Adding value to the Scottish Neighbourhoods Statistics bulk data download using R and the reproducible research, tidy data, and split-apply-combine paradigms

Dr Jon Minton,
AQMEN Research Fellow
University of Glasgow

Structure

- The Scottish Neighbourhood Statistics data collection
- Usability for casual users and power users
- The Bulk data download
- Data management paradigms:
  - Paradigm 1: tidy data
  - Paradigm 2: reproducible research
  - Paradigm 3: split-apply-combine
- Challenges and processes
- Example outputs
- Summary

Scottish Neighbourhood Statistics

- http://www.sns.gov.uk/

"The Scottish Government's on-going programme to improve the availability, consistency and accessibility of small area statistics in Scotland."

Small area focus: Uses 2001 datazones (agglomorations of 2001 census output areas) for many different years, each containing between around 500 and 1000 people

Domains include:

- Health
- Crime
- Labour Markets
- Education
- Etc

Data from many sources, including but not just the census.

http://www.sns.gov.uk/Guide/SnsInfo.aspx?Page=About

SNS for casual/occasional users

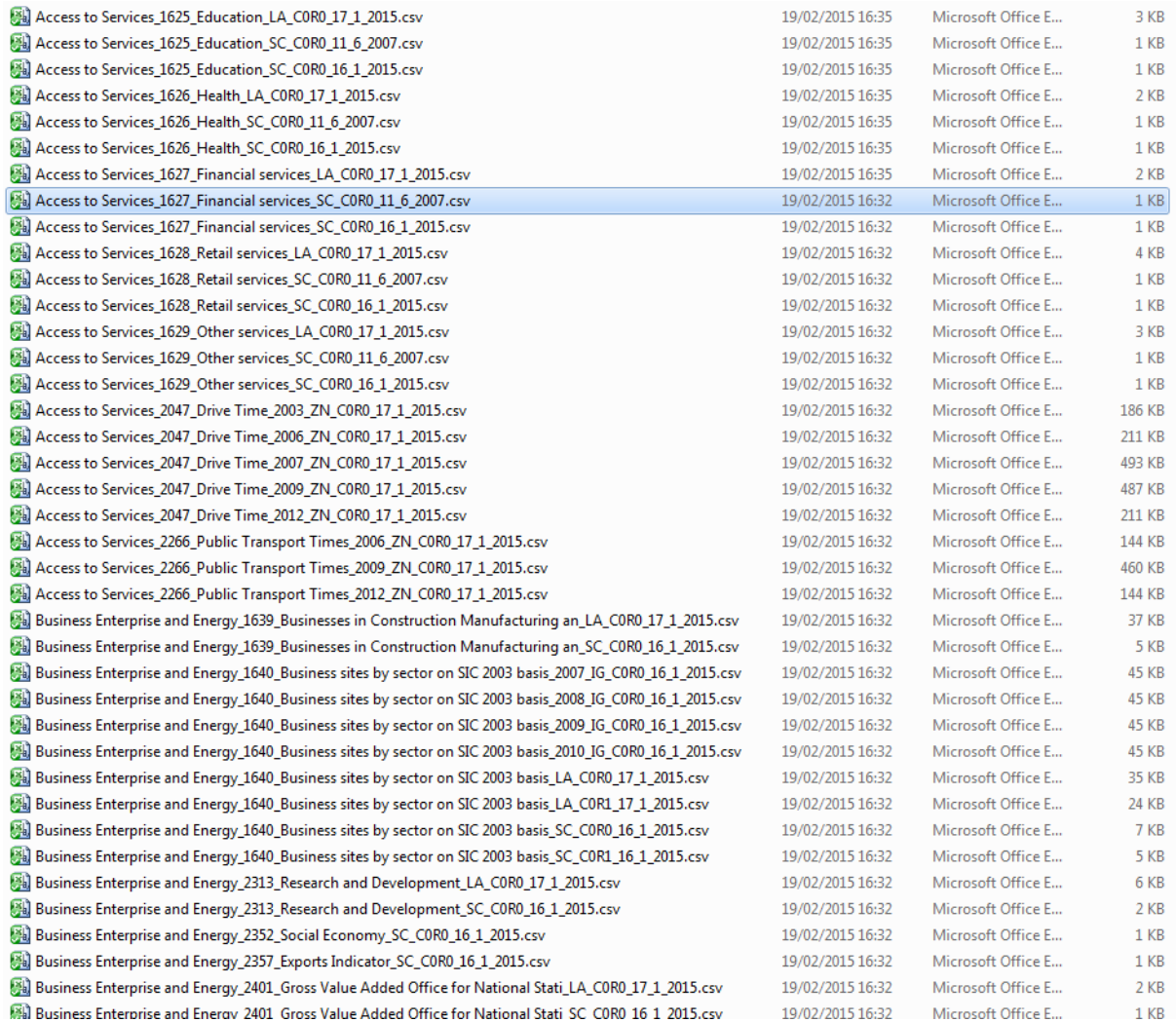
- Well designed:
  - Easy to use user interface,
  - drop down menus,
  - choropleths,
  - automatically generated reports.
- http://www.sns.gov.uk/default.aspx

