Processing, cleaning and saving NZ GREEN Grid project time use diary data

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

Last run at: 2018-05-22 09:46:35

Contents

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 time use diary data, University of Otago: Dunedin, NZ.

2 Introduction

Report circulation:

- Restricted to: NZ GREEn Grid project partners and contractors.

2.1 Purpose

This report is intended to:

- load and clean the two time use survey datasets
- save the cleaned data out to /Volumes/hum-csafe/Research Projects/GREEN Grid/Clean_data/safe/TUD/ as two seperate files, one for each survey
- produce summary data quality statistics

2.2 Requirements:

Time use survey data held in /Volumes/hum-csafe/Research Projects/GREEN Grid/_RAW DATA/Time Use Diaries/:

- PowerCo
- Unison

A lookup table to correct mis-coding of household IDs (/Volumes/hum-csafe/Research Projects/GREEN Grid/_RAW DATA/TUD_2_GridSpyLookup.xlsx).

2.3 History

Generally tracked via our git.soton repo:

- history
- issues

2.4 Support

This work was supported by:

- The University of Otago
- The New Zealand Ministry of Business, Innovation and Employment (MBIE)
- SPATIALEC - a Marie Skłodowska-Curie Global Fellowship based at the University of Otago's Centre for Sustainability (2017-2019) & the University of Southampton's Sustainable Energy Research Group (2019-202).

We do not 'support' the code but if you have a problem check the issues on our repo and if it doesn't already exist, open one. We might be able to fix it :-)

3 PowerCo

This consists of 1 file found in /Volumes/hum-csafe/Research Projects/GREEN Grid/_RAW DATA/Time Use Diaries/Powerco/Powerco Annexes/:

- TUD (Merged data)_BA.csv

This is a version of TUD (Merged data).csv with:

- small edits to correct dates
- redundant rows removed from file header

3.1 Load & process

```
## [1] "Found 352 rows of data"
```

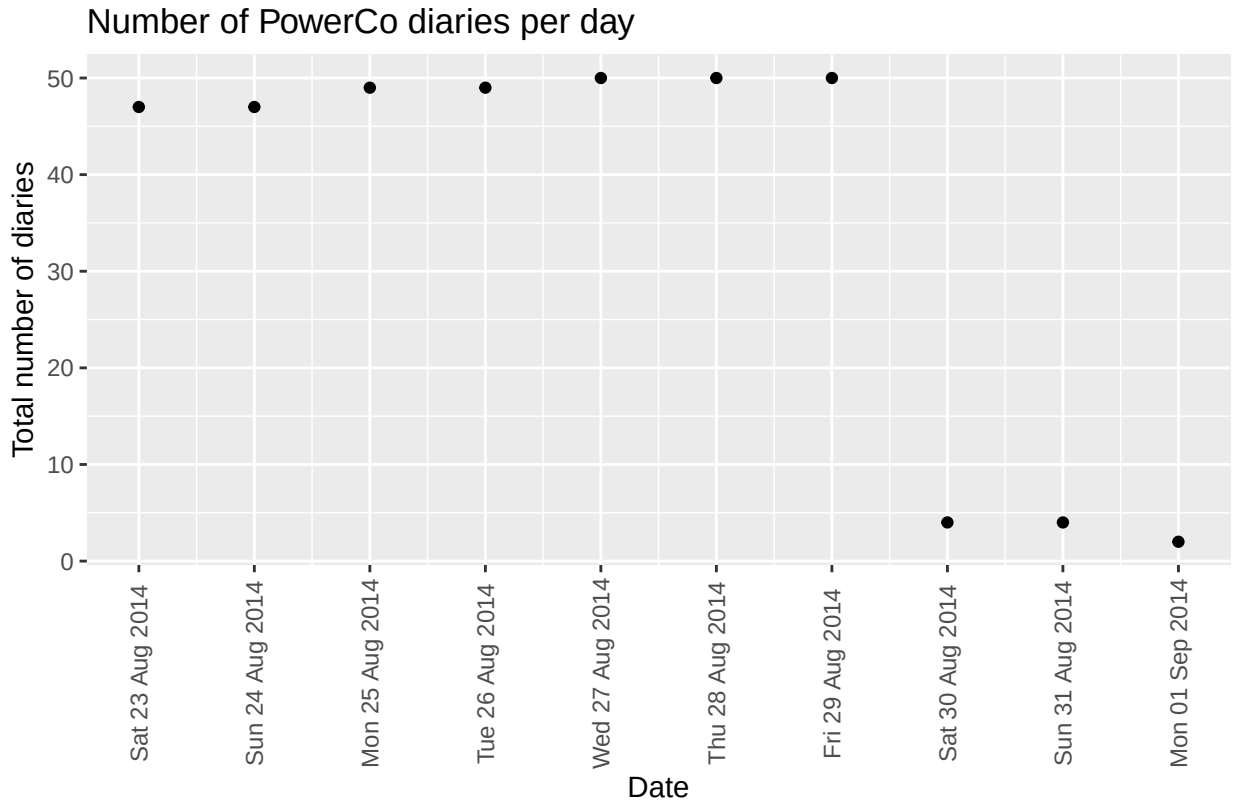
Table 1: Summary of PowerCo diaries by household

