

# Reproducible Methods in Urban Data Science – Ideas and Examples

Chris Brunsdon

National Centre for Geocomputation  
Maynooth University

contact: [christopher.brunsdon@nuim.ie](mailto:christopher.brunsdon@nuim.ie)



Amsterdam 4-12-17 (1st European Seminar on Urban Data  
Science)

# What is Reproducible Research?

## The Holy Grail

Full details of any results reported and the methods and data used to obtain them should be made available, so that others following the same methods can obtain identical results.

- Recently considered in terms of
  - Statistics
  - Econometrics
  - Signal Processing
  - Epidemiology
  - Data Science More Generally?

# Some more background. . .

Article in Nature, 2010:

## Publish your computer code: it is good enough

Programs written by scientists may be small scripts to draw charts and calculate correlations, trends and significance, larger routines to process and filter data in more complex ways... What they have in common is that, after a paper's publication, they often languish in an obscure folder or are simply deleted. Although the paper may include a brief mathematical description of the processing algorithm, it is rare for science software to be published or even reliably preserved.

Nick Barnes, Nature (467), pp753 (2010)

# Why it Matters

- Not just an Academic Issue
  - Open Data / Open Government
  - Accountability - How did you reach your conclusions or recommendations?



The screenshot shows a web browser window displaying a news article from The Guardian. At the top, there is a navigation bar with links for News, Sport, Comment, Culture, Business, Money, and Life & style. Below this, a green banner highlights the 'Environment' section, with a sub-header 'Hacked climate science emails'. The main headline reads: 'Climategate' review clears scientists of dishonesty over data. The sub-headline states: 'Rigour and honesty' of scientists not in doubt but Sir Muir Russell says UEA's Climatic Research Unit was not sufficiently open. Below the sub-headline, there are two bullet points: 'Follow the latest developments on our Climategate live blog' and 'Read the full text of the review here'. At the bottom of the article, it says 'David Adam, environment correspondent' and 'theguardian.com, Wednesday 7 July 2010 13.02 BST'. There is also a red speech bubble icon with the text 'Jump to comments (328)'.

News | Sport | Comment | Culture | Business | Money | Life & style

Environment > Hacked climate science emails

## 'Climategate' review clears scientists of dishonesty over data

'Rigour and honesty' of scientists not in doubt but Sir Muir Russell says UEA's Climatic Research Unit was not sufficiently open

- Follow the latest developments on our Climategate live blog
- Read the full text of the review here

---

David Adam, environment correspondent  
theguardian.com, Wednesday 7 July 2010 13.02 BST

 Jump to comments (328)



... and more!

### US National Academy of Sciences:

...the default assumption should be that research data, methods (including the techniques, procedures and tools that have been used to collect, generate or analyze data, such as models, computer code and input data) and other information integral to a publically reported result will be publically accessible when results are reported...


Ensuring the Integrity, Accessibility and Stewardship of Research Data in the Digital Age (cited in the Russell report on the CRU)

- Rogoff and Reinhart (the 'Excel Economists')

The Opinion Pages | OP-ED COLUMNIST

The Excel Depression

APRIL 18, 2013



Paul Krugman

Email

Share

Tweet

Save

More

ONLY

In this age of information, math errors can lead to disaster. NASA's [Mars Orbiter crashed](#) because engineers forgot to convert to metric measurements; JPMorgan Chase's "[London Whale](#)" [venture went bad](#) in part because modelers divided by a sum instead of an average. So, did an Excel coding error destroy the economies of the Western world?

The story so far: At the beginning of 2010, two Harvard economists, Carmen Reinhart and Kenneth Rogoff, circulated a paper, "[Growth in a Time of Debt](#)," that purported to identify a critical "threshold," a tipping point, for government indebtedness. Once debt exceeds 90 percent of gross domestic product, they claimed, economic growth drops off sharply.

Ms. Reinhart and Mr. Rogoff had credibility thanks to a widely admired earlier book on the history of financial crises, and their timing was impeccable. The paper came out just after Greece went into crisis and played right into the desire of many officials to "pivot" from stimulus to austerity. As a result, the paper instantly became famous; it was, and is, surely the most influential economic analysis of recent years.

