1 Citation
2 Introduction
3 Obtain listing of files
4 Load data files
5 Data quality analysis
6 Runtime
7 R environment

Code ▾

Processing, cleaning and saving NZ GREEN Grid project 1 minute electricity consumption data

Ben Anderson (b.anderson@soton.ac.uk (mailto:b.anderson@soton.ac.uk), `@dataknut`)

Last run at: 2018-05-03 17:04:15

1 Citation

If you wish to use any of the material from this report please cite as:

- Anderson, B. (2018) Processing, cleaning and saving NZ GREEN Grid project 1 minute electricity consumption data, University of Otago: Dunedin, NZ.

2 Introduction

Report circulation:

- Restricted to: NZ GREEn Grid (https://www.otago.ac.nz/centre-sustainability/research/energy/otago050285.html) project partners and contractors.

2.1 Purpose

This report is intended to:

- load and clean the project electricity consumption data (Grid Spy)
- save the cleaned data out as a single file per household
- produce summary data quality statistics

2.2 Requirements:

- grid spy 1 minute data downloads

1 Citation
2 Introduction
3 Obtain listing of files
4 Load data files
5 Data quality analysis
6 Runtime
7 R environment

2.3 History

Generally tracked via our git.soton repo (https://git.soton.ac.uk/ba1e12/nzGREENGrid).

2.4 Support

This work was supported by:

- The University of Otago (https://www.otago.ac.nz/)
- The New Zealand Ministry of Business, Innovation and Employment (MBIE) (http://www.mbie.govt.nz/)
- SPATIALEC (http://www.energy.soton.ac.uk/tag/spatialec/) - a Marie Skłodowska-Curie Global Fellowship (http://ec.europa.eu/research/mariecurieactions/about-msca/actions/if/index_en.htm) based at the University of Otago’s Centre for Sustainability (http://www.otago.ac.nz/centre-sustainability/staff/otago673896.html) (2017-2019) & the University of Southampton’s Sustainable Energy Research Group (2019-202).

3 Obtain listing of files

In this section we generate a listing of all 1 minute data files that we have received. If we are running over the complete dataset then we will be using data from:

- /hum-csafe/Research Projects/GREEN Grid/_RAW DATA/GridSpyData/

In this run we are using data from:

- /Volumes/hum-csafe/Research Projects/GREEN Grid/_RAW DATA/GridSpyData/

If these do not match then this may be a test run.

Code

```
## [1] "Looking for 1 minute data using pa
ttern = *at1.csv$ in /Volumes/hum-csafe/Re
search Projects/GREEN Grid/_RAW DATA/GridS
pyData/ - could take a while..."
```

1 Citation

2 Introduction

3 Obtain listing of files

4 Load data files

5 Data quality analysis

6 Runtime

7 R environment

Code

```
##    user  system elapsed
##   0.751   5.491 352.092
```

Code

```
## [1] "Found 21,176 files"
```

Code

```
## [1] "Processing file list and getting f
ile meta-data (please be patient)"
## [1] "All files checked"
## [1] "Saving 1 minute data files metadat
a to /Volumes/hum-csafe/Research Projects/
GREEN Grid/Clean_data/gridSpy/fListComplet
eDT.csv"
## [1] "Done"
## [1] "Saving final 1 minute data files m
etadata to /Volumes/hum-csafe/Research Pro
jects/GREEN Grid/Clean_data/gridSpy/fListC
ompleteDT.csv"
## [1] "Done"
```

Code

```
## [1] "Overall we have 21176 files from 4
4 households."
```

Code

Overall we have 21,176 files from 44 households. Of the 21,176, 12,306 (58.11%) were *not* loaded/checked as their file sizes indicated that they contained no data.

We now need to check how many of the loaded files have an ambiguous or default date - these could introduce errors.

