Accelerometry: Data structure and analysis

Andrew Leroux & Jacek Urbanek & Ciprian Crainiceanu

February 22, 2019

NHANES accelerometry: Reproducing these Analyses

- All analyses presented here can be replicated using the "CSS_NHANES.R" script located at `https://www.github.com/andrew-leroux/CSS_NHANES/`
- Steps:
  - Download or clone
  - Open R project ("CSS_NHANES.Rproj")
  - Open R script "CSS_NHANES.R"
  - Run code

NHANES accelerometry

- The National Health and Nutrition Survey is a cross-sectional study of the US population performed in 2-year waves
- Complex survey structure (beyond the scope of this talk)
- Accelerometry data available for the 2003-2004 and 2005-2006 waves
  - Acceleration summarized into minute-level "activity counts"
  - Up to 7 days of data for each participant
  - Study protocol: remove the device at bedtime

NHANES accelerometry: data structure

- Accelerometry data downloadble from NHANES is in long format
- Very large file sizes ($\approx$ 2.5 GB)
- Unintuitive data structure

SEQN	PAXSTAT	PAXCAL	PAXDAY	PAXN	PAXHOUR	PAXMINUT	PAXINTEN	PAXSTEP
31128	1	1	1	1	0	0	166	4
31128	1	1	1	2	0	1	27	0
31128	1	1	1	3	0	2	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

NHANES accelerometry: proposed data strucutre

- Wide format instead of long format[1] ($\approx$ 60 MB)
- 7 rows per participant, descending cronologoical order

	Unique Identifier		Quality Flags		NHANES wave	Activity Counts				
	SEQN	PAXDAY	PAXCAL	PAXSTAT	SDDSRVYR	MIN1	MIN2	MIN3	$\cdots$	MIN1440
(a)	31128	1	1	1	4	166	27	0	$\cdots$	0
	31128	2	1	1	4	0	0	0	$\cdots$	0
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	31128	7	1	1	4	0	0	0	$\cdots$	0
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
(b)	31193	2	2	1	4	0	0	0	$\cdots$	1921
	31193	3	2	1	4	335	2598	2185	$\cdots$	46
	31193	4	2	1	4	0	0	0	$\cdots$	0
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
(c)	31880	2	2	2	4	32767	32767	32767	$\cdots$	32767
	31880	3	2	2	4	32767	32767	32767	$\cdots$	32767
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
(d)	32008	5	1	2	4	0	0	0	$\cdots$	0
	32008	6	1	2	4	NA	NA	NA	$\cdots$	NA

[1] Leroux, A., Di, J., Smirnova, E. et al. Stat Biosci (2019). https://doi.org/10.1007/s12561-018-09229-9

NHANES accelerometry: *rnhanesdata* package

(1) Processed data	
processed physical activity data	"PAXINTEN_C.rda" and "PAXINTEN_D.rda"
wear/non-wear flags data	"Flags_C.rda" and "Flags_D.rda"
covariates data	"Covariate_C.rda" and "Covariate_D.rda"
mortality data	"Mortality_2011_C.rda" and "Mortality_2011_D.rda"
(2) Data processing functions	
NHANES activity processing code	"process_accel()"
NHANES wear/non-wear flag code	"process_flags()"
NHANES mortality	"process_mort()"
NHANES data merging	"process_covar()"
(3) Helper functions	
Calculate survey weights on subsets	"reweight_accel()"
Identify "good" days of accelerometry data	"exclude_accel()"
(4) Raw data	
NHANES covariate data	"ALQ_C.XPT","ALQ_D.XPT", "BMX_C.XPT","BMX_D.XPT", ...
NHANES linked mortality data	"NHANES_2005_2006_MORT_2011_PUBLIC.dat" "NHANES_2003_2004_MORT_2011_PUBLIC.dat"

[1] Leroux, A., Di, J., Smirnova, E. et al. Stat Biosci (2019). https://doi.org/10.1007/s12561-018-09229-9