# A few recent issues (possibly interlinked)

- Open data, Open Source
  - Open Analytics?
- Big Data
  - Complicated data
  - Real Time
  - Dashboards for urban data

# So, why bother?

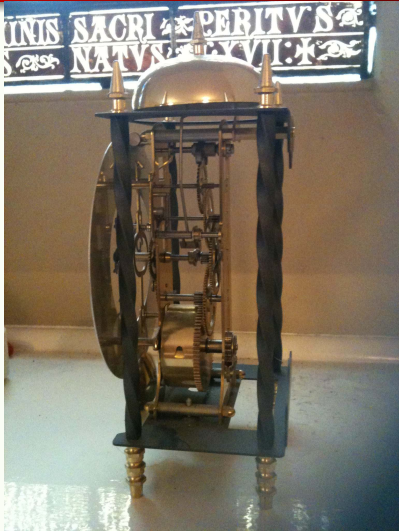
Some scenarios:

- You see a dashboard showing analysis of some transport data - the analysis technique is outlined briefly, but no explicit algorithm is given. Although you have access to the data they used, you are unable to reproduce the analysis.
- A third party finds your analysis helpful - possibly an example of best practice - and want to apply it to local data.
- You wish to access the same data as another site, but want to modify the analysis in some way.

# Some barriers to addressing these problems

- The data used in the original study is available for a fee, and you do not already own it.
- The steps used in the computation are explicitly stated, but require software that is not free, and you do not already own it.
- The data used in the original study is freely available, but the original study does not state the source precisely, or provide a copy.
- The steps used in the computation are not explicitly stated.
- The steps used in the computation are explicitly stated, but the software required is not open source, so that certain details of procedures carried out are not available.
- By adopting the *Reproducible Research* paradigm these barriers can be overcome...

# Reproducibility vs. Free Consultancy – A physical analogy



- Although *free* is helpful, reproducibility is more about *open source*.

# Practical Issues: Why some research ends up losing reproducibility

- Document and computation get separated!
- Particularly with GUI-based software and cut-and-paste.
- Pasting a picture (map, table) into a WP document severs computation from documentation.
- Ideal scenario  $\Rightarrow$  integration of:
  - Data access
  - Code / Script
  - Documentation (possibly on a web site)



# Rmarkdown - a tool for reproducibility

- Uses 'markdown' - a simple markup language
- Simpler than  $\text{\LaTeX}$  or HTML

```
. ### Structure of the talk
.   - Ideas of Reproducible Research
.   - Tools for Reproducible Research
.   - Practical Applications
.   - Loose ends
.   ```{r sample_code, echo=FALSE}
. x <- runif(1000)
. plot(x)
.   ```
```



# Demonstrating Data Lineage - How data was created for an analysis?

```
require(rgdal)
require(maptools)

raw.source <- readLines('ftp://ftp.ncdc.noaa.gov/pub/data/paleo/phenology/north_america_lilac.txt')
inp <- textConnection(gsub('^','',gsub(' +','',raw.source[162:15233])))
leaf.bloom <- read.table(inp,sep=',')
close(inp)
colnames(leaf.bloom) <- c("ID","Year","Type","First.Leaf","First.Bloom")
inp <- textConnection(gsub('^','',gsub(' +','',raw.source[15249:16375])))
station.locs <- read.csv(inp)
close(inp)

phen <- cbind(leaf.bloom,station.locs[match(leaf.bloom$ID,station.locs$ID),-1])
phen$First.Leaf[phen$First.Leaf == 999] <- NA
phen$First.Bloom[phen$First.Bloom == 999] <- NA

p4s <- CRS("+proj=longlat")
phen <- SpatialPointsDataFrame(phen[,9:8],phen,proj4string=p4s)

save(phen,file="phen.RData")
```

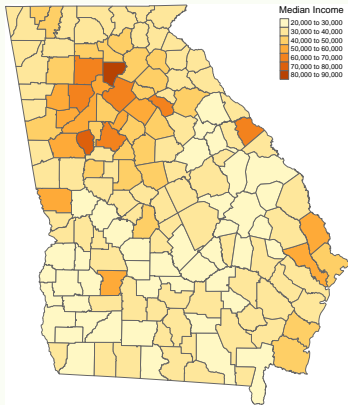
# Or Real Time - Where is it streamed from?

```
if (! is.null(x)) {  
  GET('https://data.dublinked.ie/cgi-bin/rtpi/realtimebusinformation',  
      query=list(stopid=x)) %>%  
    content -> buses_arriving  
}
```

