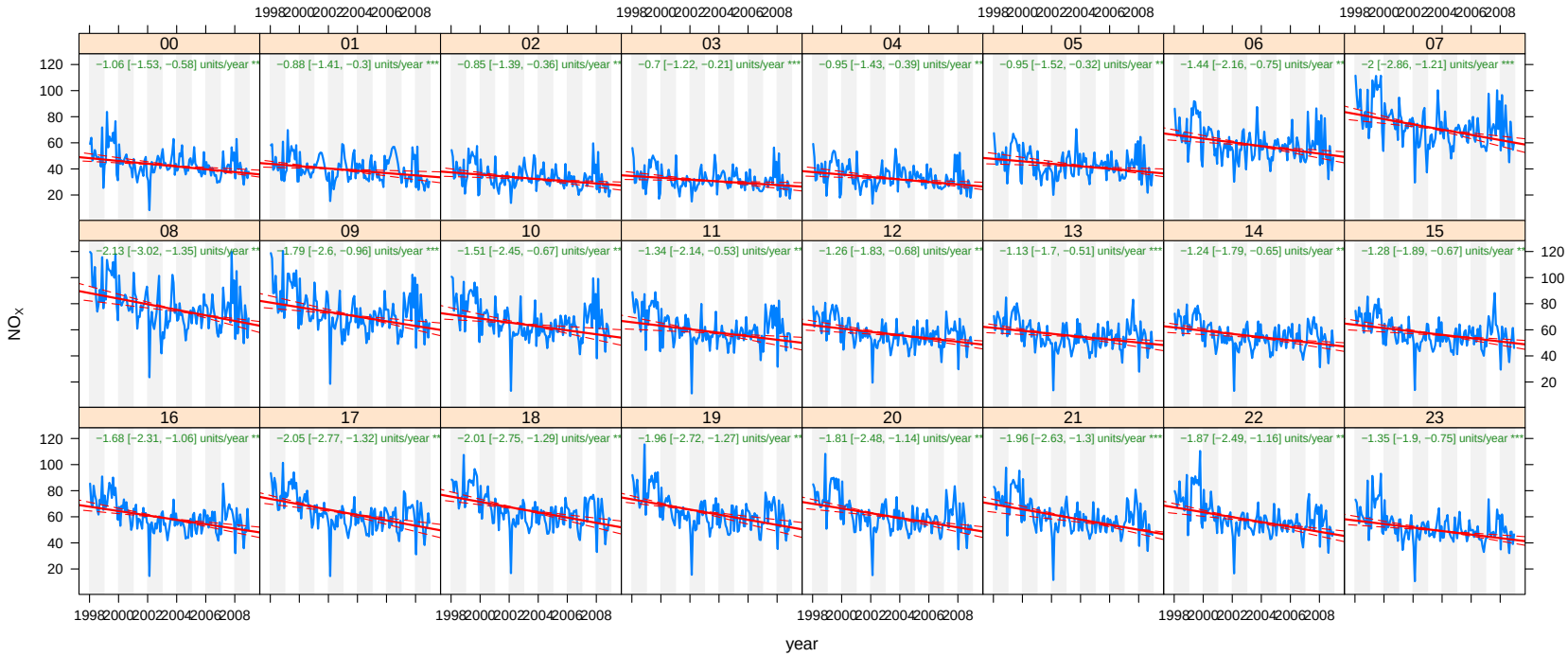


Example

MannKendall(mydata, pollutant = "nox", deseason = TRUE, type = "weekday")