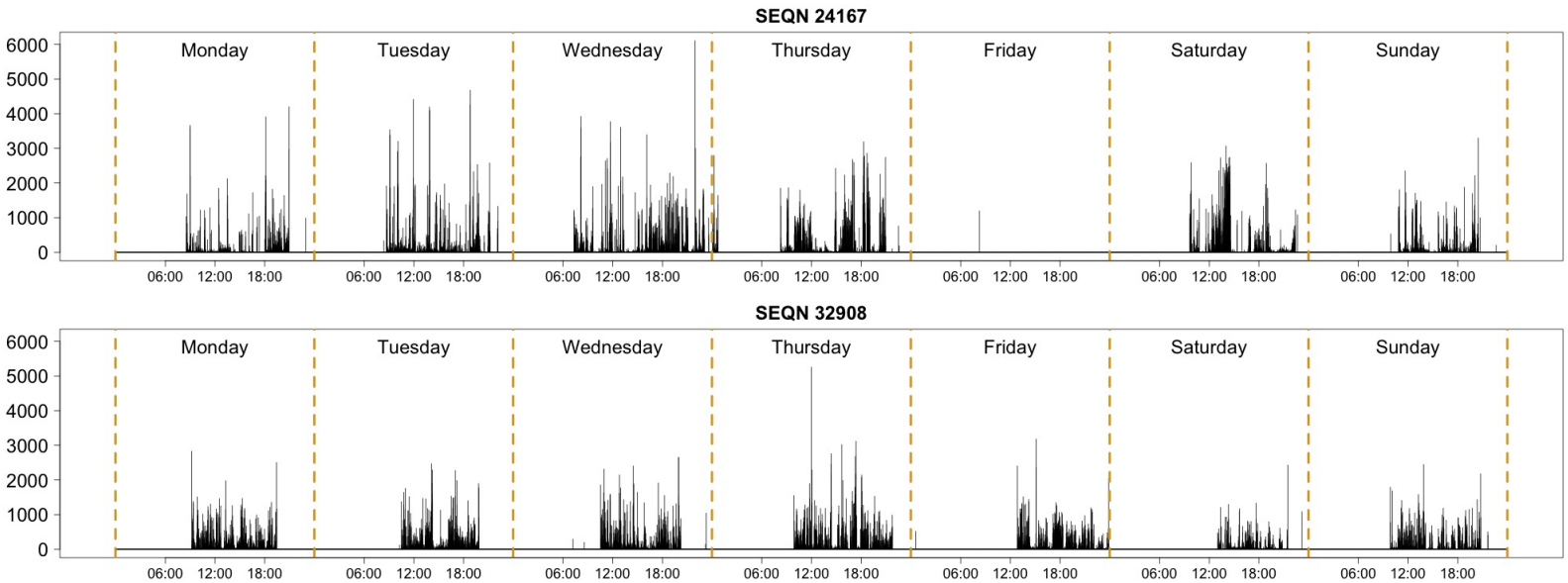
rnahnesdata: Package Installation

- Package installation may take a few minutes due to the size of the processed data.
- Requires the devtools package
- See ?"rnhanesdata-package" for details

```
if(!require("rnhanesdata")){
    devtools::install_github("andrew-leroux/rnhanesdata")
    require("rnhanesdata")
}
```