- Here data is from a real time API
- Dublinked <http://dublinked.ie>

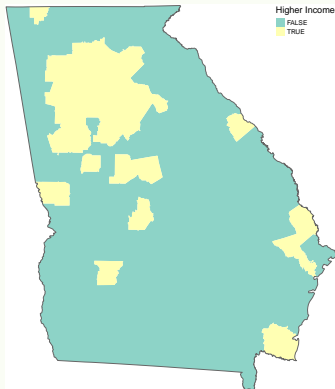
# Using the tmap package for geographical data

```
library(tmap)
data(georgia)
tm_shape(georgia) +
  tm_polygons(col='MedInc',title="Median Income") +
  tm_layout(frame=FALSE)
```



# GIS type operations - via sf and rmapshaper

```
georgia %>% mutate(hi_inc=MedInc > 45000) %>%  
  ms_dissolve(field='hi_inc') -> georgiam  
tm_shape(georgiam) +  
  tm_fill(col='hi_inc',title='Higher Income') +  
  tm_layout(frame=FALSE) + tm_shape(ms_dissolve(georgia)) +  
  tm_borders()
```



- Several possibilities
  - GWR
  - Spatial Regression
  - Microsimulation
  - Local labour market areas
  - Point pattern analysis
- To name a small number

# An Open and Reproducible Geodemographic Classification For The Republic of Ireland

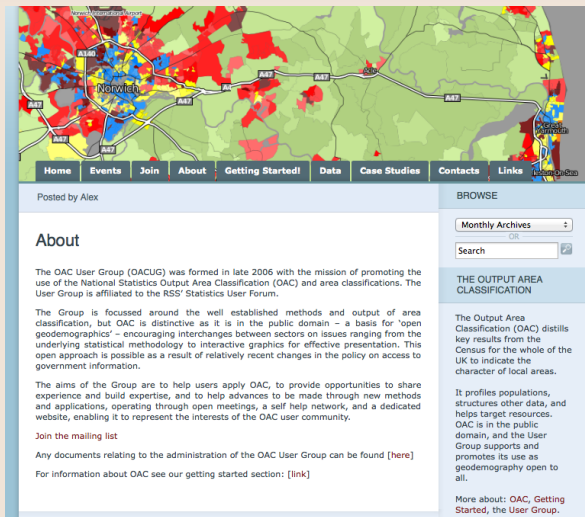
## Background and Motivation:

A geodemographic classification is essentially a grouping of geographical neighbourhoods, or other small areas, in terms of their social and economic characteristics. The classification is generally achieved by applying a *clustering algorithm* such as *k-means* to a data set of social and demographic variables computed for each of the areas.

- Initially used for marketing
- More recently used for social applications
  - eg. targeting health initiatives
  - Profiling university recruitment

# Free Geodemographics and OAC - 1

2001 Census Output Area Classification (OAC) system produced by Vickers, Rees, and Birkin



# Free Geodemographics and OAC - 2

## 2011 Census Output Area Classification (OAC) follow up from UCL

2011 Census Output Area Classification | SASPAC

saspac.org/2013/09/25/2011-census-output-area-classification/ Reader

Bonjour NUI Maynoo Resources Itinerary: le-upon-Tyne RStudio Sign In

WordPress SASPAC Follow Reblog

# SASPAC

making sense of the census

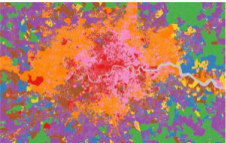
Blog About Software Training Data Support Contact FAQs

## 2011 Census Output Area Classification

A new 2011 UK Output Area Classification (OAC) is currently being produced through a collaborative project between the Office for National Statistics and University College London. Due to the timing of the data releases in Northern Ireland and Scotland, a preliminary OAC has been constructed for England & Wales. ONS/UCL are currently seeking users' views on this work via a short questionnaire available from [here](#).

In short, the 2011 OAC:

- o updates the previous 2001 OAC ([more details](#));
- o is freely available;
- o indicates the character of small areas;
- o contains a three-tiered hierarchical classification of 8 Supergroups, 24 Groups and 67 Subgroups (e.g. "Hard-pressed Households");



**WHAT IS SASPAC?**

SASPAC is a software package designed to store, interrogate, analyse and present UK Census datasets and other small area statistics. [More...](#)

**EMAIL SUBSCRIPTION**

Click to follow this blog and receive notifications of new posts by email.

