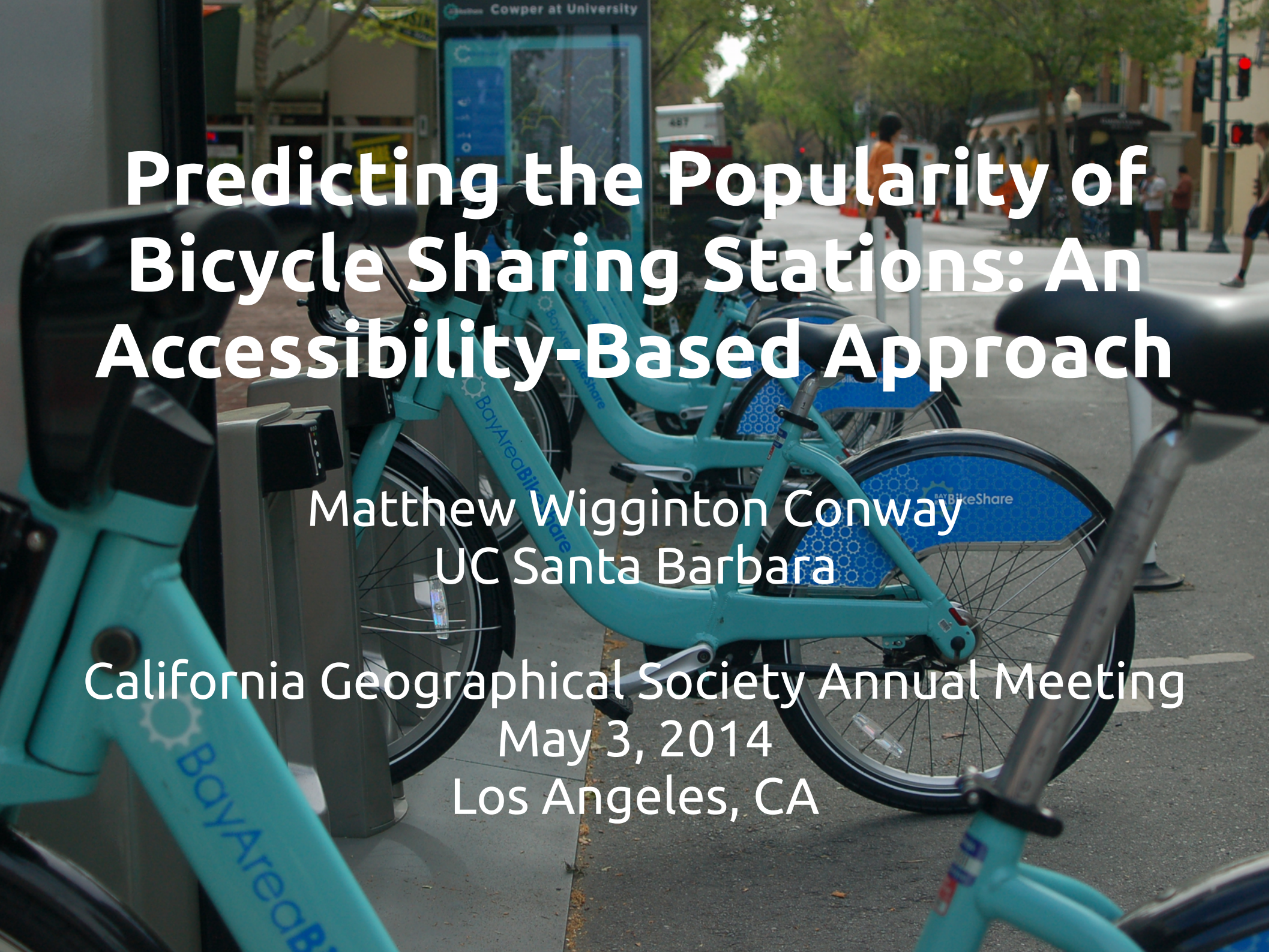


Predicting the Popularity of Bicycle Sharing Stations: An Accessibility-Based Approach

