Predicting the Popularity of Bicycle Sharing Stations: An Accessibility-Based Approach

Matthew Wigginton Conway
UC Santa Barbara

California Geographical Society Annual Meeting
May 3, 2014
Los Angeles, CA

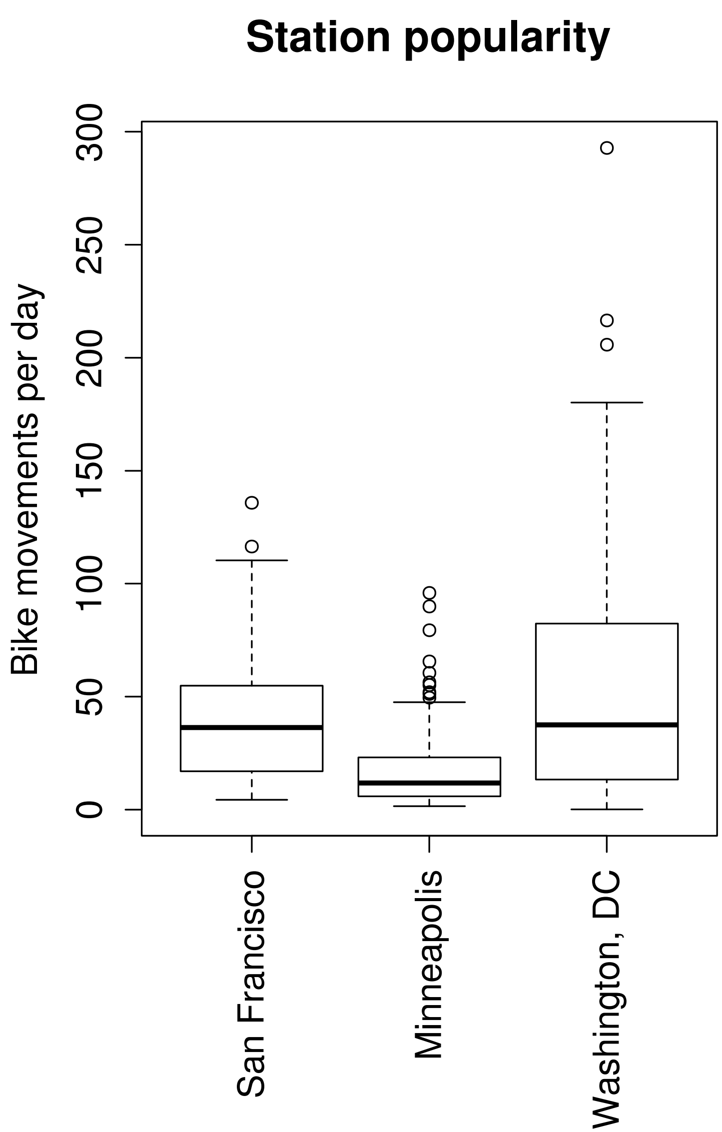
What is bikeshare?