SNS and power users

- Power users:
  - Command line based
  - Automated
  - Reproducible
  - Data stored and organised so that only limited pre-processing required.

- SNS bulk data option
  - All the tables, but in what format?
  - http://www.sns.gov.uk/Downloads/DownloadHome.aspx

Contents of download



Access to Services_1625_Education_LA_C0R0_17_1_2015.csv	19/02/2015 16:35	Microsoft Office E...	3 KB
Access to Services_1625_Education_SC_C0R0_11_6_2007.csv	19/02/2015 16:35	Microsoft Office E...	1 KB
Access to Services_1625_Education_SC_C0R0_16_1_2015.csv	19/02/2015 16:35	Microsoft Office E...	1 KB
Access to Services_1626_Health_LA_C0R0_17_1_2015.csv	19/02/2015 16:35	Microsoft Office E...	2 KB
Access to Services_1626_Health_SC_C0R0_11_6_2007.csv	19/02/2015 16:35	Microsoft Office E...	1 KB
Access to Services_1626_Health_SC_C0R0_16_1_2015.csv	19/02/2015 16:35	Microsoft Office E...	1 KB
Access to Services_1627_Financial services_LA_C0R0_17_1_2015.csv	19/02/2015 16:35	Microsoft Office E...	2 KB
Access to Services_1627_Financial services_SC_C0R0_11_6_2007.csv	19/02/2015 16:32	Microsoft Office E...	1 KB
Access to Services_1627_Financial services_SC_C0R0_16_1_2015.csv	19/02/2015 16:32	Microsoft Office E...	1 KB
Access to Services_1628_Retail services_LA_C0R0_17_1_2015.csv	19/02/2015 16:32	Microsoft Office E...	4 KB
Access to Services_1628_Retail services_SC_C0R0_11_6_2007.csv	19/02/2015 16:32	Microsoft Office E...	1 KB
Access to Services_1628_Retail services_SC_C0R0_16_1_2015.csv	19/02/2015 16:32	Microsoft Office E...	1 KB
Access to Services_1629_Other services_LA_C0R0_17_1_2015.csv	19/02/2015 16:32	Microsoft Office E...	3 KB
Access to Services_1629_Other services_SC_C0R0_11_6_2007.csv	19/02/2015 16:32	Microsoft Office E...	1 KB
Access to Services_1629_Other services_SC_C0R0_16_1_2015.csv	19/02/2015 16:32	Microsoft Office E...	1 KB
Access to Services_2047_Drive Time_2003_ZN_C0R0_17_1_2015.csv	19/02/2015 16:32	Microsoft Office E...	186 KB
Access to Services_2047_Drive Time_2006_ZN_C0R0_17_1_2015.csv	19/02/2015 16:32	Microsoft Office E...	211 KB
Access to Services_2047_Drive Time_2007_ZN_C0R0_17_1_2015.csv	19/02/2015 16:32	Microsoft Office E...	493 KB
Access to Services_2047_Drive Time_2009_ZN_C0R0_17_1_2015.csv	19/02/2015 16:32	Microsoft Office E...	487 KB
Access to Services_2047_Drive Time_2012_ZN_C0R0_17_1_2015.csv	19/02/2015 16:32	Microsoft Office E...	211 KB
Access to Services_2266_Public Transport Times_2006_ZN_C0R0_17_1_2015.csv	19/02/2015 16:32	Microsoft Office E...	144 KB
Access to Services_2266_Public Transport Times_2009_ZN_C0R0_17_1_2015.csv	19/02/2015 16:32	Microsoft Office E...	460 KB
Access to Services_2266_Public Transport Times_2012_ZN_C0R0_17_1_2015.csv	19/02/2015 16:32	Microsoft Office E...	144 KB
