# Introduction to transport data science

Robin Lovelace University of Leeds, 2020-02-05


Note: before you run this tutorial, ensure that you have recently updated R and RStudio on your computer. Furthermore, you will need to have installed a number of packages, as described here: https://docs.ropensci.org/stats19/articles/stats19-training-setup.html (https://docs.ropensci.org/stats19/articles/stats19-training-setup.html)

There is one additional package that you will need that is not available on CRAN which can be installed as follows (see Chapter 2 of Geocomputation With R (https://geocompr.robinlovelace.net/spatial-class.html) for details, this requires the package remotes):

```
# install.packages("remotes")
remotes::install_github("Nowosad/spDataLarge")
```

# Thinking about (transport) data science (30 minutes)

- Based on the contents of the lecture, come up with *your own* definition of data science
- Name 2 advantages and 2 disadvantages of this approach to transport research
- How do you see yourself using data science over the next 1 year, 5 years, 20 years
- Quick go around: what is your name, level and background?
- Get into groups of 2-3 and discuss:

## In groups of 2-4

- What do you hope to get out of it personally?
- In terms of future work in an evolving job market?
- In terms of the kinds of problems you want to solve?

# Sketching research methods (in groups of 2-4, 30 minutes)

Starting with the 1000 ‘desire lines’ dataset of Leeds, sketch-out some research ideas that cover

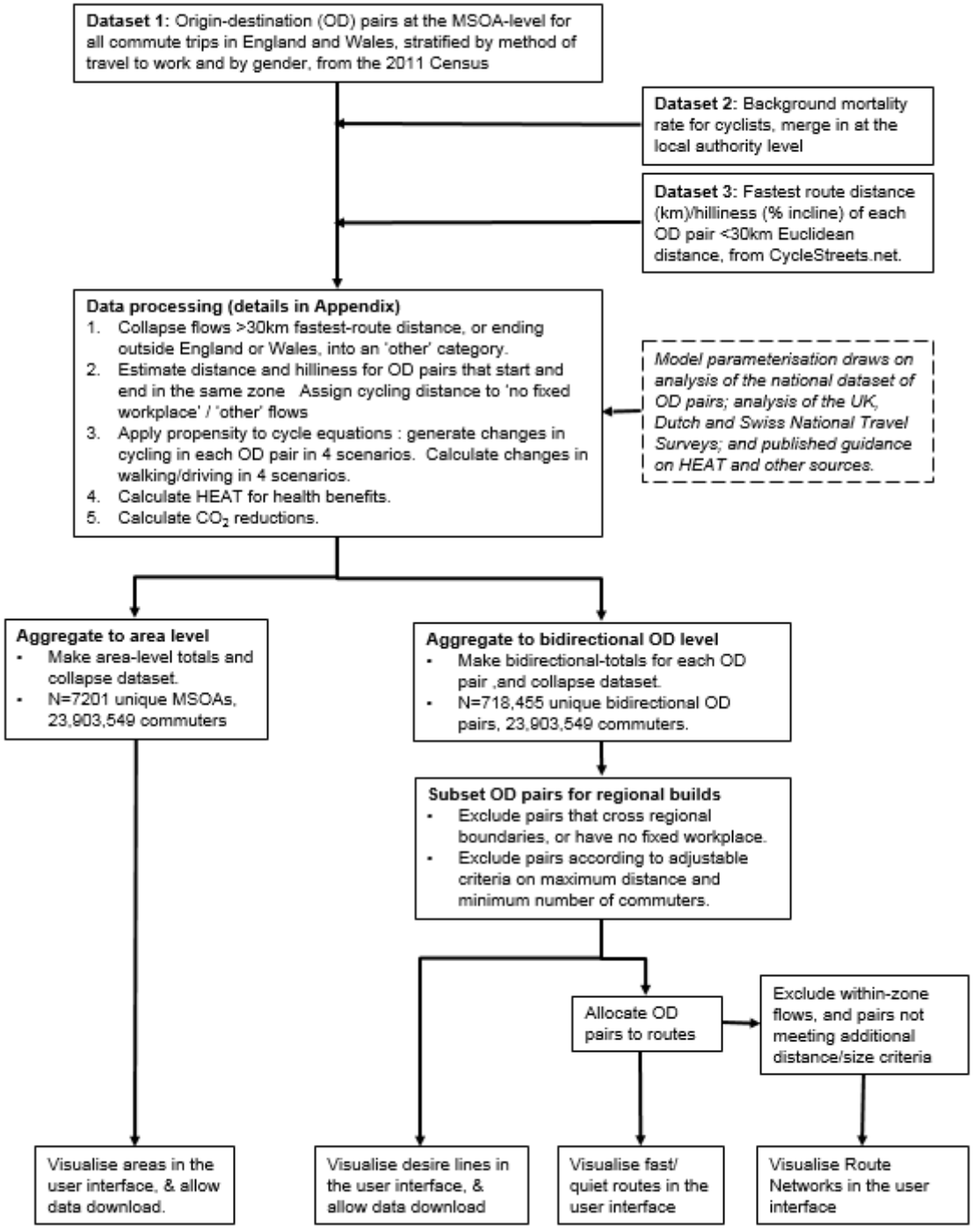
1. Hypotheses: generate two hypotheses that are falsifiable and 2 hypotheses that are not falsifiable
2. Input data: draw schematic representations of additional datasets that you could use alongside the desire lines dataset, with at least one at each of these levels:

- Zones

- Points
- Routes
- Route networks
- Individual

What temporal and spatial resolution could each one have?

3. Methods: using a flow diagram (e.g. as shown below)




# Practical, group computer task (30 minutes)

Create a github account (all). See: https://github.com (https://github.com)

Building on the follow code chunk (but with no copy-and-pasting), create a data frame that contains the names, coffee habits and like/dislike of bus travel for everyone in your group (just 1 computer per group):

```
person_name = c(
  "robin",
  "malcolm",
  "richard"
)
n_coffee = c(
  5,
  1,
  0
)
like_bus_travel = c(
  TRUE,
  FALSE,
  TRUE
)
personal_data = data.frame(person_name, n_coffee, like_bus_travel)
personal_data
```

```
##   person_name n_coffee like_bus_travel
## 1       robin        5            TRUE
## 2     malcolm        1           FALSE
## 3     richard        0            TRUE
```

When you are complete, add your code to https://github.com/ITSLeeds/TDS/blob/master/code-r/01-person-data.R (https://github.com/ITSLeeds/TDS/blob/master/code-r/01-person-data.R)

# Learning outcomes

```
# Identify available datasets and access and clean them
# Combine datasets from multiple sources
# Understand what machine learning is, which problems it is appropriate for compared wit
h traditional statistical approaches, and how to implement machine learning techniques
# Visualise and communicate the results of transport data science, and know about settin
g-up interactive web applications
# Deciding when to use local computing power vs cloud services
```

- Articulate the relevance and limitations of data-centric analysis applied to transport problems, compared with other methods