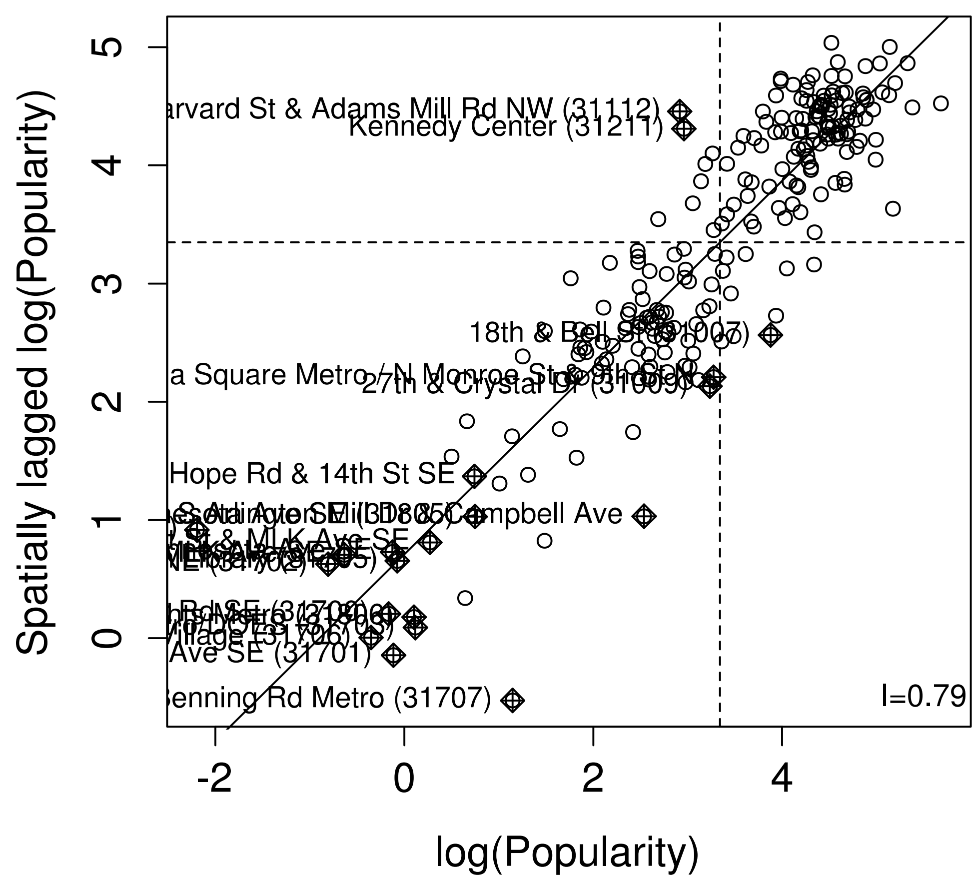
- Bicycles distributed throughout a city
- Electronic stations, automated rental
- Intended for short, point-to-point trips
- Memberships grant unlimited short trips

Bikeshare data

- Every time a bike is checked in or out, that is recorded
- Several bikeshare operators provide anonymized trip-level data to the public
- This analysis
  - Washington, DC (Capital Bikeshare)
  - Minneapolis/St. Paul (Nice Ride MN)
  - San Francisco Bay Area (Bay Area Bikeshare)