Business Enterprise and Energy_1639_Businesses in Construction Manufacturing an_LA_C0R0_17_1_2015.csv	19/02/2015 16:32	Microsoft Office E...	37 KB
Business Enterprise and Energy_1639_Businesses in Construction Manufacturing an_SC_C0R0_16_1_2015.csv	19/02/2015 16:32	Microsoft Office E...	5 KB
Business Enterprise and Energy_1640_Business sites by sector on SIC 2003 basis_2007_IG_C0R0_16_1_2015.csv	19/02/2015 16:32	Microsoft Office E...	45 KB
Business Enterprise and Energy_1640_Business sites by sector on SIC 2003 basis_2008_IG_C0R0_16_1_2015.csv	19/02/2015 16:32	Microsoft Office E...	45 KB
Business Enterprise and Energy_1640_Business sites by sector on SIC 2003 basis_2009_IG_C0R0_16_1_2015.csv	19/02/2015 16:32	Microsoft Office E...	45 KB
Business Enterprise and Energy_1640_Business sites by sector on SIC 2003 basis_2010_IG_C0R0_16_1_2015.csv	19/02/2015 16:32	Microsoft Office E...	45 KB
Business Enterprise and Energy_1640_Business sites by sector on SIC 2003 basis_LA_C0R0_17_1_2015.csv	19/02/2015 16:32	Microsoft Office E...	35 KB
Business Enterprise and Energy_1640_Business sites by sector on SIC 2003 basis_LA_C0R1_17_1_2015.csv	19/02/2015 16:32	Microsoft Office E...	24 KB
Business Enterprise and Energy_1640_Business sites by sector on SIC 2003 basis_SC_C0R0_16_1_2015.csv	19/02/2015 16:32	Microsoft Office E...	7 KB
Business Enterprise and Energy_1640_Business sites by sector on SIC 2003 basis_SC_C0R1_16_1_2015.csv	19/02/2015 16:32	Microsoft Office E...	5 KB
Business Enterprise and Energy_2313_Research and Development_LA_C0R0_17_1_2015.csv	19/02/2015 16:32	Microsoft Office E...	6 KB
Business Enterprise and Energy_2313_Research and Development_SC_C0R0_16_1_2015.csv	19/02/2015 16:32	Microsoft Office E...	2 KB
Business Enterprise and Energy_2352_Social Economy_SC_C0R0_16_1_2015.csv	19/02/2015 16:32	Microsoft Office E...	1 KB
Business Enterprise and Energy_2357_Exports Indicator_SC_C0R0_16_1_2015.csv	19/02/2015 16:32	Microsoft Office E...	1 KB
Business Enterprise and Energy_2401_Gross Value Added Office for National Stati_LA_C0R0_17_1_2015.csv	19/02/2015 16:32	Microsoft Office E...	2 KB
Business Enterprise and Energy_2401_Gross Value Added Office for National Stati_SC_C0R0_16_1_2015.csv	19/02/2015 16:32	Microsoft Office E...	1 KB

Contents of a file

- Access to services_1629_Other services_LA_C0R0_17_1_2015.csv

	CS-pcthhcac0to30	CS-pcthhcac30to60	CS-pcthhcac60plus	CS-pcthhjob0to30	CS-pcthhjob30to60	CS-pcthhjob60plus	C
GeographyCode	2001	2001	2001	2001	2001	2001	
S12000005	100	0	0	100	0	0	
S12000006	56.6	23.4	20	95.8	4.2	0	
S12000008	89.8	10.2	0	100	0	0	
S12000010	99.9	0.1	0	100	0	0	
S12000011	100	0	0	100	0	0	
S12000013	73.5	24	2.5	53.2	16.6	30.2	
S12000014	100	0	0	100	0	0	

‘Tidy data’