Mann-Kendall analysis of trends

Trends by hour of the day

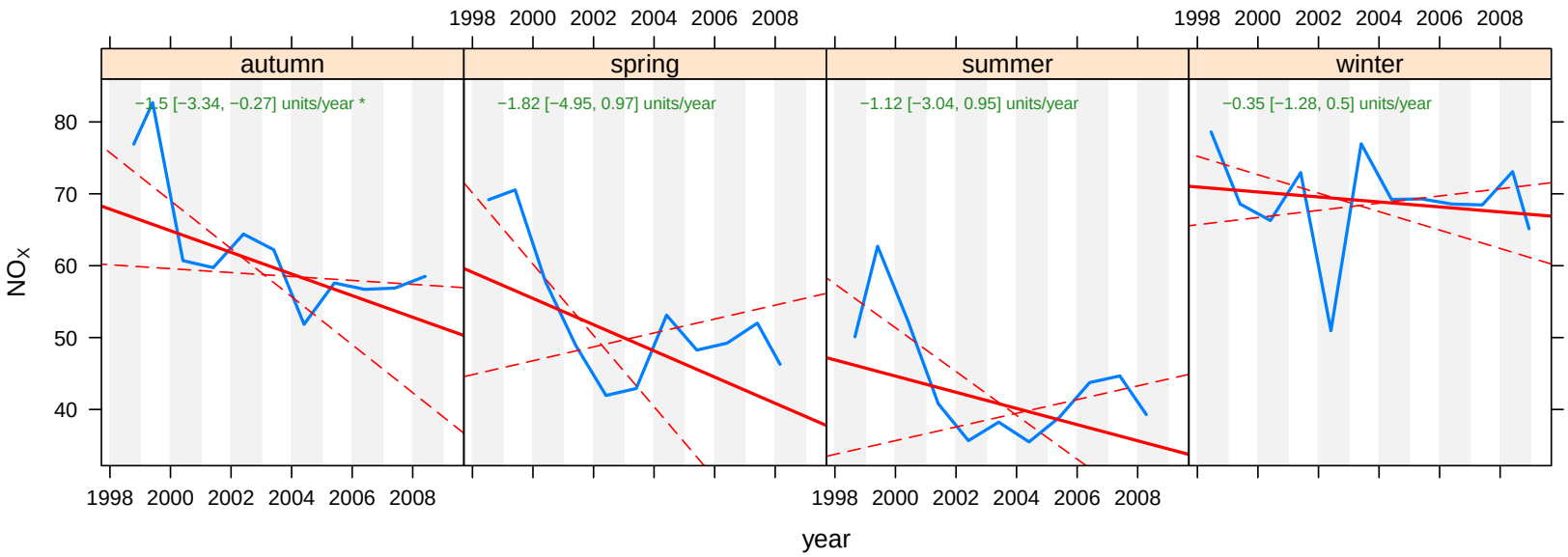


Example

MannKendall(mydata, pollutant = “nox”, deseason = TRUE, type = “hour”)