hhID	nDiaries	familySize	minDiaryDate	maxDiaryDate
rf_06	14	2.000000	2014-08-23	2014-08-29
rf_07	14	3.000000	2014-08-25	2014-08-31
rf_08	7	1.000000	2014-08-23	2014-08-29
rf_09	14	2.000000	2014-08-23	2014-08-29
rf_10	14	2.000000	2014-08-23	2014-08-29
rf_11	7	1.000000	2014-08-23	2014-08-29
rf_12	14	3.000000	2014-08-23	2014-08-29
rf_13	12	2.000000	2014-08-23	2014-08-29
rf_14	43	5.906977	2014-08-23	2014-08-29
rf_15	14	3.000000	2014-08-23	2014-08-29
rf_16	14	3.000000	2014-08-23	2014-08-29
rf_17	14	2.000000	2014-08-26	2014-09-01
rf_18	14	2.000000	2014-08-23	2014-08-29
rf_19	14	3.000000	2014-08-23	2014-08-29
rf_20	35	6.000000	2014-08-23	2014-08-29
rf_21	14	2.000000	2014-08-23	2014-08-29
rf_22	14	2.000000	2014-08-23	2014-08-29
rf_23	14	4.000000	2014-08-23	2014-08-29
rf_24	28	4.000000	2014-08-23	2014-08-29
rf_25	21	4.000000	2014-08-23	2014-08-29
rf_26	7	1.000000	2014-08-23	2014-08-29
rf_27	10	4.000000	2014-08-23	2014-08-29

```
## [1] "Saving PowerCo cleaned time use diary to /Volumes/hum-csafe/Research Projects/GREEN Grid/Clean_
## [1] "Done"
```

3.2 Tests

Should all be in August 2014...



ɔe: /Volumes/hum−csafe/Research Projects/GREEN Grid/_RAW DATA/Time Use Diaries/Powerco/Powerco Annexes/

```
## Saving 6.5 x 4.5 in image
```

In total we have 352 diaries from 22 PowerCo households.

4 Unison

This consists of 5 files found in /Volumes/hum-csafe/Research Projects/GREEN Grid/_RAW DATA/Time Use Diaries/Unison/Unison Raw Data/Raw data with paper diaries included/Cleaned excel data files/:

- TUDAdult_ONE_Child_Unison_forSAS_BA.xlsx
- TUDAdult_TWO_Children_Unison_forSAS_BA.xlsx
- TUDAdult-THREE-Children-Unison_forSAS_BA.xlsx
- TUDAdult-Unison-forSAS_BA.xlsx
- TUDTeenagerorChild-Unison_forSAS_BA.xlsx

As before these are copies of the original versions with slight editing to correct dates and for ease of processing. The relationship between them is currently unclear!