1. *Each variable forms a column*
2. *Each observation forms a row*
3. *Each type of observation unit forms a table*

http://vita.had.co.nz/papers/tidy-data.pdf

The way I think about this:

- ‘Where’ variables
- ‘What’ variables

Contents of a file

- Access to services_1629_Other services_LA_C0R0_17_1_2015.csv

	CS-pcthhcac0to30	CS-pcthhcac30to60	CS-pcthhcac60plus	CS-pcthhjob0to30	CS-pcthhjob30to60	CS-pcthhjob60plus	C
GeographyCode	2001	2001	2001	2001	2001	2001	
S12000005	100	0	0	100	0	0	
S12000006	56.6	23.4	20	95.8	4.2	0	
S12000008	89.8	10.2	0	100	0	0	
S12000010	99.9	0.1	0	100	0	0	
S12000011	100	0	0	100	0	0	
S12000013	73.5	24	2.5	53.2	16.6	30.2	
S12000014	100	0	0	100	0	0	

Where and what?

- Where variables
  - **Physical location: datazone code**
  - **Year or other time period**
  - *(Demographic groupings etc)*
- What variables
  - Observations relating to a **specific physical, temporal**, (demographic) ‘location’

Automation and split-apply-combine

- An earlier contribution to the Wickhamverse
  - http://www.jstatsoft.org/v40/i01/paper
  - plyr

  > *Many data analysis problems involve the application of a split-apply-combine strategy, where you break up a big problem into manageable pieces, operate on each piece independently and then put all the pieces back together. This insight gives rise to a new R package that allows you to smoothly apply this strategy, without having to worry about the type of structure in which your data is stored.”*

- Plyr can be used to automate the reading in and writing out of data files, so to automate the conversion of the SNS bulk data download to a ‘tidy data’ format

What can be automated, and what can't be?

- Access to services_1629_Other services_LA_C0R0_17_1_2015.csv

	CS-pcthhcac0to30	CS-pcthhcac30to60	CS-pcthhcac60plus	CS-pcthhjob0to30	CS-pcthhjob30to60	CS-pcthhjob60plus	C
GeographyCode	2001	2001	2001	2001	2001	2001	
S12000005	100	0	0	100	0	0	
S12000006	56.6	23.4	20	95.8	4.2	0	
S12000008	89.8	10.2	0	100	0	0	
S12000010	99.9	0.1	0	100	0	0	
S12000011	100	0	0	100	0	0	
S12000013	73.5	24	2.5	53.2	16.6	30.2	
S12000014	100	0	0	100	0	0	

Quasi-tidy automatable form

- Ideal tidy form:

areal_unit	year	age_group	hcac	Hjob	etc
'Where' variables			*'What' variables*		

- Automatable quasi-tidy form:

areal_ unit	year	Hcacage1	Hcacage2	Hcacage3	Hjobage1	Hjobage2	Hjobage3	etc
'Where' variables		*'What' /'where' variables*						

Reproducible research

- R studio projects and github repositories
  - Code and (small) data repository
- Dropbox and Repmis for reading larger files
- (SQL server interfaces as ideal end stage)

Target form specifications

- Datazones where possible
  - Identifiable by _dz_ in title
- Years and data which can be aggregable to years
  - Quarters, months
  - Stage 1: identify
  - Stage 2: extract
  - Stage 3: aggregate

The code itself

- https://github.com/JonMinton/sns_fetch_and_recast/blob/master/script.r

Output data

- Quasi-tidied data
  - https://www.dropbox.com/sh/ofv701lxrswfjsy/AABj6QWL_eR8I2SYxuwEkA5Ca?dl=0

- Population count data, with additional tidying:
  - https://www.dropbox.com/s/17r8nz8k77w958m/populations_by_age_year_sex.csv?dl=0

Summary & Conclusions

- Always a trade off
  - Time invested in automating
  - Time saved by automating
  - Degree of automation
- The tidy data paradigm makes it easier to think about the target data form
- Plyr makes it easier to break down a task into standardised pieces
- Github and dropbox mean the outputs of this work can benefit others (including my future self) more quickly and easily
- R is always changing and evolving
  - I might not write the code now the way I did then
  - dplyr, maggrittr, and %>%

Thanks for listening!

Jon

Jonathan.minton@glasgow.ac.uk