[Sign up!](#)

[Search](#)

**SASPAC TWITTER UPDATES**

- o Check out September eNews letter [bit.ly/1uweMOK](#) 4 days ago
- o New SASPAC OD-data training day added.. taking place at City Hall take a look now! [saspac.org/training/booki...](#) 1 week ago



# Reproducible Geodemographics - One Step beyond

- Information relating to the data and clustering method used is freely available
- Advantages
  - Others able to scrutinise the approach
  - Others able to adapt the methodology
    - Use different clustering method
    - Use different areal units
    - Update with new data
  - Awareness of variables used
    - Avoid 'faux-pas' of using geodemographic classes to predict a variable already used in the classification system

# Variables Used

## Age Structure

- Age 0 to 4
- Age 5 to 14
- Age 25 to 44
- Age 45 to 64
- Age 65 and over

## Internet Access

- Broadband
- Internet

## Nationality

- EU National
- ROW National
- Born outside Ireland

## Wellbeing

- HE Qualification
- Two Cars
- JTW Public Transport
- Home Workers
- LLTI
- Unpaid Carers
- Unemployed
- Economically Inactive Families

## Occupation

- Students
- Agricultural
- Construction
- Manufacturing
- Commerce
- Transport
- Public Sector
- Professional

## Housing

- Rent Public
- Rent Private
- Flats
- No Central Heating
- Rooms per HH
- People per Room
- Septic Tank

## Household Structure

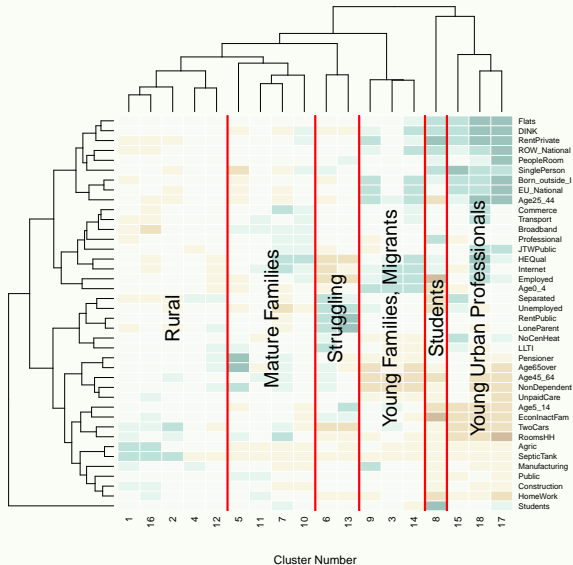
- Separated
- Single Person
- Pensioner
- Lone Parent
- Double Income no Children (DINK)
- Non Dependent Children

## In Reproducibility Terms

- Details (incl. code) - <http://rpubs.com/chrisbrunsdon/14998>
- Reproducibility via *knitr* and *rpubs*