Code

Number of files with given date column names by inferred date format

dateColName	dateFormat	nFiles
date NZ	dmy - definite	1
date NZ	mdy - definite	2
date NZ	ymd - default (but day/month value <= 12)	12

1 Citation
2 Introduction
3 Obtain listing of files
4 Load data files
5 Data quality analysis
6 Runtime
7 R environment

dateColName	dateFormat	nFiles
date NZ	ymd - definite	67
date UTC	ambiguous	28
date UTC	ymd - default (but day/month value <= 12)	3413
date UTC	ymd - definite	5347
unknown - file not loaded (fsize = 2751)	NA	1812
unknown - file not loaded (fsize = 43)	NA	10494

Results to note:

- There are 28 ambiguous files
- The non-loaded files only have 2 distinct file sizes, confirming that they are unlikely to contain useful data.

We now inspect the ambiguous and (some of) the default files.

To help with data cleaning the following table lists files that are ambiguous.

Code

Files with ambigious date formats

file	dateColName	dateExample	dateFormat
rf_06/15Jul2014-25May2016at1.csv	date UTC	14/07/14	ambiguous
rf_07/15Jul2014-25May2016at1.csv	date UTC	14/07/14	ambiguous
rf_08/15Jul2014-25May2016at1.csv	date UTC	14/07/14	ambiguous
rf_10/15Jul2014-25May2016at1.csv	date UTC	14/07/14	ambiguous
rf_11/15Jul2014-25May2016at1.csv	date UTC	14/07/14	ambiguous
rf_13/15Jul2014-25May2016at1.csv	date UTC	14/07/14	ambiguous
rf_19/15Jul2014-25May2016at1.csv	date UTC	14/07/14	ambiguous
rf_21/15Jul2014-25May2016at1.csv	date UTC	14/07/14	ambiguous
rf_22/15Jul2014-25May2016at1.csv	date UTC	14/07/14	ambiguous
rf_23/15Jul2014-25May2016at1.csv	date UTC	14/07/14	ambiguous
rf_24/15Jul2014-25May2016at1.csv	date UTC	27/07/14	ambiguous

1 Citation
2 Introduction
3 Obtain listing of files
4 Load data files
5 Data quality analysis
6 Runtime
7 R environment

file	dateColName	dateExample	dateFormat
rf_25/12Oct2016-20Nov2017at1.csv	date UTC	11-10-16	ambiguous
rf_26/15Jul2014-25May2016at1.csv	date UTC	14/07/14	ambiguous
rf_27/15Jul2014-25May2016at1.csv	date UTC	27/07/14	ambiguous
rf_29/24Mar2015-25May2016at1.csv	date UTC	25/03/15	ambiguous
rf_30/15Feb2016-25May2016at1.csv	date UTC	14/02/16	ambiguous
rf_30/24Mar2015-25May2016at1.csv	date UTC	27/03/15	ambiguous
rf_31/24Mar2015-25May2016at1.csv	date UTC	25/03/15	ambiguous
rf_34/18Jan2016-25May2016at1.csv	date UTC	17/01/16	ambiguous
rf_34/20Jul2015-25May2016at1.csv	date UTC	19/07/15	ambiguous
rf_34/24Mar2015-25May2016at1.csv	date UTC	26/03/15	ambiguous
rf_35/24Mar2015-25May2016at1.csv	date UTC	23/03/15	ambiguous
rf_39/24Mar2015-25May2016at1.csv	date UTC	27/03/15	ambiguous
rf_43/24Mar2015-25May2016at1.csv	date UTC	26/03/15	ambiguous
rf_43/27Mar2015-18Oct2015at1.csv	date UTC	26/03/15	ambiguous
rf_44/24Mar2015-25May2016at1.csv	date UTC	24/03/15	ambiguous
rf_46/12Oct2016-20Nov2017at1.csv	date UTC	11-10-16	ambiguous
rf_47/24Mar2015-25May2016at1.csv	date UTC	24/03/15	ambiguous

Looking at the file names we will assume they are dmy.

Code

The following table lists 'date NZ' files which are set by default only - do they look OK to assume dateFormat?

