Station Usage in Seattle Bike Share

Arman Bilge[1,2], Yanbo Ge[3], and Xiaoliu Wu[2]

[1]Computational Biology Program, Fred Hutchinson Cancer Research Center and Departments of [2]Statistics and [3]Civil and Environmental Engineering, University of Washington

Background

- Pronto! has been Seattle's bike share system for 2 years
- Users must start and end their trips at various stations across the city
- Program is shutting down at the end of March due to insufficient use
- Many reasons suggested for its failure; e.g., poor placement of stations

Questions

- Which features of a station affect its daily usage?
- Can we use these to decide where to place stations?

Data

- Pronto! (October 2014–August 2016)
  - record for every trip taken (197810 trips in total)
  - daily weather report (temperature, precipitation, wind, etc.)
- Google Maps API
  - elevation
  - nearby points-of-interest (transportation, etc.)
- Puget Sound Regional Council Census
  - socio-economic data (population density, job density, income, etc.)

Challenges (and solutions)

- Stark contrast between weekday and weekend usage patterns
  - Only analyze data from weekdays
- Overall system usage depends heavily on daily weather
  - Control for temperature and precipitation in our model
- Some stations were decomissioned or installed at later dates
  - Remove stations that have been active for less than 20% of time frame
- System has complex, network dependencies between stations
  - Only count daily departures (and not arrivals) per station, called *outflow*
  - Try to find a subset of independent stations
- No data on total membership of the system
  - Conceded limitation of our analysis

Assumptions

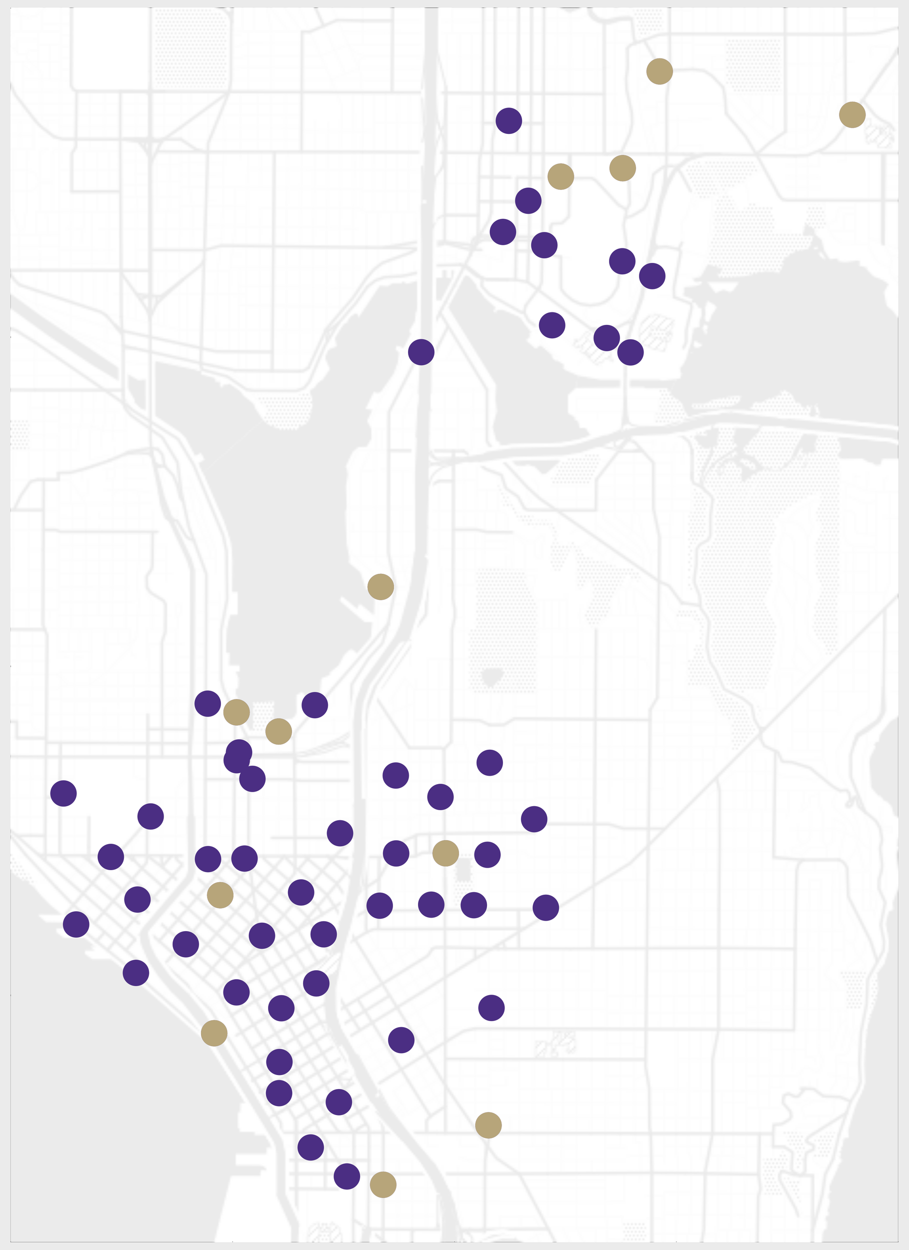
Let $\underset{\sim}{y}_i = [y_{ij}]^\intercal$, where y_{ij} is the (transformed) total outflow of the jth station on the ith day. Then we assume that