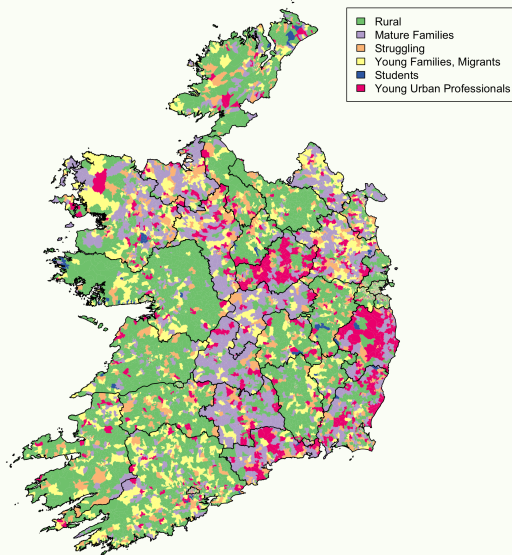


# Labels of Broad Clusters

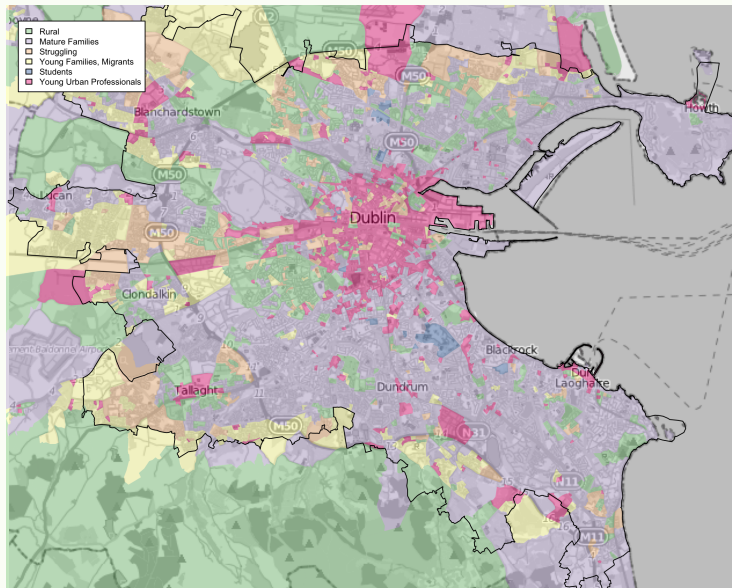


Nb. Brown  $\Rightarrow z < -2.0$ ; Blue  $\Rightarrow z > 2.0$

# Broad Clusters



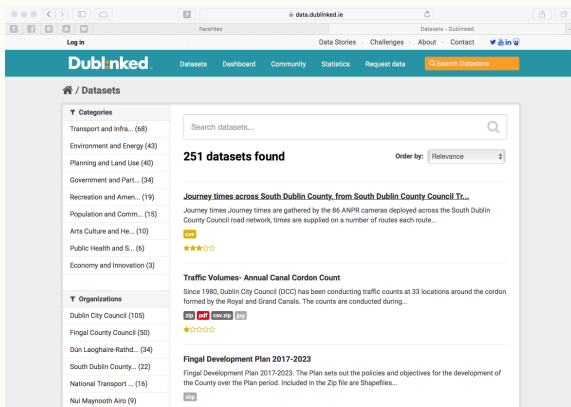
# Dublin Area





- 'Student' Cluster
- Right on the north campus accomodation block!

- Directly related to open data via Dublicked



The screenshot shows the Dublicked website interface. The header includes the Dublicked logo and navigation links: Datasets, Dashboard, Community, Statistics, Request data, and a Search Datasets button. The main content area is titled "/ Datasets" and features a search bar with the text "Search datasets...". Below the search bar, it states "251 datasets found" and "Order by: Relevance". The results list includes:

- Journey times across South Dublin County, from South Dublin County Council Tr...**  
Journey times Journey times are gathered by the 86 ANPR cameras deployed across the South Dublin County Council road network, times are supplied on a number of routes each route...  
CSV  
★★★★☆
- Traffic Volumes- Annual Canal Cordon Count**  
Since 1980, Dublin City Council (DCC) has been conducting traffic counts at 33 locations around the cordon formed by the Royal and Grand Canals. The counts are conducted during...  
ZIP PDF CSV XLS  
★★★★☆
- Fingal Development Plan 2017-2023**  
Fingal Development Plan 2017-2023. The Plan sets out the policies and objectives for the development of the County over the Plan period. Included in the Zip file are Shapfiles...  
ZIP  
★★★★☆

On the left side, there is a sidebar with "Categories" and "Organizations". Categories include Transport and Infra... (68), Environment and Energy (43), Planning and Land Use (40), Government and Part... (34), Recreation and Amen... (19), Population and Comm... (15), Arts Culture and He... (10), Public Health and S... (6), and Economy and Innovation (3). Organizations include Dublin City Council (105), Fingal County Council (50), Dún Laoghaire-Rathd... (34), South Dublin County... (22), National Transport ... (16), and Nui Maynooth Airo (9).