4.1 Load & process

```
## [1] "Found 352 rows in total"
```

Table 2: Test diaryDates that did not parse

ResponseID	r_diaryDate	tudCode	StartDate	EndDate
NA	NA	NA	NA	NA

ResponseID	r_diaryDate	tudCode	StartDate	EndDate
NA	NA	NA	NA	NA
NA	NA	NA	2015-07-21 21:12:46	2015-07-21 21:13:00

Table 3: Report diaries with edited diary dates (done in .xlsx before loading)

r_diaryDate	tudCode	dateNote	sourceFile
2015-07-20	28	imputed	TUDAdult_ONE_Child_Unison_forSAS_BA.xlsx
2015-07-21	28	imputed	TUDAdult_ONE_Child_Unison_forSAS_BA.xlsx
2015-07-20	33	imputed	TUDAdult_ONE_Child_Unison_forSAS_BA.xlsx
2015-07-20	39	imputed	TUDAdult_ONE_Child_Unison_forSAS_BA.xlsx
2015-07-23	39	imputed	TUDAdult_ONE_Child_Unison_forSAS_BA.xlsx
2015-07-24	39	imputed	TUDAdult_ONE_Child_Unison_forSAS_BA.xlsx
2015-07-26	39	imputed	TUDAdult_ONE_Child_Unison_forSAS_BA.xlsx
2015-07-20	39	imputed	TUDAdult_ONE_Child_Unison_forSAS_BA.xlsx
2015-07-20	41	might actually be the 20th	TUDAdult_TWO_Children_Unison_forSAS_BA.xlsx
2015-07-21	41	might actually be the 21st	TUDAdult_TWO_Children_Unison_forSAS_BA.xlsx
2015-07-20	41	imputed from StartDate	TUDAdult_TWO_Children_Unison_forSAS_BA.xlsx
2015-07-21	41	imputed from StartDate	TUDAdult_TWO_Children_Unison_forSAS_BA.xlsx
2015-07-22	41	imputed from StartDate	TUDAdult_TWO_Children_Unison_forSAS_BA.xlsx
2015-07-23	41	imputed from StartDate	TUDAdult_TWO_Children_Unison_forSAS_BA.xlsx
2015-07-24	41	imputed from StartDate	TUDAdult_TWO_Children_Unison_forSAS_BA.xlsx
2015-07-25	41	imputed from StartDate	TUDAdult_TWO_Children_Unison_forSAS_BA.xlsx
2015-07-26	41	imputed from StartDate	TUDAdult_TWO_Children_Unison_forSAS_BA.xlsx
2015-07-21	31	corrected to July from Feb	TUDTeenagerorChild-Unison_forSAS_BA.xlsx
2015-07-26	45	25/7/2015 missing in original	TUDTeenagerorChild-Unison_forSAS_BA.xlsx

Table 4: Summary of Unison diaries by household

tudCode	nDiaries	minDiaryDate	maxDiaryDate
NA	3	NA	NA
28	21	2015-07-20	2015-07-26
29	14	2015-07-20	2015-07-26
30	14	2015-07-20	2015-07-26
31	21	2015-07-20	2015-07-26
32	21	2015-07-20	2015-07-26
33	14	2015-07-20	2015-07-26
34	14	2015-07-20	2015-07-26
35	14	2015-07-20	2015-07-26
36	14	2015-07-20	2015-07-26
37	14	2015-07-20	2015-07-26
38	21	2015-07-20	2015-07-26
39	21	2015-07-20	2015-07-26
40	14	2015-07-20	2015-07-26
41	21	2015-07-20	2015-07-26
42	21	2015-07-20	2015-07-26
43	14	2015-08-03	2015-08-09
44	14	2015-07-20	2015-07-26
45	37	2015-07-20	2015-07-26

tudCode	nDiaries	minDiaryDate	maxDiaryDate
46	11	2015-07-20	2015-07-26
47	14	2015-07-20	2015-07-26

The tudCodes are *not* the gridSpy ids, we need to create these from the unison sheet in /Volumes/humcsafe/Research Projects/GREEN Grid/_RAW DATA/TUD_2_GridSpyLookup.xlsx.