Mann-Kendall analysis of trends

Trends by season



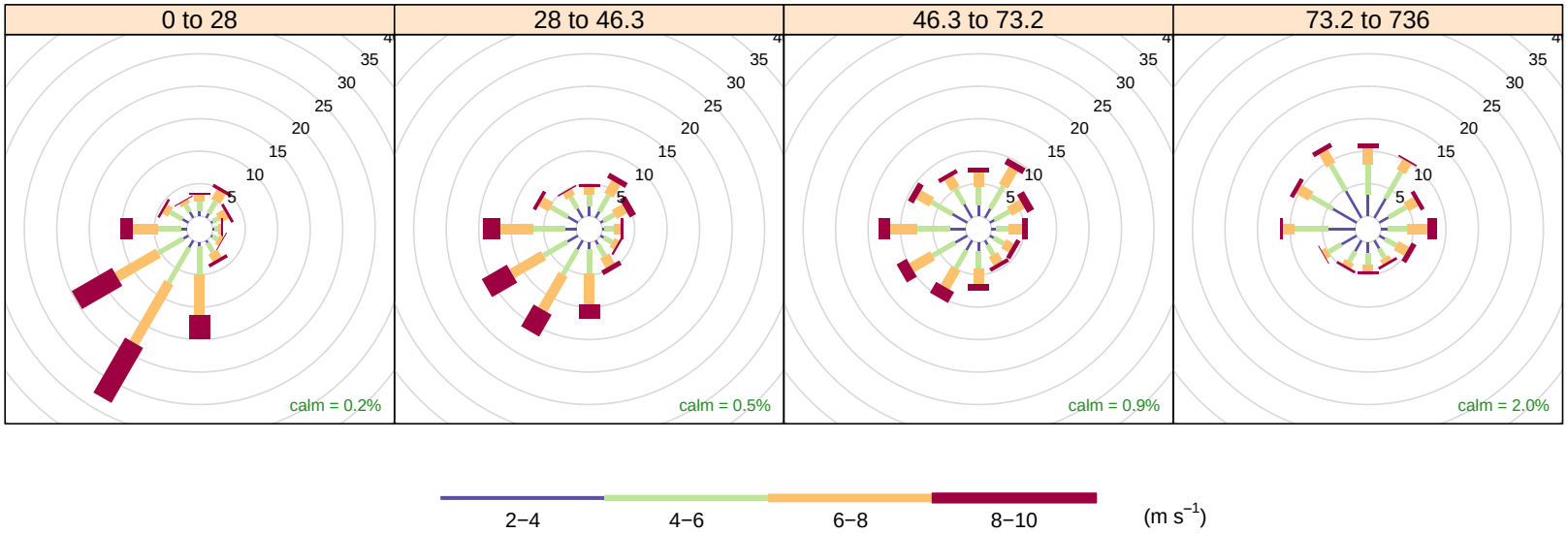
Example

MannKendall(mydata, pollutant = "nox", deseason = TRUE, type = "season")

A different kind of wind rose

How do met conditions vary by...

Consider meteorology dependent on concentration of NO_X

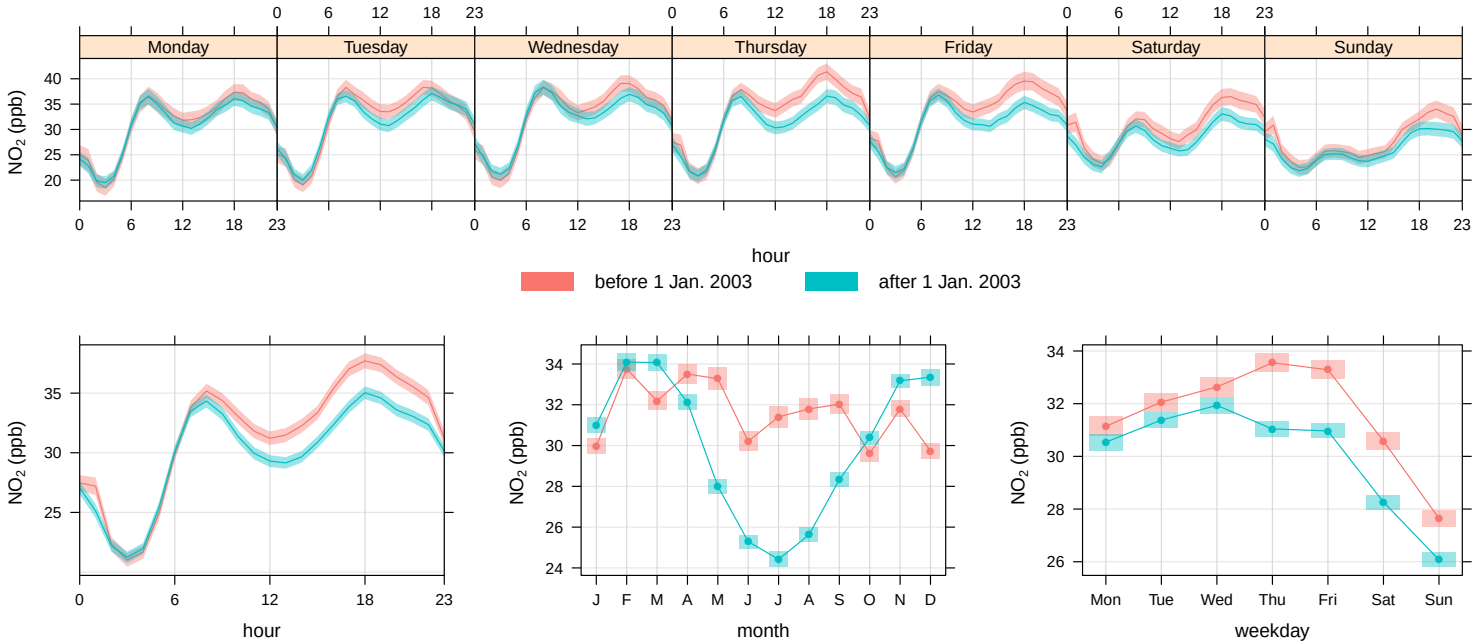


Example

wind.rose(mydata, type = "nox")

How do concentrations vary by time?