Code

Files with inferred default date formats

file	fSize	dateColName	dateExample	dateFormat
rf_01/1Jan2014-24May2014at1.csv	6255737	date NZ	2014-01-06	ymd - default (but day/month value <= 12)
rf_02/1Jan2014-24May2014at1.csv	6131625	date NZ	2014-03-03	ymd - default (but day/month value <= 12)

1 Citation
2 Introduction
3 Obtain listing of files
4 Load data files
5 Data quality analysis
6 Runtime
7 R environment

file	fSize	dateColName	dateExample	dateFormat
rf_06/24May2014-24May2015at1.csv	19398444	date NZ	2014-06-09	ymd - default (but day/month value <= 12)
rf_10/24May2014-24May2015at1.csv	24386048	date NZ	2014-07-09	ymd - default (but day/month value <= 12)
rf_11/24May2014-24May2015at1.csv	23693893	date NZ	2014-07-08	ymd - default (but day/month value <= 12)
rf_12/24May2014-24May2015at1.csv	21191785	date NZ	2014-07-09	ymd - default (but day/month value <= 12)

These look OK if we compare the file names with the dateExample.

The following table lists 'date NZ' files which are set by default only - do they look OK to assume dateFormat?

Code

Files with inferred default date formats

file	fSize	dateColName	dateExample	dateFormat
rf_06/10Apr2018-11Apr2018at1.csv	156944	date UTC	2018-04-09	ymd - default (but day/month value <= 12)
rf_06/10Dec2017-11Dec2017at1.csv	156601	date UTC	2017-12-09	ymd - default (but day/month value <= 12)
rf_06/10Feb2018-11Feb2018at1.csv	153353	date UTC	2018-02-09	ymd - default (but day/month value <= 12)
rf_06/10Jan2018-11Jan2018at1.csv	153982	date UTC	2018-01-09	ymd - default (but day/month value <= 12)

1 Citation
2 Introduction
3 Obtain listing of files
4 Load data files
5 Data quality analysis
6 Runtime
7 R environment

file	fSize	dateColName	dateExample	dateFormat
rf_06/10Mar2018-11Mar2018at1.csv	156471	date UTC	2018-03-09	ymd - default (but day/month value <= 12)
rf_06/10Nov2017-11Nov2017at1.csv	155639	date UTC	2017-11-09	ymd - default (but day/month value <= 12)

These also look OK so we will stick with the following derived date formats:

Code

Number of files with given date column names by final imputed date format

dateColName	dateFormat	nFiles
date NZ	dmy - definite	1
date NZ	mdy - definite	2
date NZ	ymd - default (but day/month value <= 12)	12
date NZ	ymd - definite	67
date UTC	dmy - inferred	28
date UTC	ymd - default (but day/month value <= 12)	3413
date UTC	ymd - definite	5347
unknown - file not loaded (fsize = 2751)	NA	1812
unknown - file not loaded (fsize = 43)	NA	10494

3.1 Data file quality checks

The following chart shows the distribution of these files over time using their sizes. Note that white indicates the presence of small files which may not contain observations.