Table 5: Linking table

CODE	tag_gridSpy_Hhid	source
28	rf_33	unison
29	rf_46	unison
30	rf_37	unison
31	rf_28	unison
32	rf_39	unison
33	rf_29	unison
34	rf_30	unison
35	rf_31	unison
36	rf_43	unison
37	rf_35	unison
38	rf_44	unison
39	rf_41	unison
40	rf_36	unison
41	rf_42	unison
42	rf_34	unison
43	rf_38	unison
43	rf_38	unison
44	rf_32	unison
45	rf_47	unison
46	rf_45	unison
47	rf_40	unison

Table 6: Check linkage: there should be 1 or 2 diaries for each combination

linkCode	hhID	nDiaries
28	rf_33	21
29	rf_46	14
30	rf_37	14
31	rf_28	21
32	rf_39	21
33	rf_29	14
34	rf_30	14
35	rf_31	14
36	rf_43	14
37	rf_35	14
38	rf_44	21
39	rf_41	21
40	rf_36	14
41	rf_42	21
42	rf_34	21

linkCode	hhID	nDiaries
43	rf_38	28
44	rf_32	14
45	rf_47	37
46	rf_45	11
47	rf_40	14

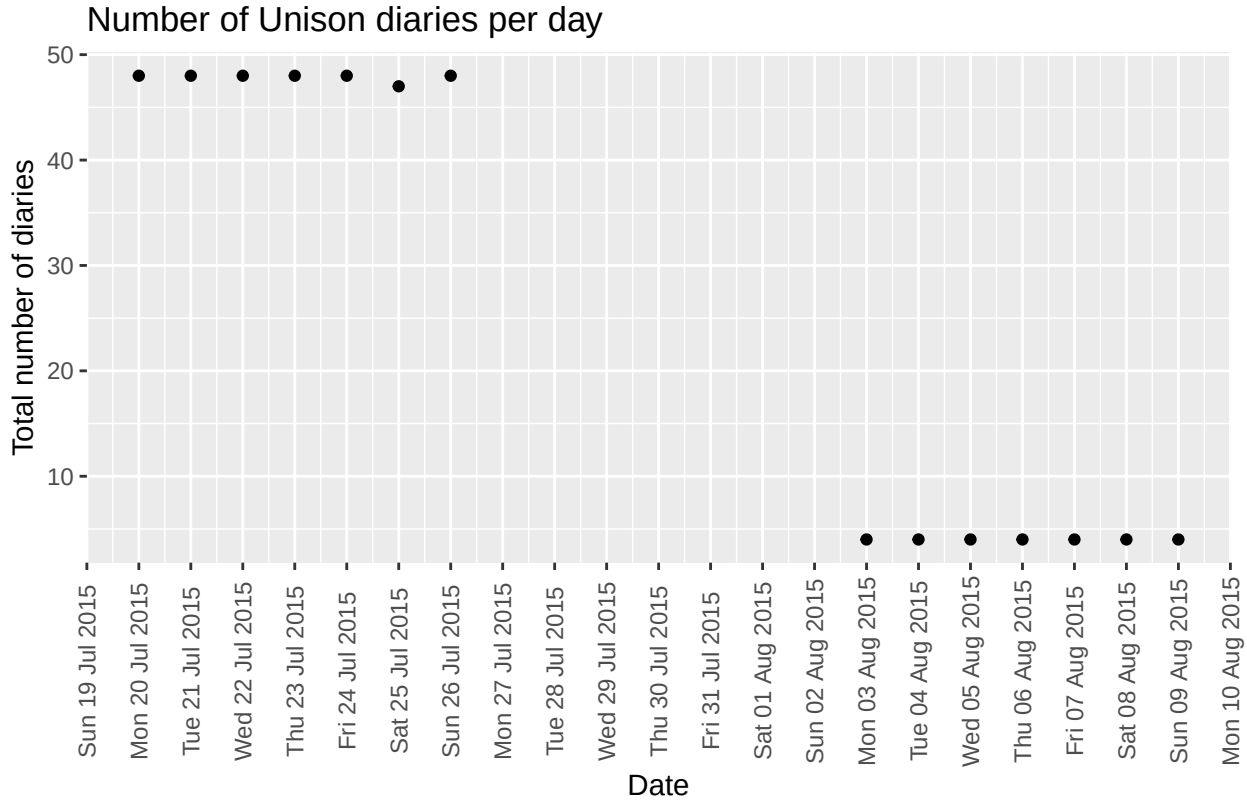
In total we have 363 diaries from 20 Unison households.

```
## [1] "Saving Unison cleaned time use diary to /Volumes/hum-csafe/Research Projects/GREEN Grid/Clean_da
```