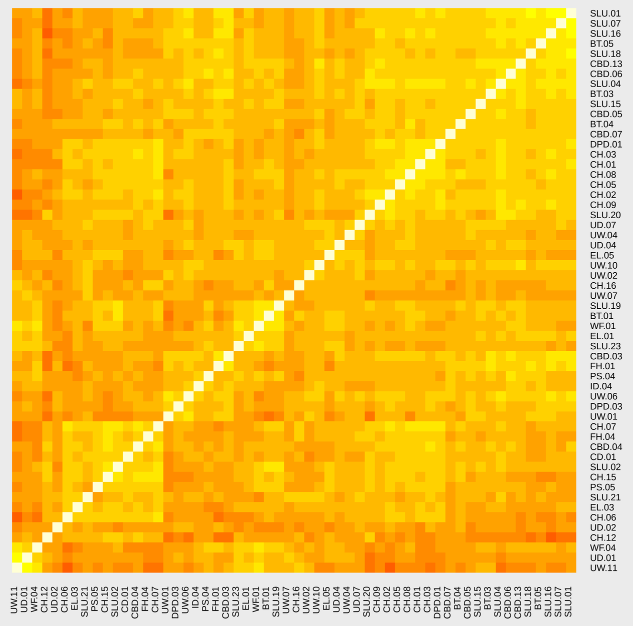
$$\underset{\sim}{y}_i = X\underset{\sim}{\beta}_s + \underset{\sim}{w}_i^\intercal \underset{\sim}{\beta}_w \underset{\sim}{1} + \underset{\sim}{\epsilon}_i, \underset{\sim}{\epsilon}_i \sim \mathcal{N}(\underset{\sim}{0}, \Sigma)$$

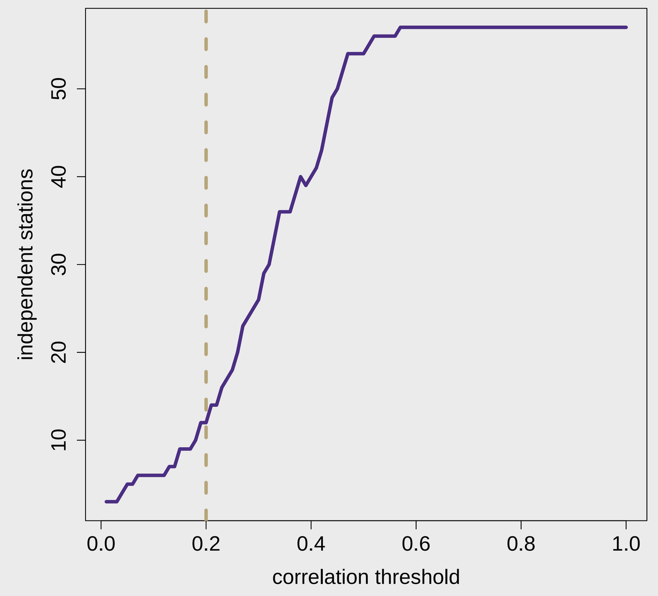
where X are the station features and $\underset{\sim}{w}_i$ is the weather for the ith day. Note that we do not assume Σ is diagonal to account for inter-station correlations. Suppose we find a set of stations S_δ such that $\forall i, j \in S_\delta : \sigma_{ij} < \delta$. Then our model simplifies to