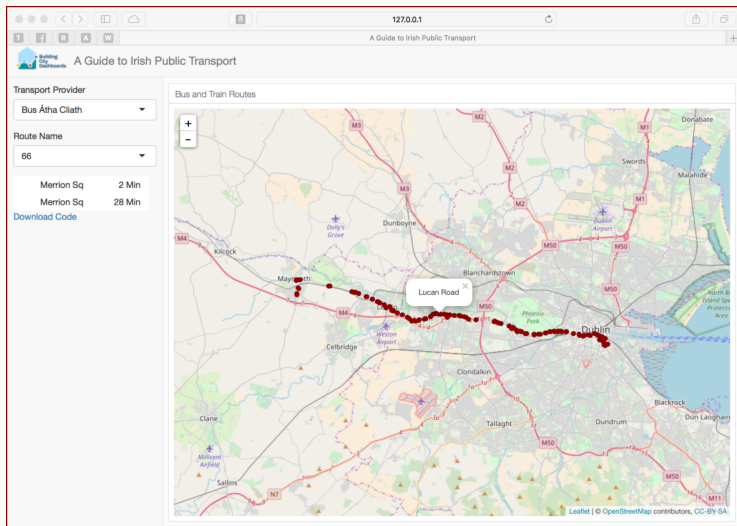


# Using shiny and flexdashboard

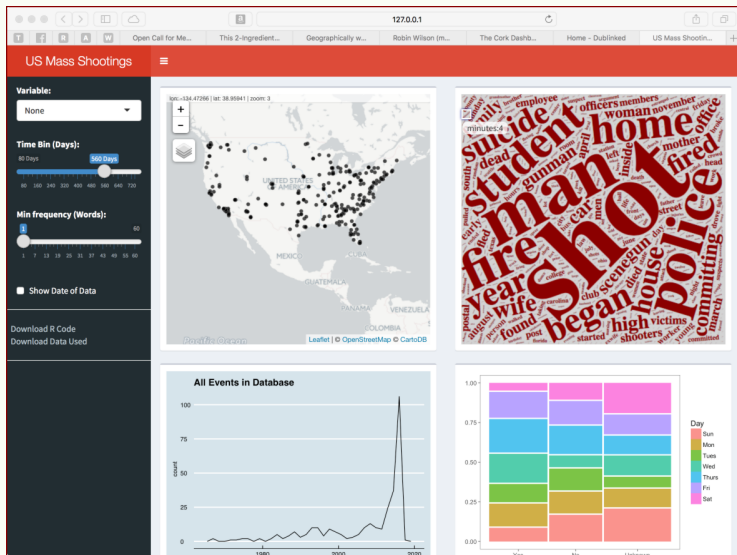
- shiny writes interactive web pages in R
- flexdashboard embeds shiny into Rmarkdown

```
. Column {data-width=650}
. -----
.
.
. ### Bus and Train Routes
.
. ```{r, eval=FALSE}
. observe({
.   x <- input$tp_name
.   routes <- sort(unique(stop_routes$route[stop_routes$name==x]))
.   routes <- c("All",routes)
.   this() %>% updateSelectInput("route",label="Route Name",
.     choices=routes)
. })
.
. # More stuff
. ```
```

# The App...



# Another example



# Opening Data Science

The screenshot shows a web browser window with the URL `eventbrite.ie`. The page features a navigation bar with the Eventbrite logo, a search bar, and links for browsing events, help, sign in, and creating an event. The main content area is for the 'Megadojo 2017 - Maynooth' event, organized by Coderdojo Limerick. The event banner includes logos for Maynooth University, CoderDojo, and Bank of Ireland. The event details are as follows:

Event Title	Organizer	Date	Time	Location	Price	Action
Megadojo 2017 - Maynooth	by Coderdojo Limerick	DEC 02	10am to 5pm (10.30am, 12pm, 2pm & 3.30pm workshops*)	Maynooth University	Free	<a href="#">REGISTER</a>

**DESCRIPTION**

We're inviting young people to join us for a day of technology-based fun in Maynooth University. We're looking for 1024 kids to take part. We'll have coding workshops, demos from companies (big and small), talks and a hackathon. And all this FREE!

We will be running beginner and intermediate Scratch workshops for the younger attendees along with WebDev and hardware workshops for the older age groups. These are aimed

**DATE AND TIME**

Sat, December 2, 2017  
10:00 AM – 5:00 PM GMT  
[Add to Calendar](#)

**LOCATION**

Maynooth University

The contribution of Science Foundation Ireland  
(Investigators Programme Grant 15/IA/3090 - Building City Dashboards)  
is gratefully acknowledged.