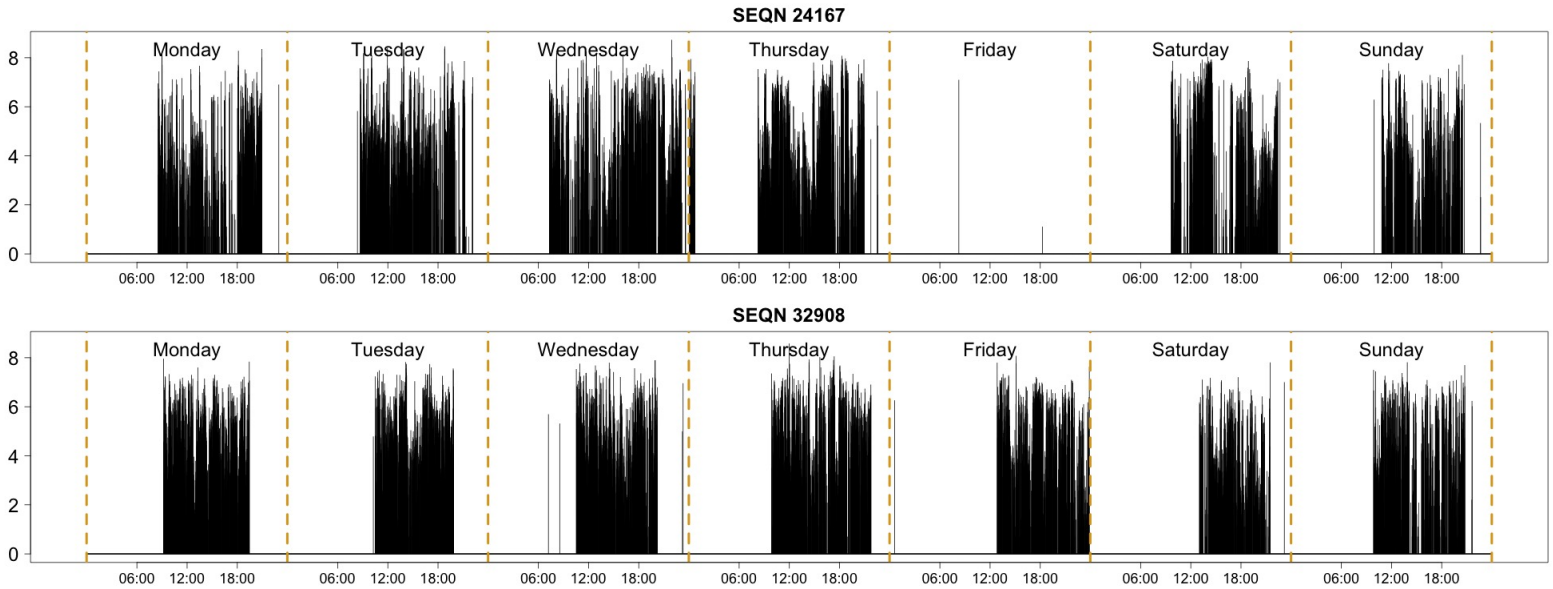
NHANES accelerometry: EDA

- 7 days of data for two participants at the minute level
- Estimated non-wear time has been imputed as 0
- Dominated by a few very large values