Patterns in station popularity





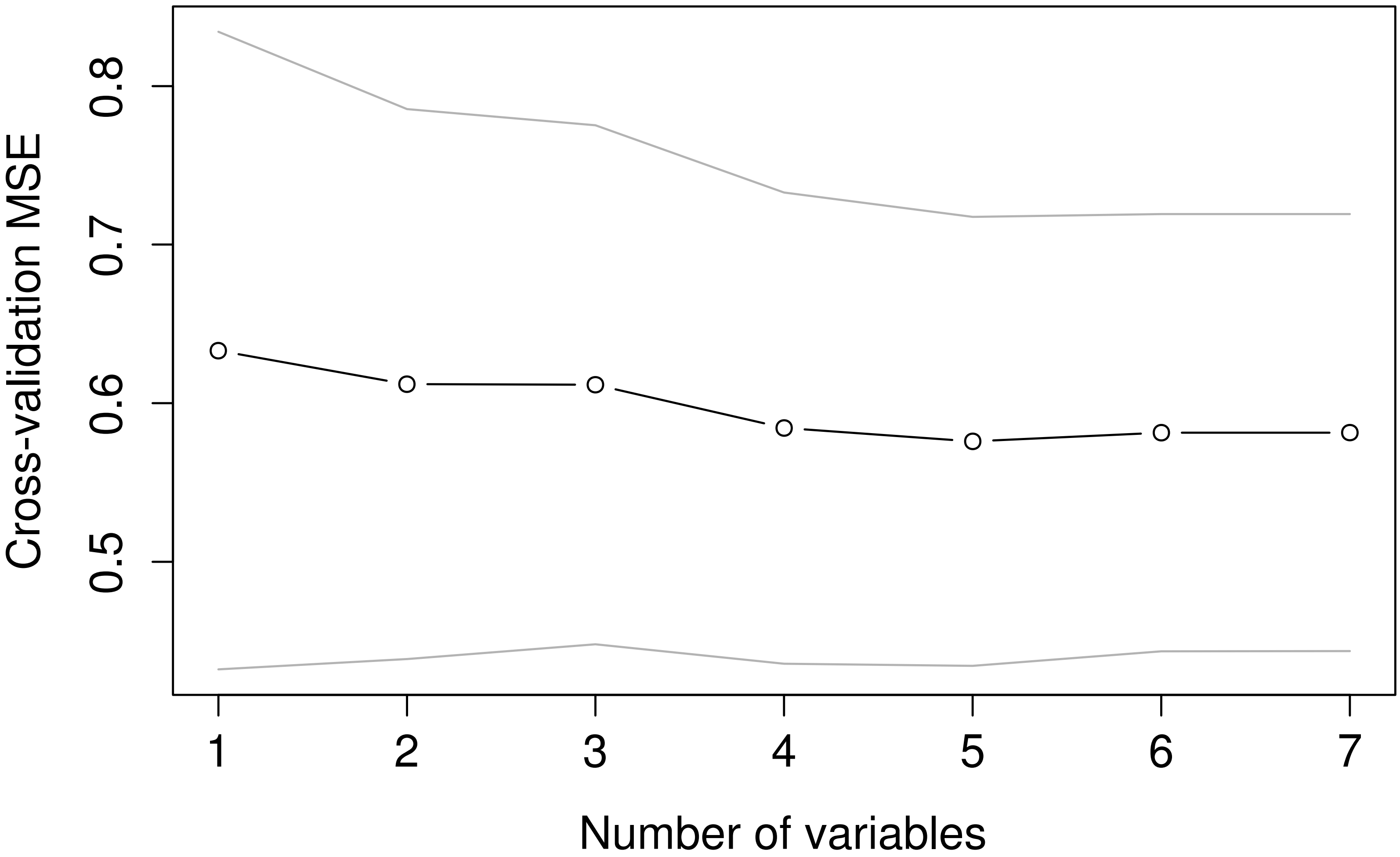
Modeling popularity

- Hypothesis: accessibility drives station popularity
- Independent variables
  - accessibility to jobs and residents by walking and transit
  - accessibility to other stations by bike
- Dependent variable: log(Popularity)
- Modeling philosophy: fit model in Washington, DC, and try to transfer to other cities

Data sources

- Station popularity can be extracted from trip data in Washington and Minneapolis, and from real-time availability data in San Francisco
- Block-level population and jobs data is available in the 2010 Census
- Street network data is available from OpenStreetMap
- Transit schedule data is available from transit providers
- Accessibility can be calculated using OpenTripPlanner

Linear regression



Linear regression

	Coefficients		Mean Squared Error		R^2		Moran's *I*	
Model	**Intercept**	**Predictor ☂**	**Cross-validation ☀‡**	**Test**	**Training**	**Test ♣**	**Response**	**Residuals**
Linear model (DC)	1.64	0.06	0.63	–	0.68	–	0.79	0.50
Direct transfer (MN)	1.64	0.06	–	0.61	–	0.31	0.69	0.55
Direct transfer (SF)	1.64	0.06	–	0.87	–	-0.15	0.49	0.53
Refit linear model (MN)	1.40	0.07	0.62	–	0.32	–	0.69	0.53
Refit linear model (SF)	2.65	0.03	0.54	–	0.33	–	0.49	0.23

not statistically significant ($\alpha = 0.05$)