Looking at interventions



Example

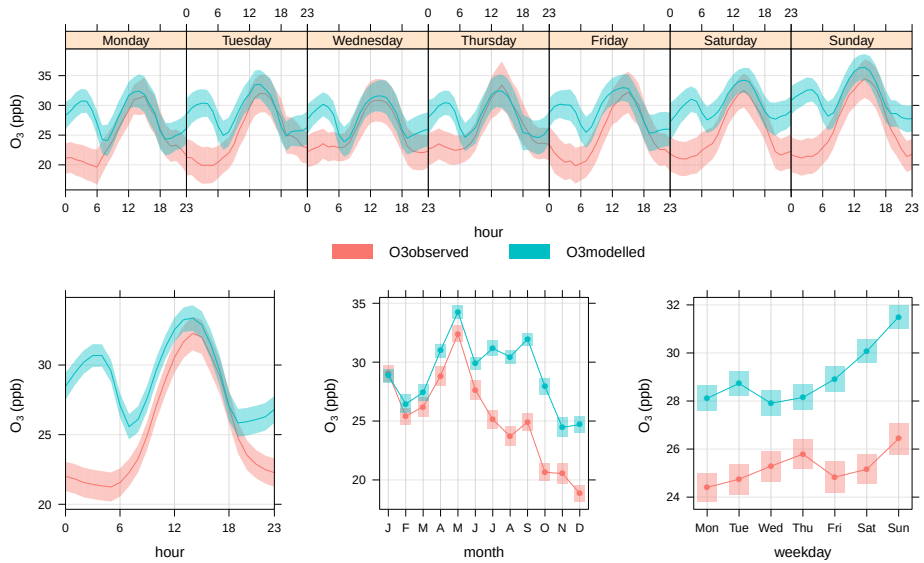
mydata = split.by.date(mydata, date = "1/1/2003", labels = c("before 1 Jan. 2003", "after 1 Jan. 2003"))

time.variation(mydata, pollutant = "no2", type = "site", ylab = "no2 (ppb)")

Tools for model evaluation

CMAQ model output

- Evaluating models is important
- Emphasis is on quantitative 'metrics' e.g. fractional bias
- Scope for better *understanding* of model performance



Example

time.variation(cmaq, pollutant = "o3", type = "site")[a]

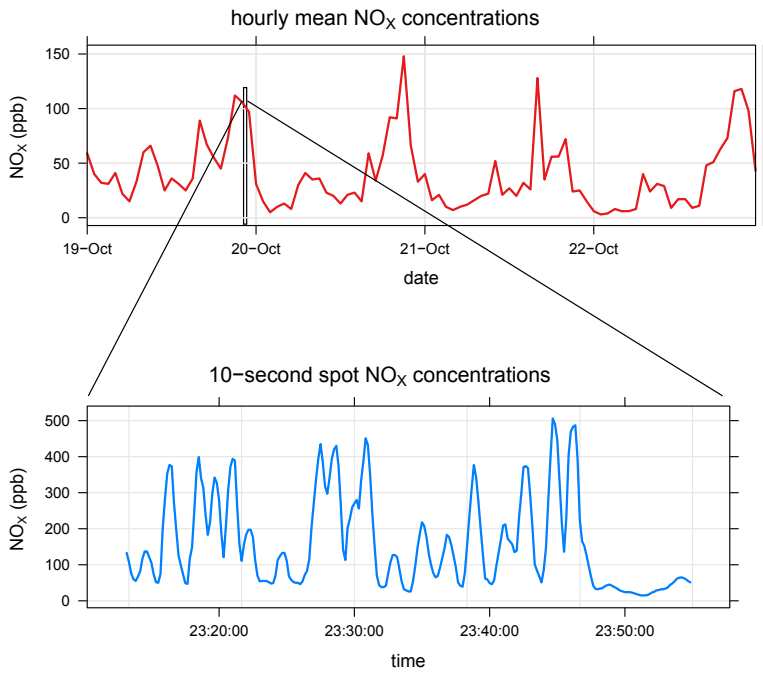
[a] Thanks to Sean Beevers, King's College London for CMAQ output.

Outline

Higher time resolution measurements

What are the benefits?