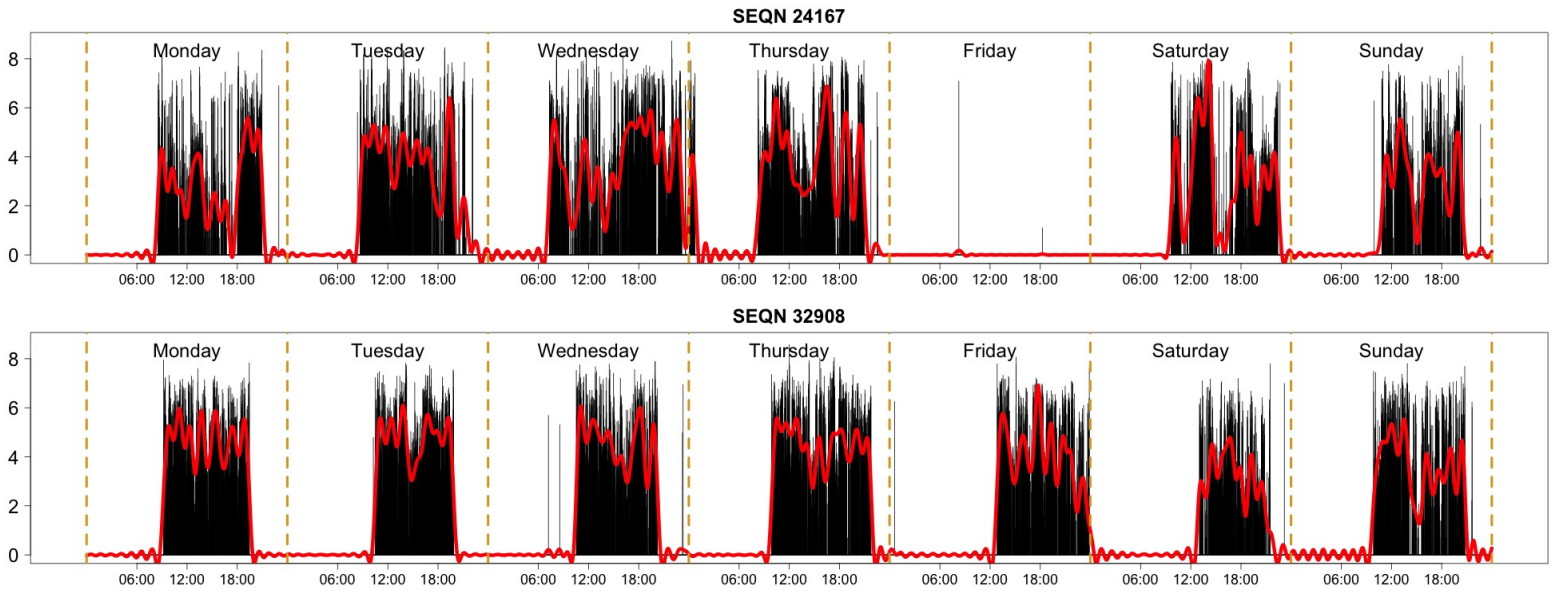
NHANES accelerometry: EDA

- Apply a $\log(1 + x)$ transformaiton at the minute level
- Still a high degree of minute-to-minute variability