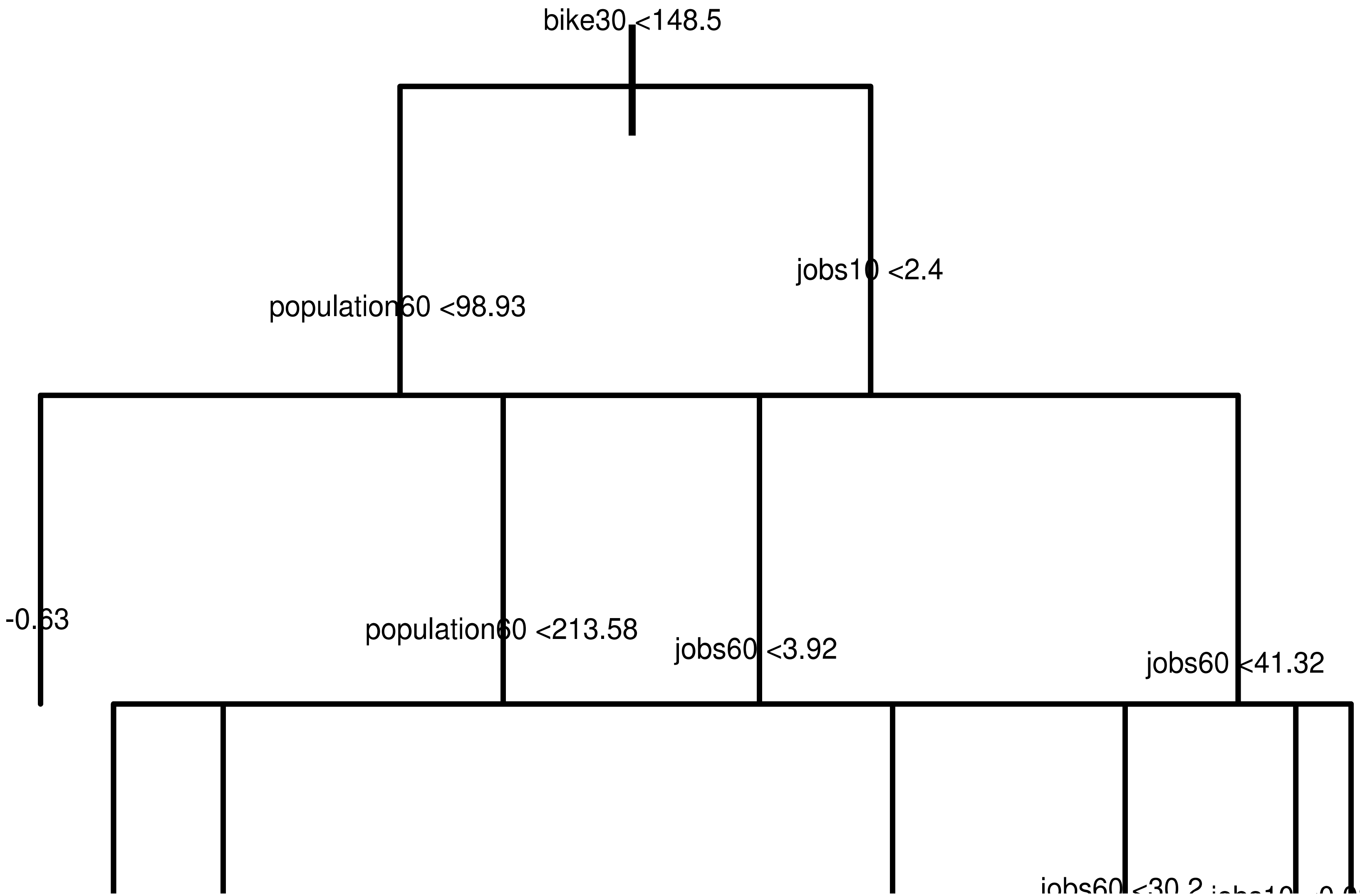
☀ 5-fold

‡ These models and measures are stochastic; parameters and values may vary slightly if refit, even with the same data.

