WACL R Training

Training for air pollution data analysis in R

Will Drysdale and Stuart Lacy

14th & 21st Nov 2023

University of York

Introduction

Welcome!

A 2-day course introducing the R statistical programming language

- Introduction to R, RStudio and programming for beginners
- Building a script; the benefits of programming over spreadsheets
- Reading, manipulating and visualising data, with tips and tricks to solve common problems
- A focus on the types of skills you'll need for working with air quality data, using real-world datasets
- Chance to practise skills with us on hand to help out

Approaches

- Authentic, live coding
- All course material is available on GitHub
  - Includes all data and script files produced during this course
  - A bespoke self-teaching document will also be made available
  - Useful for post-course learning
- All material used in this course will be **entirely reproducible**
  - This means that you will be able to recreate all the outputs shown during the course (and afterwards)
- Questions are encouraged, and one of us will always be at hand to solve problems

Topics to be covered

Tuesday 14th November 2023, 10:00-16:00

- Introduction to R for Air Quality Data
    - Reading and interrogating data within R (recapping pre-course prep)
    - Introducing statistical analysis; averages and trend lines
    - Using `openair` for air quality data analysis
    - Reading and combining multiple data streams
    - Further data handling; reshaping, grouping and summarising

Tuesday 21st November 2023, 10:00-16:00

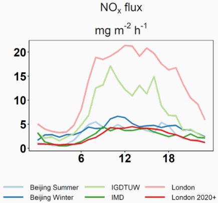
- Data visualisation
    - Introduction to the `ggplot2` plotting library
    - Examples of different plot types
    - Making publication standard visualisations

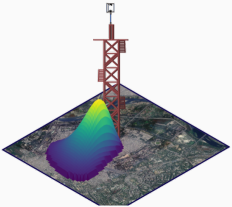
Who are we?

Will Drysdale

I use R for:

- **Eddy Covariance** - processing of high time resolution data (5 - 20 Hz) to calculate emissions using *eddy4R*
  - Perform analysis **automatically** and **reproducibly**
  - **Collaborate** with developers to add our own tools



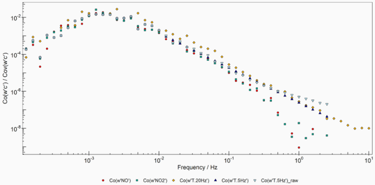


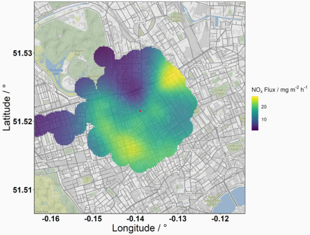
Who are we?

Will Drysdale

I also use R in many other aspects of my work:

- **Instrument data** work up
- Producing **Figures**
- **Mapping** spatial data



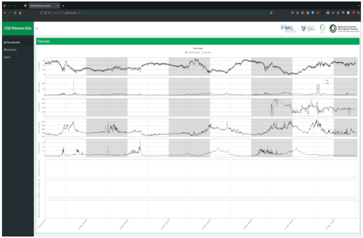


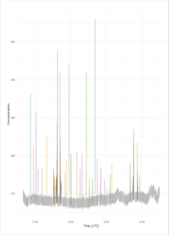
Who are we?

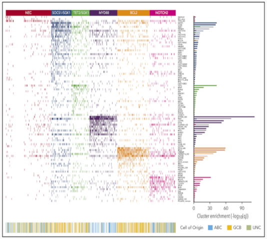
Stuart Lacy

I use R for:

- Statistical modelling / Machine learning
- Writing web app data dashboards (using `Shiny`)
- Developing reproducible data tools



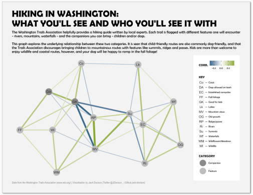




Who are we?

Jack Davison

- High quality data viz

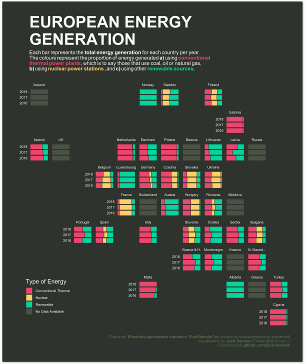


HIKING IN WASHINGTON:
WHAT YOU'LL SEE AND WHO YOU'LL SEE IT WITH

The Washington Trails Association helpfully provides a hiking guide written by local experts. Each trail is flagged with different features one will encounter - rivers, mountains, waterfalls - and the companions you can bring - children and/or dogs.

This graph explores the underlying relationship between these two categories. It is seen that child-friendly routes are also commonly dog-friendly, and that the Trails Association discourages bringing children to mountainous routes with features like summits, ridges and passes. Kids are more than welcome to enjoy wildlife and coastal routes, however, and your dog will be happy to romp in the fall foliage!

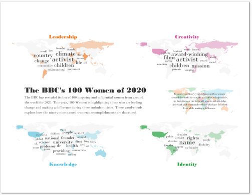
Data from the Washington Trails Association (www.wta.org) | Visualisation by Jack Davison (Twitter @JDavison_ | Github jack-davison)



EUROPEAN ENERGY GENERATION

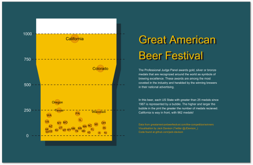
Each bar represents the **total energy generation** for each country per year. The colours represent the proportion of energy generated **a)** using **conventional thermal power plants**, which is to say those that use coal, oil or natural gas, **b)** using **nuclear power stations** , and **c)** using other **renewable sources**.