Matthew Wigginton Conway
UC Santa Barbara

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

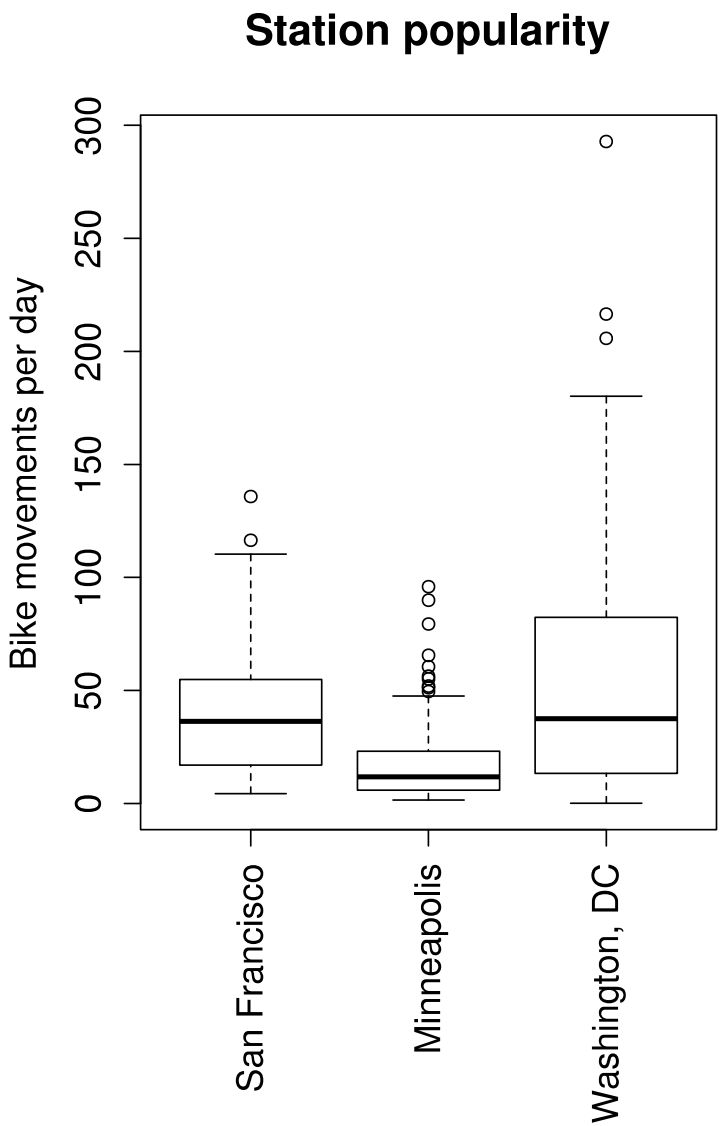
What is bikeshare?