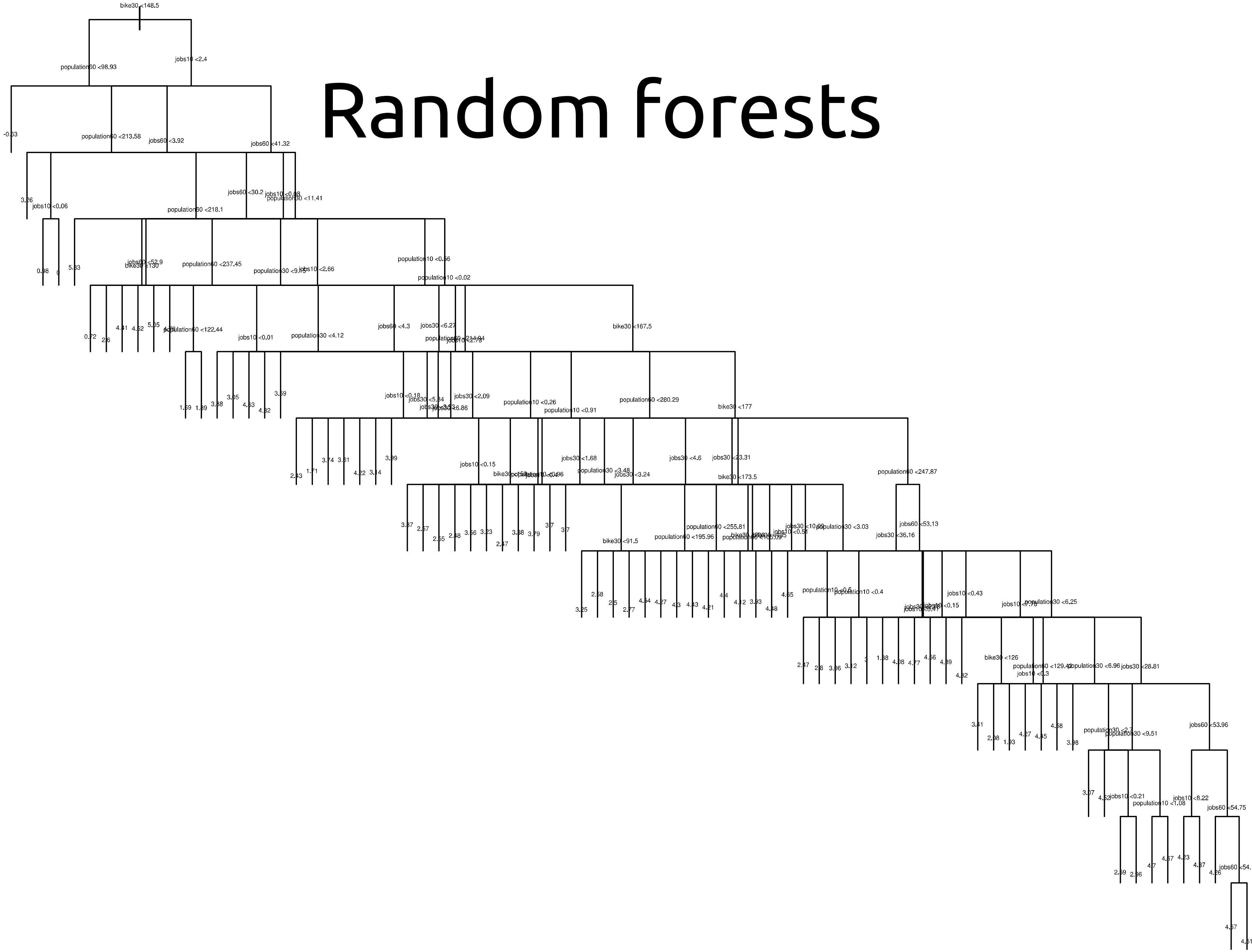
♣ Using test R^2 to evaluate the validity of transferred models is misleading, as it is based on the mean of the test observations. Thus it "sees" the test data, which the model did not see when trained.

☂ The predictor is jobs within 60 minutes by transit for the linear models and the exponentiated random forest prediction for the semilog-scaled models.