Data from 'Electricity generation statistics - First Results' (ec.europa.eu/eurostat/statistics-explained)
Visualisation by Jack Davison (Twitter @JDavison_)
Code found at github.com/jack-davison



The BBC's 100 Women of 2020

The BBC has revealed its list of 100 inspiring and influential women from around the world for 2020. This year, '100 Women' is highlighting those who are leading change and making a difference during these turbulent times. These word-clouds explore how the ninety-nine named women's accomplishments are described.

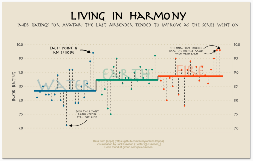


Great American Beer Festival

The Professional Judge Panel awards gold, silver or bronze medals that are recognized around the world as symbols of brewing excellence. These awards are among the most coveted in the industry and heralded by the winning brewers in their national advertising.

In this beer, each US State with greater than 25 medals since 1987 is represented by a bubble. The higher and larger the bubble in the pint the greater the number of medals recieved. California is way in front, with 962 medals!

Data from greatamericanbeerfestival.com/the-competition/winners
Visualisation by Jack Davison (Twitter @JDavison_)
Code found at github.com/jack-davison



LIVING IN HARMONY

IMDB RATINGS FOR AVATAR: THE LAST AIRBENDER TENDED TO IMPROVE AS THE SERIES WENT ON

Data from {appa} (https://github.com/averyrobbins1/appa)
Visualisation by Jack Davison (Twitter @JDavison_)
Code found at github.com/jack-davison

Who are you?

Introductions

- What is your name?
- What do you do?
- What kind of data do you use?
    - Big? Small? From the lab? Fieldwork? Modelled? Time-series? Categorical?
- What type of data analysis do you do?
- What are you hoping to get out of these sessions?

Further Help

Learning R does not finish at the end of this short course

- There are many R users in WACL who are happy to help, including ourselves.
- There are lots of resources online that we'll point you to.
- WACL has a programming Slack channel for help with R & Python.
- If there is interest, we'll look to do shorter sessions on more specific problems

Exercises

Set-up

- Ensure you have cloned this repository
- Make sure your packages are loaded!

```
library(openair)
library(dplyr)
```

Day 1 - Morning

Reading in data

Read the data into R, correct any mistakes, and plot a simple timeseries of NO2:NOx.

- They will need combining.
- Make sure the columns are correctly formatted (remember `class()`)
- You will need to create a new column (use `mutate()`)

Statistics

Find the mean and standard deviation of your NO2:NOx column, and fit a trendline of NOx $\sim$ NO2.

- You may need to drop NA values.
- Use `lm()` to fit the line, and `summary()` / `coef()` to pull coefficients.
- It may be a nice idea to plot a scatter with a trend-line.

Exploring `openair`

Read the openair book and have a go at some analysis.

- https://bookdown.org/david_carslaw/openair/
- Try out some different plots. Can you plot all the pollutants on one graph? What about by year?
- If you would like more data, `openair::mydata` contains some extra timeseries data that is ready to go!

Day 1 - Afternoon

Reading in data

Read in the same data from this morning, but using a more reproducible workflow.

- Use lists and loops, similar to the example.
- This time, you will be column-binding rather than row-binding.

Data Manipulation

***Tidy* your data using `pivot_longer()`, then find the mean and standard deviation of each pollutant using `dplyr`. Also find the hour at which each pollutant peaks in the data frame.**

- You'll not be able to work out the mean/sd *and* find the time at which the pollutants peak in the same pipeline; think about how you structure your script to avoid repetition.
- You could also use `mutate()` and `lubridate` to get an average *per year*.

Making visualisations with `ggplot2`

Use `ggplot2` to plot NO, NO2 and NOx timeseries.

- Remember you can assign colours using `aes(color = column)` and split the plot using `facet_wrap()` - what looks best?
- Is the plot too messy? Could you time average a bit to make it clearer?
- Can you plot a linear trendline using `geom_smooth()`?

Real World Exercise

Preparation

- Complete one or more of these challenges!
- These are much more open-ended than things we've done so far and gives you a taste of `real' data science.
- If you're struggling, work together and ask for help!

Challenge 0: Reading Data

Read in the data.

- To start, you'll need to read in the data.
- Remember that data often isn't read in perfect and ready to use - you may need the skills you've learned yesterday and today.
- Is the data ready to be used with `openair`?

Challenge 1: Diurnal Profiles

A key part of using time series data is plotting diurnal profiles. Can you plot some for, e.g., ozone?

- By the end, these should be plotted per airmass history (`Flag_name` column).
- To start, try plotting with `openair::timeVariation()` to see what they should look like.
- Can you recreate the `openair` plot using `dplyr` and `ggplot2`?

Challenge 2: Data Exploration and Validation

Real world instrument data can have some weird features - can you identify dodgy data in the VOCs?

- Tip: Use `select(contains("pp"))` to just select the VOCs (and some others).
- Calculate some summary statistics for the VOCs.
- Can you use simple visualisations to identify anomalous points? Can these be removed?

Challenge 3: Many Linear Models (& Elegant Scripting)

One of the most common statistical analyses is finding the correlation between two values. Can you find the correlation between CO2 and *every* VOC in the dataset?

- Recall the use of `lm()` to do linear modelling.
- You could do each VOC manually - but can you think of a better way?
- We haven't explicitly done this in the course, but you could use other things you've learned.

Summary

Read in the data and attempt any of Challenges 1, 2 or 3.

- (**Challenge 0:** Read in the data, fixing any issues.)
- **Challenge 1:** Plot reactive species diurnals in `ggplot2`.
- **Challenge 2:** Filter or flag anomalously high VOC data.
- **Challenge 3:** Find a concise way to correlate all VOC columns to CO2.