- Loss of useful information with hourly mean data
- Measurements of NO_X close to Heathrow Airport
  - Hourly means show broad variation in source emissions
  - 10-second measurements reveal individual aircraft plumes
  - Many new insights[a]

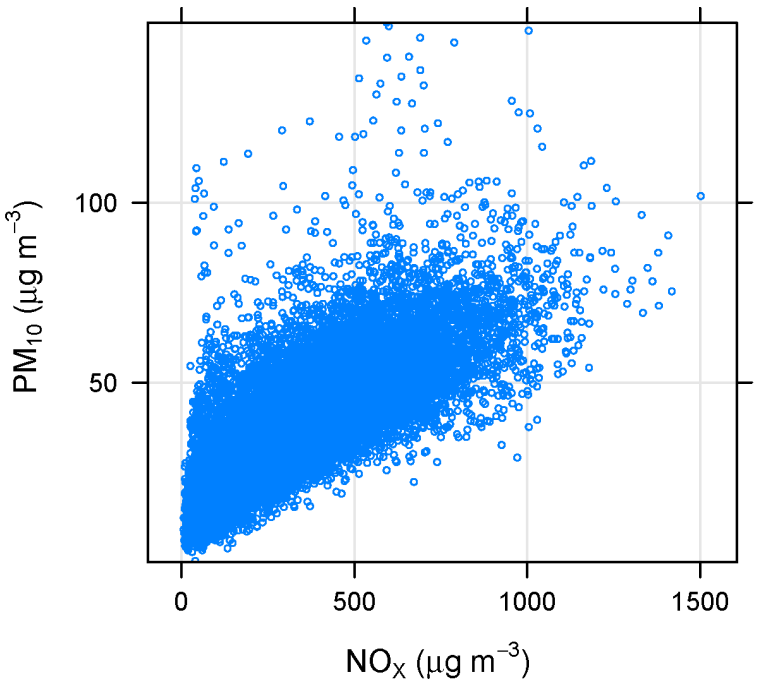


[a] Carslaw et al. (2008). Near-field commercial aircraft contribution to nitrogen oxides by engine, aircraft type and airline by individual plume sampling. *ES&T*. 42(6): 1871-1876.

Comparisons with emission inventories

Alternative look at pollutant ratios and relationships

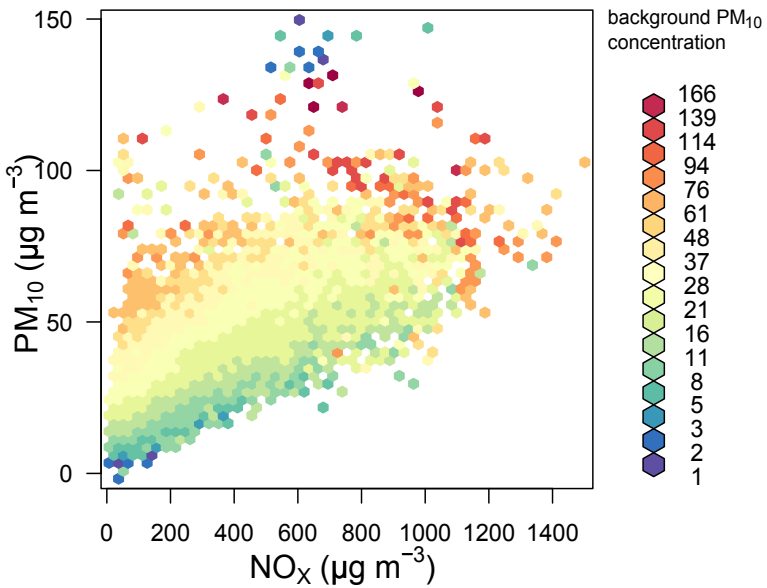
- Pollutants ratios/relationships often used for comparison with emission inventories
- Scatter plots contain potentially lots of useful information
- Much of the variation is due to background concentrations
- What to do if no background concentrations available?



Comparisons with emission inventories

Alternative look at pollutant ratios and relationships

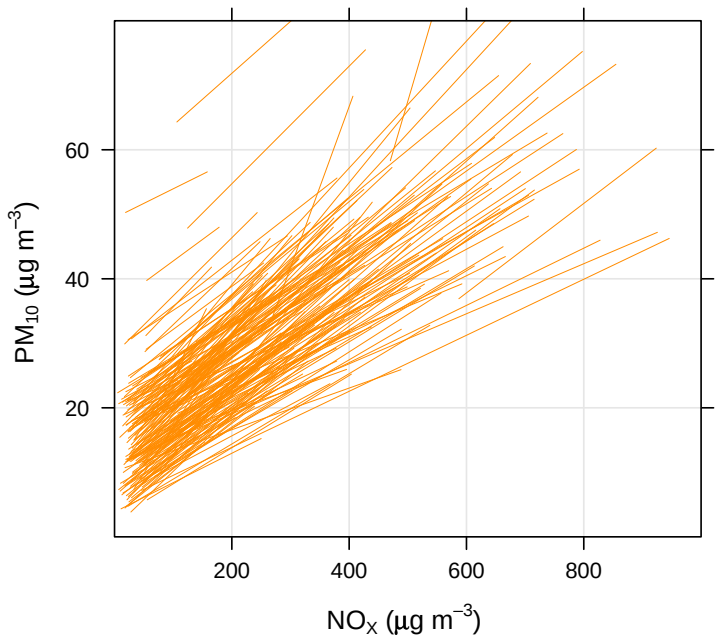
- Pollutants ratios/relationships often used for comparison with emission inventories
- Scatter plots contain potentially lots of useful information
- Much of the variation is due to background concentrations
- What to do if no background concentrations available?