Random forests



Random forests



Random forests

Random forests

	Coefficients		Mean Squared Error		R^2		Moran's *I*	
Model	**Intercept**	**Predictor** ☂	**Cross-validation** ☀‡	**Test**	**Training**	**Test** ♣	**Response**	**Residuals**
Random forest model (DC) ‡	–	–	0.31	–	0.84	–	0.79	-0.02†
Direct transfer random forest (MN) ‡	–	–	–	0.99	–	-0.12	0.69	0.63
Direct transfer random forest (SF) ‡	–	–	–	0.61	–	0.19	0.49	0.27
Double-log-scaled random forest (MN) ‡	1.39	0.44	0.75	–	0.17	–	0.69	0.63
Double-log-scaled random forest (SF) ‡	1.23	0.68	0.52	–	0.34	–	0.49	0.20
Refit random forest (MN) ‡	–	–	0.47	–	0.47	–	0.69	0.30
Refit random forest (SF) ‡	–	–	0.50	–	0.31	–	0.49	0.06†

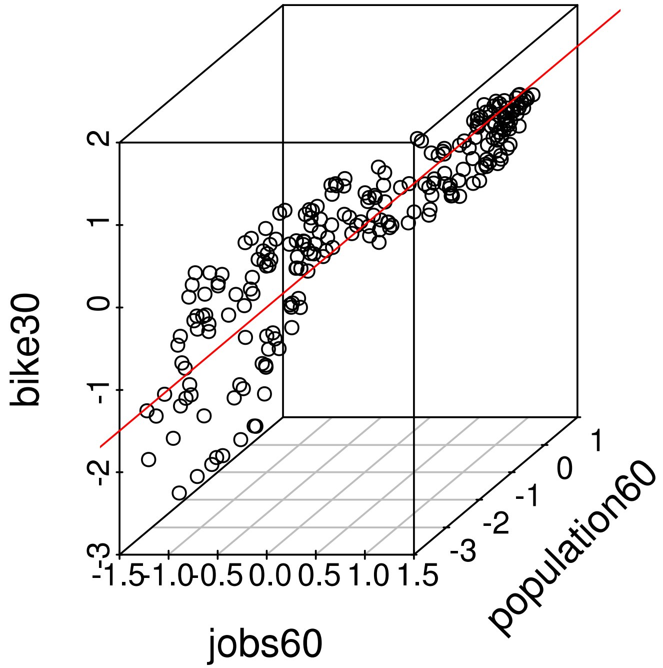
† not statistically significant ($\alpha = 0.05$)
☀ 5-fold
‡ These models and measures are stochastic; parameters and values may vary slightly if refit, even with the same data.
♣ Using test R^2 to evaluate the validity of transferred models is misleading, as it is based on the mean of the test observations. Thus it "sees" the test data, which the model did not see when trained.
☂ The predictor is jobs within 60 minutes by transit for the linear models and the random forest prediction in log units for the double-log-scaled models.

Discussion

- Accessibility is significantly correlated with bikeshare station use
- These accessibility-based models don't predict as well as might have been hoped
- Model transfer is inconsistent, suggesting city-specific factors (cf. Rixey 2013)

Further research

- Add more accessibility types (see Capital Bikeshare 2013, 23)
- Try additional statistical methods
  - Ridge regression
  - Principal components regression



Acknowledgements

- Kostas Goulias, UCSB
- Eric Fischer
- OpenTripPlanner Team
- Bikeshare operators and transit agencies
- Any errors that remain are my own

References

Capital Bikeshare. 2013. “2013 Capital Bikeshare Member Survey Report.” http://capitalbikeshare.com/assets/pdf/CABI-2013SurveyReport.pdf

Ligges, U. and Mächler, M. 2003. “Scatterplot3d - an R Package for Visualizing Multivariate Data.” *Journal of Statistical Software* 8(11), 1-20.

Rixey, R. Alexander. 2013. “Station-Level Forecasting of Bike Sharing Ridership: Station Network Effects in Three U.S. Systems.” http://docs.trb.org/prp/13-1862.pdf.

Questions, comments and contact

Matthew Wigginton Conway
University of California, Santa Barbara
www.indicatrix.org
matt@indicatrix.org