Code

1 Citation

2 Introduction

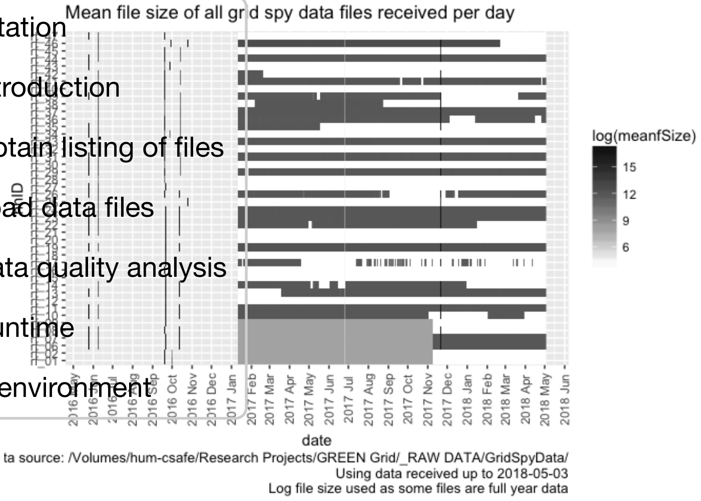
3 Obtain listing of files

4 Load data files

5 Data quality analysis

6 Runtime

7 R environment



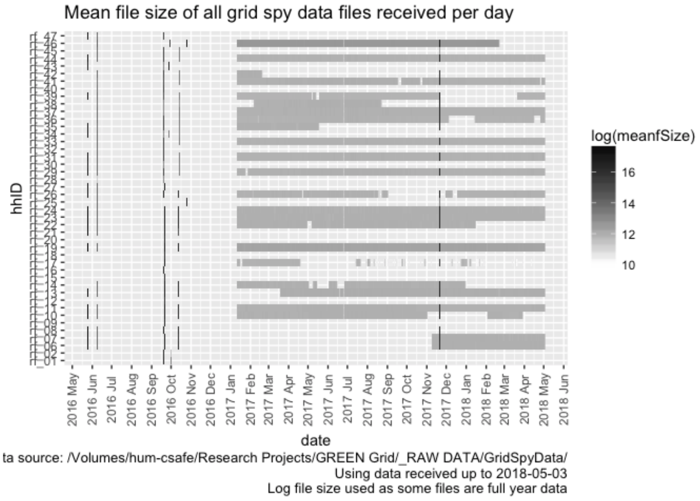
ta source: /Volumes/hum-csafe/Research Projects/GREEN Grid/_RAW DATA/GridSpyData/
Using data received up to 2018-05-03
Log file size used as some files are full year data

Code

```
## Saving 7 x 5 in image
```

The following chart shows the same chart but only for files which we think contain data.

Code



ta source: /Volumes/hum-csafe/Research Projects/GREEN Grid/_RAW DATA/GridSpyData/
Using data received up to 2018-05-03
Log file size used as some files are full year data

Code

```
## Saving 7 x 5 in image
```

4 Load data files

In this section we load the data files that have a file size > 3000 bytes. Things to note:

- We assume that any files smaller than this value have no observations. This is based on:
    - Manual inspection of several small files
    - The identical (small) file sizes involved
    - *But* we should probably test the first few

lines to double check…

- We have to deal with quite a lot of duplication some of which has caused the different date formats. See our repo issues list (https://git.soton.ac.uk/ba1e12/nzGREENGrid/issues?scope=all&utf8=%E2%9C%93&state=all).

1 Citation
2 Introduction
3 Obtain listing of files
4 Load data files
5 Data quality analysis
6 Runtime
7 R environment

The following table shows the number of files per household that we willl load.

Code

Summary of household files to load

hhID	nFiles	meanSize	minFileDate	maxFileDate
rf_01	3	15548174.7	2016-09-20	2016-09-30
rf_02	3	10134268.3	2016-09-20	2016-09-30
rf_06	180	811227.3	2016-05-25	2018-05-02
rf_07	180	872017.9	2016-05-25	2018-05-02
rf_08	5	23989121.0	2016-05-25	2017-11-21
rf_09	2	14344605.0	2016-09-21	2016-09-21
rf_10	358	525455.0	2016-05-25	2018-03-30
rf_11	482	427777.7	2016-05-25	2018-05-02
rf_12	2	10713096.0	2016-09-21	2016-09-21
rf_13	414	495372.3	2016-05-25	2018-05-02
rf_14	329	424262.0	2016-06-08	2017-12-31
rf_15	2	10553143.0	2016-09-21	2016-09-21
rf_16	1	20037376.0	2016-09-20	2016-09-20
rf_17	202	415129.2	2016-09-21	2018-04-12
rf_18	2	14374309.5	2016-09-21	2016-09-21
rf_19	482	567987.6	2016-05-25	2018-05-02
rf_20	2	14665810.0	2016-09-21	2016-09-21
rf_21	4	23058797.8	2016-05-25	2016-10-12
rf_22	371	533704.5	2016-05-25	2018-01-16
rf_23	482	443525.6	2016-05-25	2018-05-02
rf_24	482	431897.8	2016-05-25	2018-05-02
rf_25	3	12341581.3	2016-06-08	2017-11-21