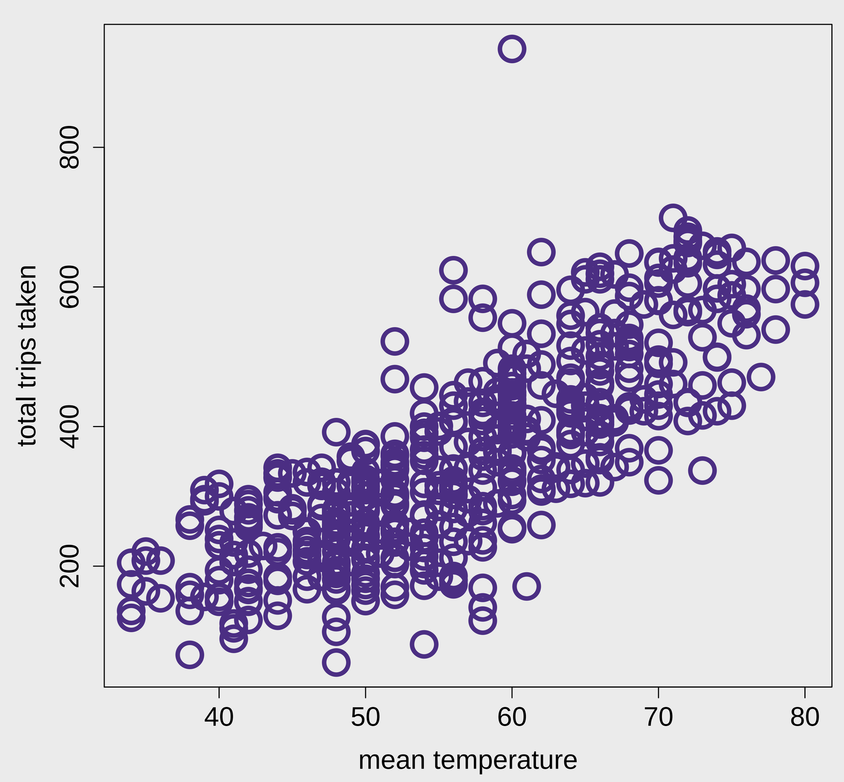
$$y_{ij} \approx \underset{\sim}{x}_j \underset{\sim}{\beta}_s + \underset{\sim}{w}_i^\intercal \underset{\sim}{\beta}_w + \epsilon ij, \epsilon ij \sim \mathcal{N}(0, \sigma_{jj})$$