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

Bikeshare data

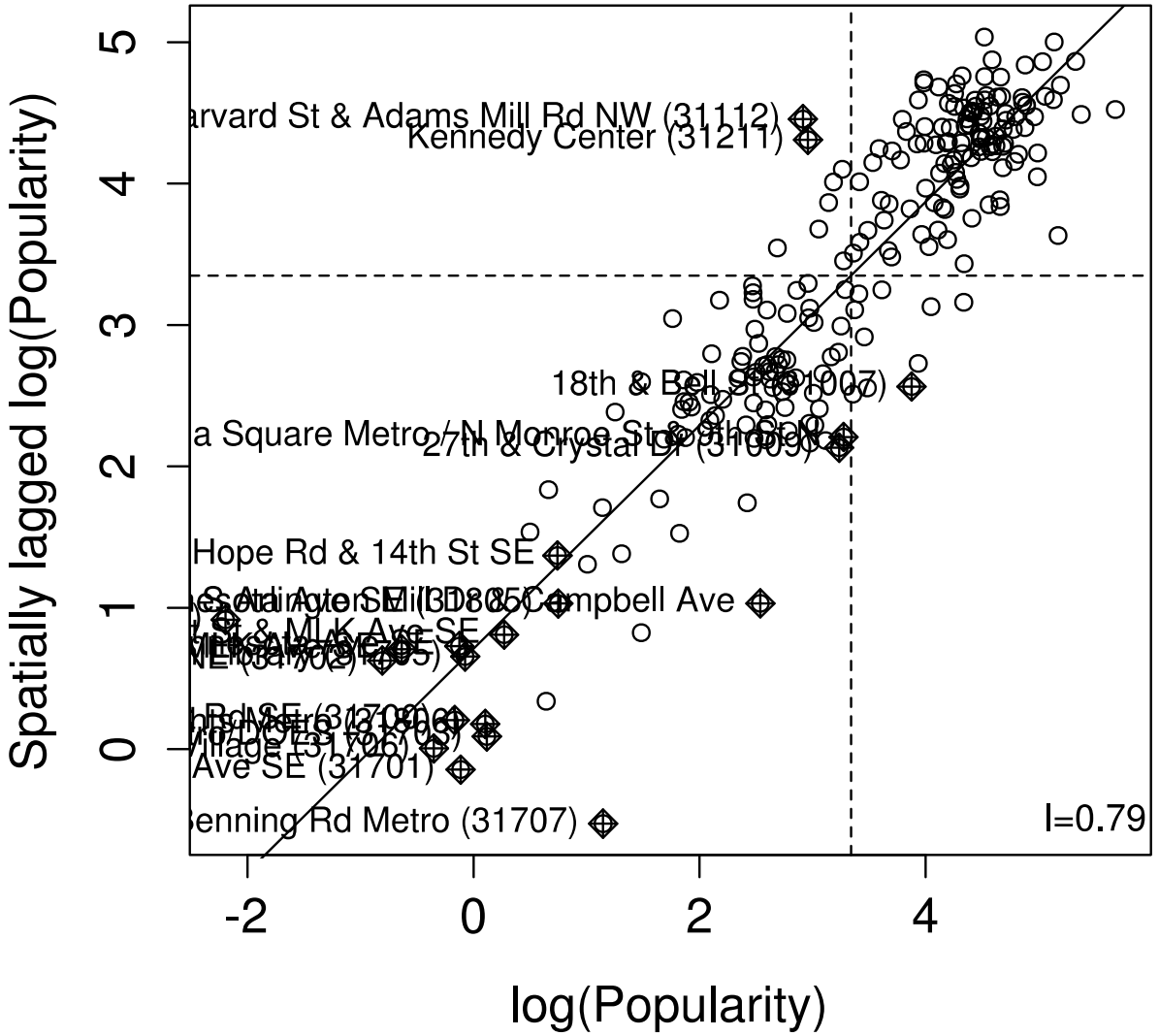
- Every time a bike is checked in or out, that is recorded
- Several bikeshare operators provide anonymized trip-level data to the public
- This analysis
  - Washington, DC (Capital Bikeshare)
  - Minneapolis/St. Paul (Nice Ride MN)
  - San Francisco Bay Area (Bay Area Bikeshare)

Variability in station popularity



- In each system studied, there is variation in station popularity
- Hypothesis: stations in more accessible locations are more popular
- Hypothesis: Models can be transferred from one city to another

Spatial autocorrelation of popularities



Modeling popularity

- Independent variables: accessibility to jobs and housing by walking and transit
- Dependent variable: popularity
- Modeling philosophy: fit model in Washington, DC, and try to transfer to other cities
- Use cross-validation to control overfitting
- Model metrics: Mean squared error, Moran's *I*