1 Citation
2 Introduction
3 Obtain listing of files
4 Load data files
5 Data quality analysis
6 Runtime
7 R environment

hhID	nFiles	meanSize	minFileDate	maxFileDate
rf_26	388	412369.7	2016-05-25	2018-05-02
rf_27	3	22607698.7	2016-05-25	2016-09-21
rf_28	2	2297483.0	2016-06-08	2016-09-19
rf_29	479	343395.5	2016-05-25	2018-05-02
rf_30	5	13695336.0	2016-05-25	2016-10-13
rf_31	482	342570.0	2016-05-25	2018-05-02
rf_32	2	13934454.0	2016-06-08	2016-09-20
rf_33	481	288981.7	2016-06-08	2018-05-02
rf_34	7	14106275.3	2016-05-25	2016-10-13
rf_35	134	573648.6	2016-05-25	2017-11-21
rf_36	432	301991.4	2016-06-08	2018-05-02
rf_37	481	302924.8	2016-06-08	2018-05-02
rf_38	201	385707.5	2016-06-08	2017-11-21
rf_39	358	385304.5	2016-05-25	2018-05-02
rf_40	2	9299902.0	2016-06-08	2016-09-20
rf_41	473	266272.2	2016-06-08	2018-05-02
rf_42	45	1315953.6	2016-06-08	2017-11-21
rf_43	4	9442492.0	2016-05-25	2016-09-28
rf_44	482	344224.9	2016-05-25	2018-05-02
rf_45	4	10513812.0	2016-06-08	2017-11-21
rf_46	411	605048.1	2016-06-08	2018-02-21
rf_47	3	17544847.0	2016-05-25	2016-09-20

Code

1 Citation

2 Introduction

3 Obtain listing of files

4 Load data files

5 Data quality analysis

6 Runtime

7 R environment

```
## [1] "Loading: rf_01"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_01_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_01_all_1min_data.csv"
## [1] "Loading: rf_02"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_02_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_02_all_1min_data.csv"
## [1] "Loading: rf_06"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_06_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_06_all_1min_data.csv"
## [1] "Loading: rf_07"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_07_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_07_all_1min_data.csv"
## [1] "Loading: rf_08"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_08_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_08_all_1min_data.csv"
## [1] "Loading: rf_09"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_09_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_09_all_1min_data.csv"
## [1] "Loading: rf_10"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_10_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_10_all_1min_data.csv"
## [1] "Loading: rf_11"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
```

1 Citation

2 Introduction

3 Obtain listing of files

4 Load data files

5 Data quality analysis

6 Runtime

7 R environment

```
n/rf_11_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_11_all_1min_data.csv"
## [1] "Loading: rf_12"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_12_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_12_all_1min_data.csv"
## [1] "Loading: rf_13"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_13_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_13_all_1min_data.csv"
## [1] "Loading: rf_14"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_14_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_14_all_1min_data.csv"
## [1] "Loading: rf_15"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_15_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_15_all_1min_data.csv"
## [1] "Loading: rf_16"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_16_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_16_all_1min_data.csv"
## [1] "Loading: rf_17"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_17_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_17_all_1min_data.csv"
## [1] "Loading: rf_18"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_18_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
```

1 Citation
2 Introduction
3 Obtain listing of files
4 Load data files
5 Data quality analysis
6 Runtime
7 R environment

```
min/rf_18_all_1min_data.csv"
## [1] "Loading: rf_19"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_19_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_19_all_1min_data.csv"
## [1] "Loading: rf_20"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_20_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_20_all_1min_data.csv"
## [1] "Loading: rf_21"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_21_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_21_all_1min_data.csv"
## [1] "Loading: rf_22"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_22_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_22_all_1min_data.csv"
## [1] "Loading: rf_23"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_23_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_23_all_1min_data.csv"
## [1] "Loading: rf_24"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_24_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_24_all_1min_data.csv"
## [1] "Loading: rf_25"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_25_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_25_all_1min_data.csv"
## [1] "Loading: rf_26"
## [1] "Saved /Volumes/hum-csafe/Research
```