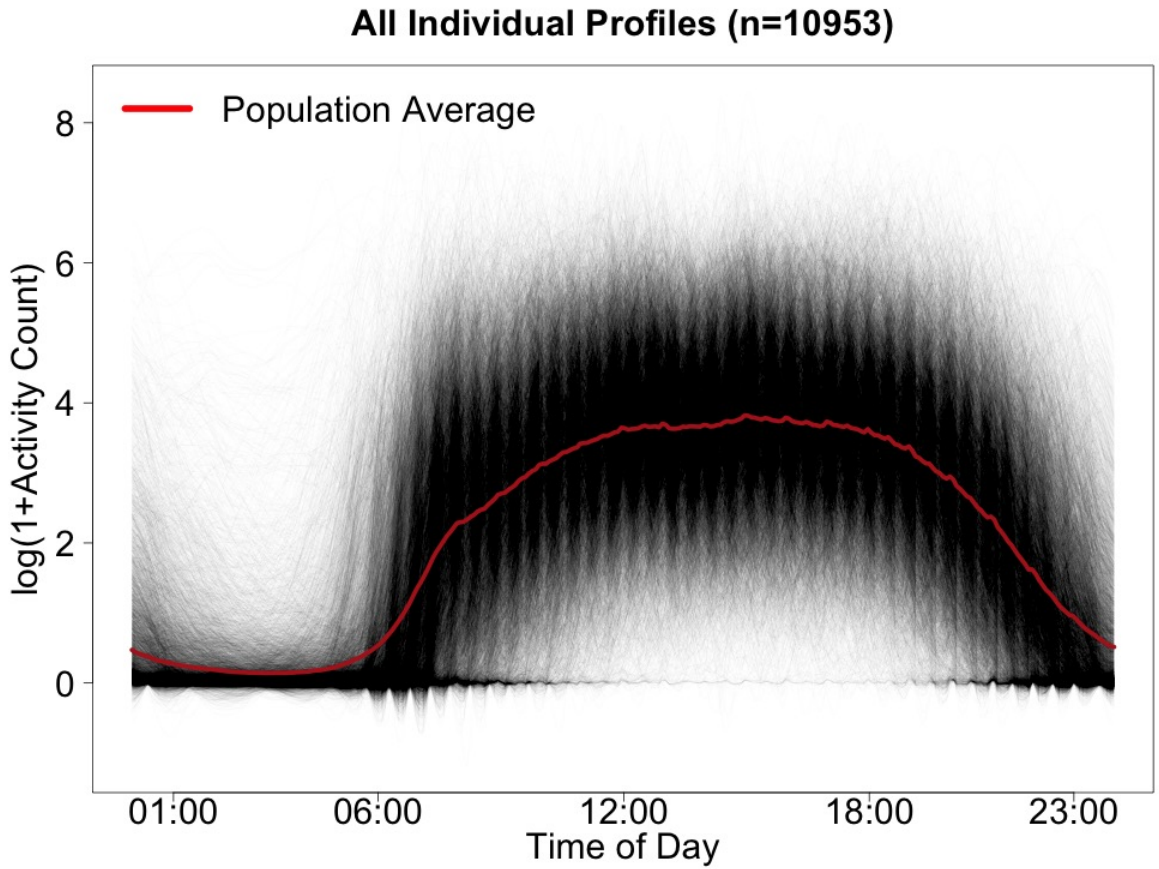
NHANES accelerometry: EDA

- Apply a $\log(1 + x)$ transformaiton at the minute level
- Still a high degree of minute-to-minute variability
- Smooth the data



NHANES accelerometry: EDA

- Visualizing the whole population
- Subset based on ≥ 10 hours of wear time, smooth the log transformed data, then average profiles across days within participants



NHANES accelerometry: Analysis Procedure

- Load and merge any relevant data by unique identifier (SEQN)
- Apply exclusion criteria
  - Data quality: 1) device calibration (PAXCAL); and 2) NHANES supplied flag (PAXSTAT)
  - Adherence to wear-time protocol. Most studies use ≥ 10 hours.
  - Sufficient number of days of data. Most studies use ≥ 3 days of data with ≥ 10 hours of wear.
  - Other criteria: missing data, etc.
- Apply binning or smoothing to the activity data if desired
- Calculate features of interest
- Incorporate survey design? Survey weights?
- Regresison, machine learning, etc.

NHANES accelerometry: Features

- What even is an "activity count"?
- Current standard: calculate single summaries of the data
  - Volume of activity[2]
    - Time spent in sedentary/light/moderate/vigorous behaviours. Require population-specific studies to determine thresholds.
    - Average daily total activity count (TAC). A proxy for total volume of moderate/vigorous activity
    - Average daily total log activity count (TLAC). A proxy for total volume of low/light activity
  - Patterns of activity
    - Fragmentation measures[3]
    - Timing of physical activity (activity profiles)
- Here, we focuse on analyzing patterns of activity using subject-specific average activity profiles.

---

[3]Varma VR, Dey D, Leroux A, et al. Total volume of physical activity: TAC, TLAC or TAC(λ). Prev Med. 2017;106:233-235.

[3]Di, J., Leroux, A., Urbanek, J., et al. Patterns of sedentary and active time accumulation are associated with mortality in US adults: The NHANES study. bioRxiv: 182337.

NHANES accelerometry: Analysis Procedure

- In the subsequent analyses presented here we work with activity profiles
- The data is smoothed, binned into 5 minute intervals, then averaged across days. This is done separately for the un-transformed and log-transformed activity counts
- Binning is done mostly to reduce computational burden
- None of the results here adjust for the survey design of NHANES

Some Thoughts

- D

Open Problems

- F

In class exercises

- Find