Search for linear patterns in the data

Estimation of pollutant ratios in the absence of background data

- Split data into 3-hour non-overlapping blocks[a]
  - Fit regression line to each 3-hour block and calculate slope
  - Filter for slopes with high R^2
- $\Rightarrow$ Linear patterns in data
- Mode of slope is a good estimate of mean PM_{10}/NO_X ratio

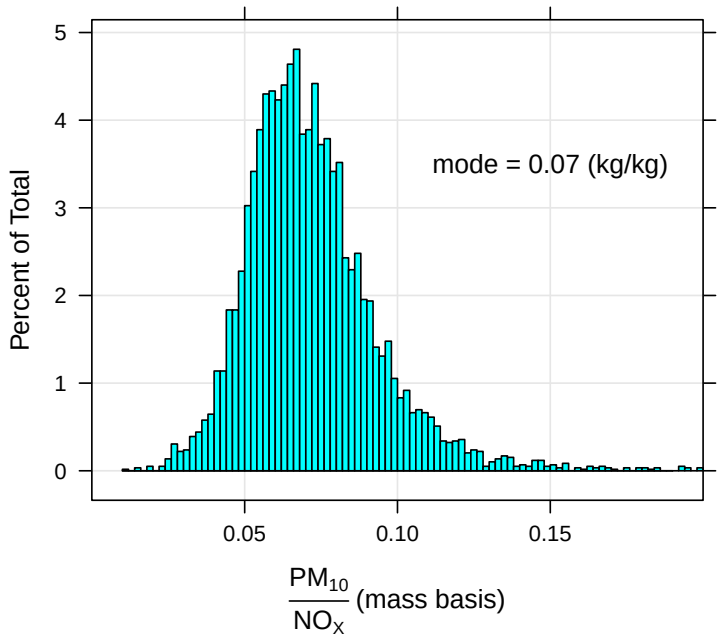


[a] Bentley, S.T. (2003). Graphical techniques for constraining estimates of aerosol emissions from motor vehicles using air monitoring network data. *Atmos. Env.*, 1491–1500.

Search for linear patterns in the data

Estimation of pollutant ratios in the absence of background data

- Split data into 3-hour non-overlapping blocks[a]
  - Fit regression line to each 3-hour block and calculate slope
  - Filter for slopes with high R^2
- $\Rightarrow$ Linear patterns in data
- Mode of slope is a good estimate of mean PM_{10}/NO_X ratio



[a] Bentley, S.T. (2003). Graphical techniques for constraining estimates of aerosol emissions from motor vehicles using air monitoring network data. *Atmos. Env.*, 1491–1500.

Developments

1. Reviewing scientific literature and will adopt promising approaches e.g.
   - Source identification and characterisation
   - Better quantitative analysis
   - Further development of tools for model evaluation
2. The **openair** package
   - Graphical-user interface (GUI)?[1]
   - Remote repository with full version control and easier installation
   - Work towards access via AURN archive and LAQN
   - Reproducible analyses using *Sweave*, R and LaTeX

[1] A researcher from another university has started this. . .

Developments

1. Reviewing scientific literature and will adopt promising approaches e.g.
   - Source identification and characterisation
   - Better quantitative analysis
   - Further development of tools for model evaluation
2. The **openair** package
   - Graphical-user interface (GUI)?[1]
   - Remote repository with full version control and easier installation
   - Work towards access via AURN archive and LAQN
   - Reproducible analyses using *Sweave*, R and LaTeX

[1] A researcher from another university has started this...

Thank you for you attention...

Questions?

David Carslaw
d.c.carslaw@its.leeds.ac.uk