1 Citation

2 Introduction

3 Obtain listing of files

4 Load data files

5 Data quality analysis

6 Runtime

7 R environment

```
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_26_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_26_all_1min_data.csv"
## [1] "Loading: rf_27"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_27_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_27_all_1min_data.csv"
## [1] "Loading: rf_28"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_28_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_28_all_1min_data.csv"
## [1] "Loading: rf_29"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_29_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_29_all_1min_data.csv"
## [1] "Loading: rf_30"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_30_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_30_all_1min_data.csv"
## [1] "Loading: rf_31"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_31_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_31_all_1min_data.csv"
## [1] "Loading: rf_32"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_32_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_32_all_1min_data.csv"
## [1] "Loading: rf_33"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_33_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
```

1 Citation
2 Introduction
3 Obtain listing of files
4 Load data files
5 Data quality analysis
6 Runtime
7 R environment

```
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_33_all_1min_data.csv"
## [1] "Loading: rf_34"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_34_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_34_all_1min_data.csv"
## [1] "Loading: rf_35"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_35_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_35_all_1min_data.csv"
## [1] "Loading: rf_36"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_36_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_36_all_1min_data.csv"
## [1] "Loading: rf_37"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_37_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_37_all_1min_data.csv"
## [1] "Loading: rf_38"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_38_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_38_all_1min_data.csv"
## [1] "Loading: rf_39"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_39_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_39_all_1min_data.csv"
## [1] "Loading: rf_40"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_40_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_40_all_1min_data.csv"
## [1] "Loading: rf_41"
```

1 Citation

2 Introduction

3 Obtain listing of files

4 Load data files

5 Data quality analysis

6 Runtime

7 R environment

```
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_41_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_41_all_1min_data.csv"
## [1] "Loading: rf_42"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_42_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_42_all_1min_data.csv"
## [1] "Loading: rf_43"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_43_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_43_all_1min_data.csv"
## [1] "Loading: rf_44"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_44_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_44_all_1min_data.csv"
## [1] "Loading: rf_45"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_45_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_45_all_1min_data.csv"
## [1] "Loading: rf_46"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_46_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_46_all_1min_data.csv"
## [1] "Loading: rf_47"
## [1] "Saved /Volumes/hum-csafe/Research
Projects/GREEN Grid/Clean_data/gridSpy/1mi
n/rf_47_all_1min_data.csv, gzipping..."
## [1] "Gzipped /Volumes/hum-csafe/Researc
h Projects/GREEN Grid/Clean_data/gridSpy/1
min/rf_47_all_1min_data.csv"
```

Code

1 Citation

2 Introduction

3 Obtain listing of files

4 Load data files

5 Data quality analysis

6 Runtime

7 R environment

```
## [1] "Saving daily observations stats by hhid to /Volumes/hum-csafe/Research Projects/GREEN Grid/Clean_data/gridSpy/hhDailyObservationsStats.csv"
```

Code

```
## [1] "Done"
```

5 Data quality analysis

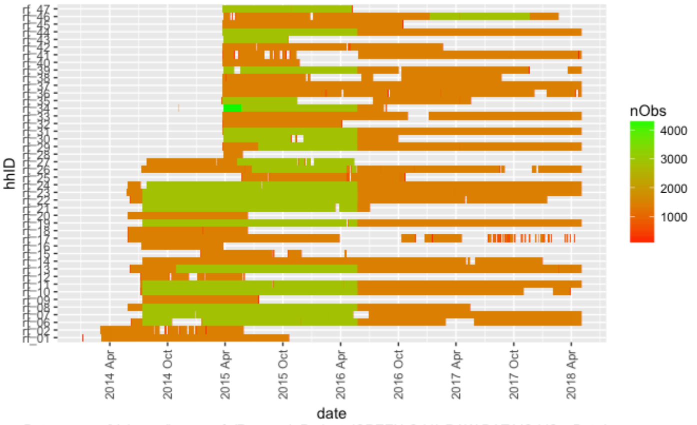
Now produce some data quality plots & tables.