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

4.2 Tests

All of the diaries should be in July/August 2015...



AW DATA/Time Use Diaries/Unison/Unison Raw Data/Raw data with paper diaries included/Cleaned excel data files/

```
## Saving 6.5 x 4.5 in image
```

If any of them are earlier than July 2015 they are flagged below for ease of fixing.

Table: Households with potential diary date errors

r_diaryDate tudCode sourceFile nDiaries ———— ——— ———— ———

5 Summary

6 Runtime

Analysis completed in 13.31 seconds (0.22 minutes) using knitr in RStudio with R version 3.4.4 (2018-03-15) running on x86_64-apple-darwin15.6.0.

7 R environment

R packages used: data.table, lubridate, ggplot2, readr, dplyr, readxl, knitr

- base R - for the basics (R Core Team 2016)
- data.table - for fast (big) data handling (Dowle et al. 2015)
- lubridate - date manipulation (Grolemund and Wickham 2011)
- ggplot2 - for slick graphics (Wickham 2009)
- readr - for csv reading/writing (Wickham, Hester, and Francois 2016)
- dplyr - for select and contains (Wickham and Francois 2016)
- knitr - to create this document (Xie 2016)
- nzGREENGrid - for local NZ GREEN Grid utilities

```
## 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/Versions/3.4/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/3.4/Resources/lib/libRlapack.dylib
##
## 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          readxl_1.1.0        dplyr_0.7.4
## [4] readr_1.1.1         ggplot2_2.2.1.9000  lubridate_1.7.4
## [7] data.table_1.10.4-3 nzGREENGrid_0.1.0
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_0.12.16      pillar_1.2.2      compiler_3.4.4
##  [4] cellranger_1.1.0  plyr_1.8.4        highr_0.6
##  [7] bindr_0.1.1       tools_3.4.4       digest_0.6.15
## [10] evaluate_0.10.1   tibble_1.4.2      gtable_0.2.0
## [13] pkgconfig_2.0.1   rlang_0.2.0.9001  yaml_2.1.18
## [16] bindrcpp_0.2.2    withr_2.1.2       stringr_1.3.0
## [19] hms_0.4.2         rprojroot_1.3-2   grid_3.4.4
## [22] glue_1.2.0        R6_2.2.2          rmarkdown_1.9
## [25] magrittr_1.5      backports_1.1.2   scales_0.5.0.9000
## [28] htmltools_0.3.6   assertthat_0.2.0  colorspace_1.3-2
## [31] labeling_0.3      stringi_1.1.7     lazyeval_0.2.1
## [34] munsell_0.4.3
```

References

Dowle, M, A Srinivasan, T Short, S Lianoglou with contributions from R Saporta, and E Antonyan. 2015. *Data.table: Extension of Data.frame*. https://CRAN.R-project.org/package=data.table.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. http://www.jstatsoft.org/v40/i03/.

R Core Team. 2016. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2009. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. http://ggplot2.org.

Wickham, Hadley, and Romain Francois. 2016. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Romain Francois. 2016. *Readr: Read Tabular Data*. https://CRAN.R-project.org/package=readr.

Xie, Yihui. 2016. *Knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://CRAN.R-project.org/package=knitr.