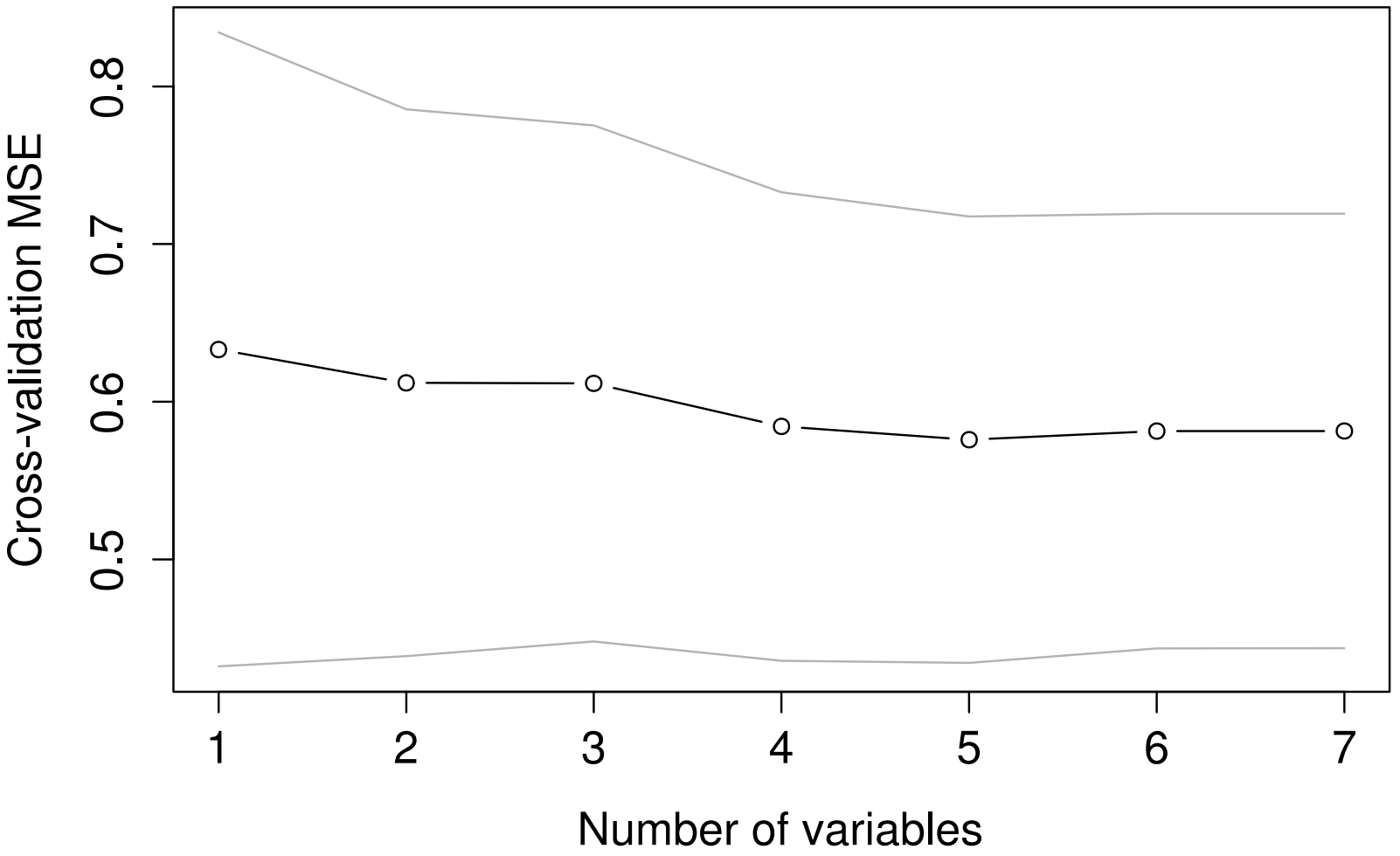
Data sources

- Station popularity can be extracted from trip data in Washington and Minneapolis, and from real-time availability data in San Francisco
- Block-level population data is available in the 2010 Census
- Jobs by Census block is available from the Census Longitudinal Employer-Household Dynamics Origin-Destination Employment Statistics (LODES)
- Street network data is available from OpenStreetMap
- Transit schedule data is available from transit providers

Accessibility calculation

- OpenTripPlanner is open-source network analysis software that can be used calculate the accessibility of a location
- For this project, accessibility to jobs and population were calculated for 10 minutes by walking and 30 and 60 minutes by transit
- A simple cumulative accessibility measure was used
- These accessibility measures were used as independent variables in regressions trying to explain station popularity