The following plots show the number of observations per day per household. In theory we should not see:

- dates before 2014 or in to the future (they indicate data conversion errors)
- more than 1440 observations per day (they indicate potentially duplicate data)

Code

N observations per household per day for all loaded grid spy data



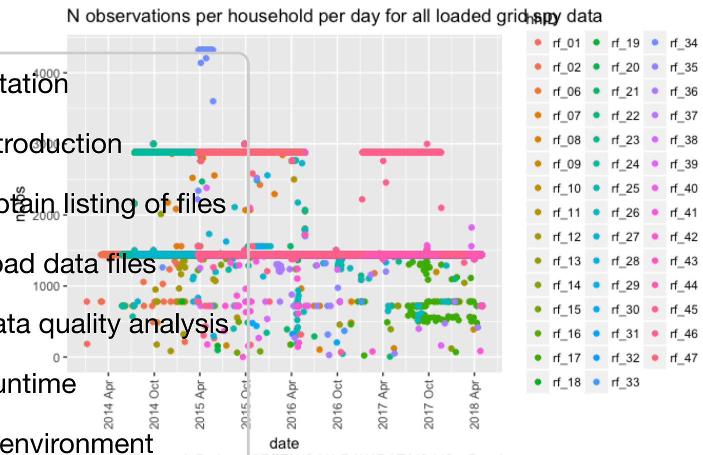
Data source: /Volumes/hum-csafe/Research Projects/GREEN Grid/_RAW DATA/GridSpyData/
Using data received up to 2018-05-03
Only files of size > 3000 bytes loaded

Code

```
## Saving 7 x 5 in image
```

Code

1 Citation
2 Introduction
3 Obtain listing of files
4 Load data files
5 Data quality analysis
6 Runtime
7 R environment



Volumes/hum-csafe/Research Projects GREEN Grid/_RAW DATA/GridSpyData/
Using data received up to 2018-05-03
Only files of size > 3000 bytes loaded

Code

```
## Saving 7 x 5 in image
```

The following table shows the min/max observations per day and min/max dates for each household. As above, we should not see:

- dates before 2014 or in to the future (indicates date conversion errors)
- more than 1440 observations per day (indicates potentially duplicate observations)
- non-integer counts of circuits as it suggests some column errors

We should also not see NA in any row (indicates date conversion errors).

If we do see any of these then we still have data cleaning work to do!

Code

Summary observation stats by hhID

hhID	minObs	maxObs	meanNDataColumns	minDate	maxDate
rf_01	171	1500	6	2014-01-05	2015-10-20
rf_02	215	1440	6	2014-03-02	2015-05-28
rf_06	486	3000	6	2014-06-08	2018-05-02
rf_07	105	3000	6	2014-07-13	2018-05-02
rf_08	123	3000	6	2014-05-28	2017-05-15
rf_09	163	1500	6	2014-07-13	2015-07-16
rf_10	389	2998	6	2014-07-08	2018-03-29
rf_11	556	3000	6	2014-07-07	2018-05-02

1 Citation
2 Introduction
3 Obtain listing of files
4 Load data files
5 Data quality analysis
6 Runtime
7 R environment

hhID	minObs	maxObs	meanNDataColumns	minDate	maxDate
rf_12	85	1500	6	2014-07-08	2015-06-02
rf_13	456	3000	6	2014-06-05	2018-05-02
rf_14	120	1500	6	2014-07-13	2017-12-30
rf_15	62	1440	6	2015-01-14	2016-04-18
rf_16	720	1500	6	2014-07-09	2015-03-25
rf_17	22	1500	6	2014-05-29	2018-04-11
rf_18	157	1500	6	2014-05-29	2015-06-11
rf_19	720	3000	9	2014-07-14	2018-05-02
rf_20	98	1500	6	2014-05-28	2015-06-11
rf_21	290	3000	6	2014-07-14	2016-07-01
rf_22	6	3000	6	2014-06-05	2018-01-14
rf_23	342	3000	6	2014-05-25	2018-05-02
rf_24	571	3000	6	2014-05-28	2018-05-02
rf_25	45	1500	6	2015-05-24	2016-10-22
rf_26	386	3000	6	2014-07-10	2018-05-02
rf_27	780	3000	6	2014-07-27	2016-05-13
rf_28	297	1440	6	2015-03-26	2015-05-26
rf_29	720	3000	6	2015-03-25	2018-05-02
rf_30	205	3000	6	2015-03-27	2016-09-29
rf_31	720	2998	6	2015-03-25	2018-05-02
rf_32	325	1500	6	2015-03-25	2016-04-05
rf_33	369	1500	6	2015-03-23	2018-05-02
rf_34	317	4320	6	2014-11-03	2016-08-24
rf_35	50	3000	6	2015-03-22	2017-05-17
rf_36	29	1500	6	2015-03-23	2018-05-02
rf_37	720	1500	6	2015-03-23	2018-05-02
rf_38	398	1500	6	2015-03-24	2017-08-22
rf_39	163	3000	5	2015-03-27	2018-05-02
rf_40	268	1500	6	2015-03-24	2015-11-22
rf_41	1	1573	6	2015-03-25	2018-05-02