Results



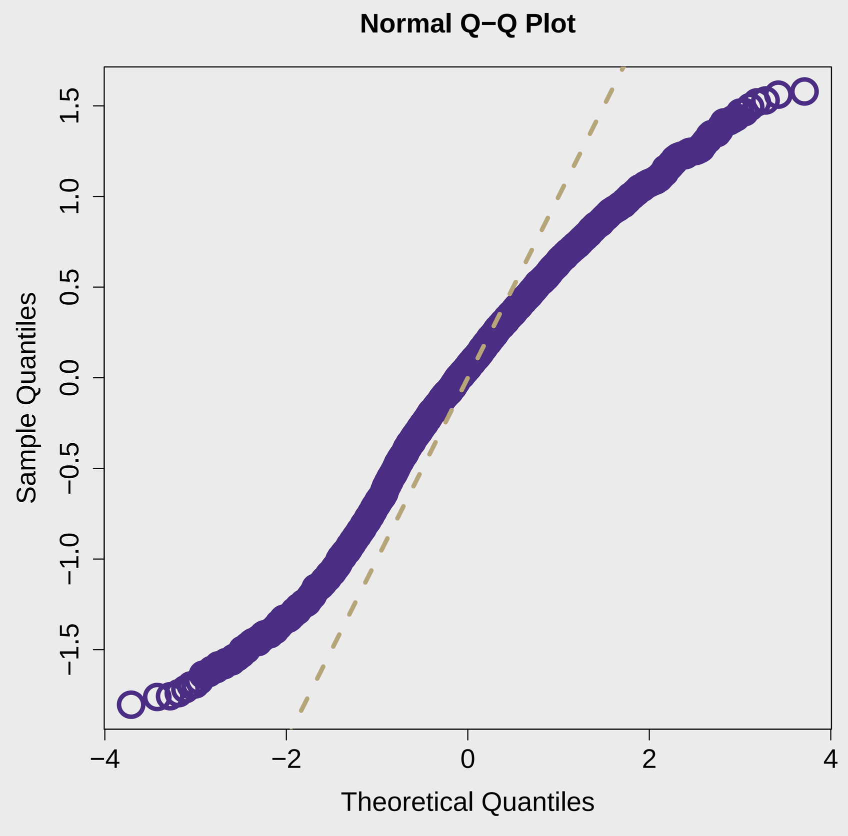


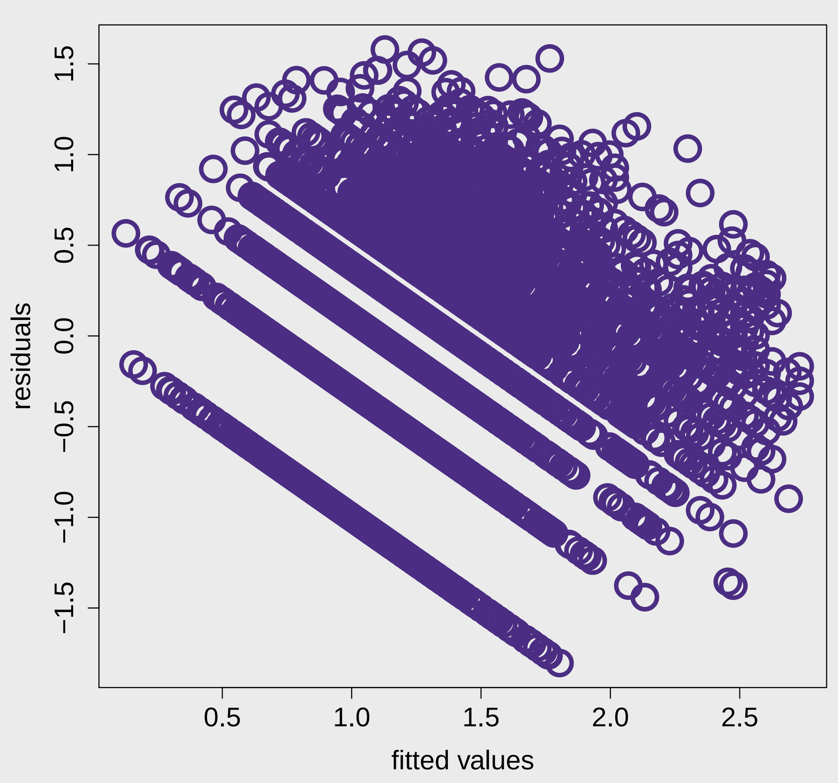




	Model 1
(Intercept)	0.924 (0.083)***
elevation	−6.239 (0.532)***
population density	−0.035 (0.003)***
job density	0.007 (0.000)***
income	−0.007 (0.004)
number metro routes	−0.010 (0.001)***
distance to metro	4.428 (0.609)***
listed on Google maps	−0.601 (0.033)***
mean temp.	0.021 (0.001)***
precip.	−0.536 (0.042)***
R^2	0.327
Adj. R^2	0.325
Num. obs.	4831
RMSE	0.628

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$





Methods

1. log-transform the per-station daily outflow to stabilize variance
2. Following technique for seemingly unrelated regression (SUR):
   (a) Regress on weather covariates using ordinary least squares (OLS)
   (b) Estimate $\hat{\Sigma}$ from the residuals
3. Use $\hat{\Sigma}$ to select independent subset of stations S_δ (e.g., with a connected-components algorithm)
4. Fit model with station features and weather covariates to data from selected stations using OLS

Validation

- Two independent code bases gave identical results
- Although not exactly normallly distributed, residuals are symmetric
- Transformed observations demonstrate homoscedasticity (constant variance)

Concluding Remarks

- Linear regression is a challenging framework for analyzing a network; we appropriately subsampled our data to minimize violation of assumptions
- Controlling for weather, all station features considered were significant
  - Surprisingly, elevation was negative correlated with departures
  - Station proximity to transit decreased departures
  - Also surprisingly, stations listed on Google maps had decreased outflow
- Future work should analyze *inflow* (i.e., daily number of arrivals per station) and compare to our results

References and Acknowledgements

T Amemiya (1985). *Advanced econometrics*. Harvard University Press.
E Fishman (2016). *Transport Reviews*, 36:92–113.
K Gebhard and RB Noland (2013). *Transporation Research Board* 92.
D Kahle and H Wickham (2013). *The R Journal* 5:144–161.
DJ Kim et al. (2012). *Transporation Research Board* 91.
P Leifeld (2013). *Journal of Statistical Software*, 55:1–24.
A Small (2017). The Four Horsemen of the Bike Share Apocalypse. *CityLab*.

Thanks to Elena Erosheva and Michael Karcher for helpful discussions and our STAT 504 class for their feedback.

Download this poster

Fork the source code

https://git.io/vyPjX

http://git.io/vyPjw