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

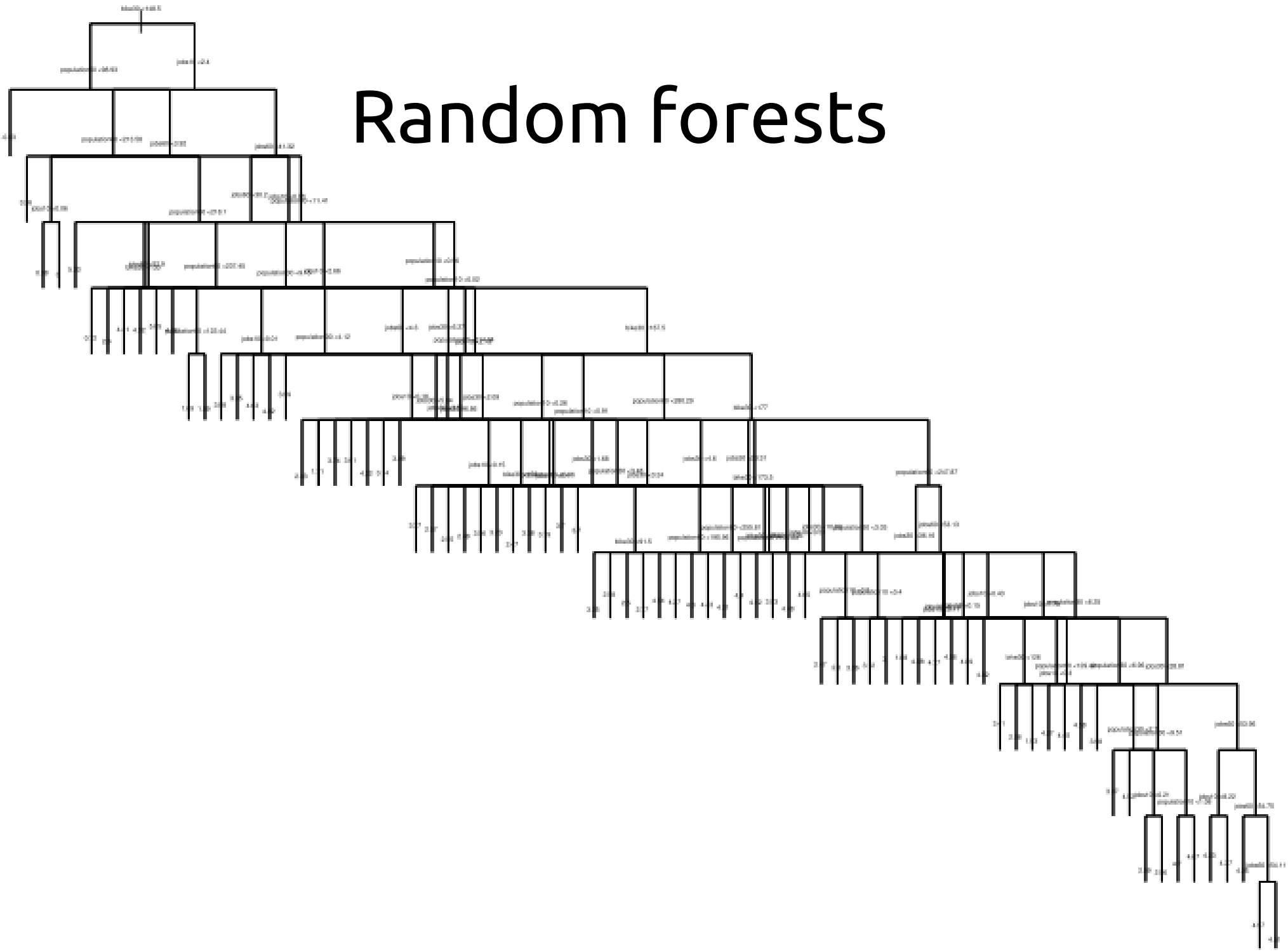
‡ These models and measures are stochastic; parameters and values may vary slightly if refit, even with the same data.

♣ 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
Semilog-scaled random forest (MN) ‡	1.49	0.07	0.60	–	0.33	–	0.69	0.54
Semilog-scaled random forest (SF) ‡	2.61	0.03	0.59	–	0.29	–	0.49	0.24
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 exponentiated random forest prediction for the semilog-scaled models.

Discussion

- Accessibility-based models don’t predict as well as might have been hoped
- Model transfer is inconsistent, suggesting city-specific factors (cf. Rixey 2013)
- Small number of stations in any city constrains results

Further research

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

Acknowledgements

- Kostas Goulias and Stuart Sweeney, UCSB
- Eric Fischer
- OpenTripPlanner Team
- Bikeshare operators and transit agencies

References

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

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