hhID	minObs	maxObs	meanNDataColumns	minDate	maxDate
rf_42	79	1500	6	2015-03-23	2017-02-18
rf_43	1560	2990	6	2015-03-26	2015-10-18
rf_44	720	3000	6	2015-03-24	2018-05-02
rf_45	69	1499	6	2015-03-24	2016-10-15
rf_46	305	3000	13	2015-03-26	2018-02-19
rf_47	318	3000	6	2015-03-24	2016-05-08

1 Citation
2 Introduction
3 Obtain listing of files
4 Load data files
5 Data quality analysis
6 Runtime
7 R environment

6 Runtime

Code

Analysis completed in 1.000239710^{4} seconds (166.71 minutes) using knitr (https://cran.r-project.org/package=knitr) in RStudio (http://www.rstudio.com) with R version 3.4.4 (2018-03-15) running on x86_64-apple-darwin15.6.0.

7 R environment

R packages used:

- base R - for the basics [@baseR]
- data.table - for fast (big) data handling [@data.table]
- ggplot2 - for slick graphics [@ggplot2]
- dplyr - for select and contains [@dplyr]
- lubridate - date manipulation [@lubridate]
- knitr - to create this document [@knitr]
- greenGridr - for local NZ GREEN Grid utilities

Code

1 Citation
2 Introduction
3 Obtain listing of files
4 Load data files
5 Data quality analysis
6 Runtime
7 R environment

```
## R version 3.4.4 (2018-03-15)
## Platform: x86_64-apple-darwin15.6.0 (64
-bit)
## Running under: macOS High Sierra 10.13.
4
##
## Matrix products: default
## BLAS: /Library/Frameworks/R.framework/V
ersions/3.4/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework
/Versions/3.4/Resources/lib/libRlapack.dyl
ib
##
## locale:
## [1] en_GB.UTF-8/en_GB.UTF-8/en_GB.UTF-8
/C/en_GB.UTF-8/en_GB.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils
datasets  methods   base
##
## other attached packages:
## [1] knitr_1.20          dplyr_0.7.4
readr_1.1.1
## [4] ggplot2_2.2.1       lubridate_1.7.4
data.table_1.10.4-3
## [7] greenGridr_0.1.0
##
## loaded via a namespace (and not attache
d):
##  [1] Rcpp_0.12.16      bindr_0.1.1
magrittr_1.5
##  [4] hms_0.4.2         munsell_0.4.3
colorspace_1.3-2
##  [7] R6_2.2.2          rlang_0.2.0.9001
highr_0.6
## [10] stringr_1.3.0     plyr_1.8.4
tools_3.4.4
## [13] grid_3.4.4        gtable_0.2.0
htmltools_0.3.6
## [16] assertthat_0.2.0  yaml_2.1.18
lazyeval_0.2.1
## [19] rprojroot_1.3-2   digest_0.6.15
tibble_1.4.2
## [22] bindrcpp_0.2.2    glue_1.2.0
evaluate_0.10.1
## [25] rmarkdown_1.9     labeling_0.3
stringi_1.1.7
## [28] compiler_3.4.4    pillar_1.2.2
scales_0.5.0.9000
## [31] backports_1.1.2   pkgconfig_2.0.1
```