Predicting the Popularity of Bicycle Sharing Stations: An Accessibility-Based Approach Using Linear Regression and Random Forests

Matthew Wigginton Conway

www.indicatrix.org / matt@indicatrix.org

May 4, 2014

Abstract

Bikesharing systems have been growing quickly of late. They consist of stations distributed throughout a city where one can check out a bike for short trips. Some stations are more popular than others. This project develops statistical models to predict station popularity based on the accessibility of stations to jobs and housing. Linear regression and random forest regression are employed. The models are then transferred to other systems to evaluate their efficacy as planning tools for new systems. Models do not predict as well as might have been hoped, nor is there any one model that performs uniformly well in all cities. Spatial autocorrelation is present both in the popularities and the residuals from most of the models, indicating that spatial effects (such as accessibility) have not been fully explained.

Bikesharing systems have become popular of late, being installed in many cities. They consist of electronic stations distributed at regular intervals throughout a city. Members of the system can check out bikes at any station and return them to any other station. Some stations are, of course, more popular than others. One would expect stations with the highest accessibility to jobs and residents to be the most popular. This article explores whether measures of accessibility can be used to explain and predict bikeshare popularity and explain the autocorrelation seen in popularities. A model of bikeshare popularity is first developed in Washington, DC (Capital Bikeshare), and then transferred to Minneapolis (Nice Ride Minnesota) and the San Francisco Bay Area (Bay Area Bikeshare).[1] This

1 Throughout this article, San Francisco will refer to the entire Bay Area Bikeshare system, not only the stations located in the City of San Francisco. Minneapolis will refer to entire Nice Ride Minnesota system, and Washington, DC will refer to the entire Capital Bikeshare system.

article does find a significant connection between accessibility and station popularity, but falls short of finding a reliable general model for predicting bikeshare popularity in disparate cities.

Literature Review

Rixey (2013) has done a thorough analysis of the popularity of bikeshare stations, using multiple linear regression to explain station popularity based on a multitude of factors. His work did not focus exclusively on accessibility, although his use of buffers around stations to calculate variables such as the number of people and jobs nearby constitute accessibility measures (albeit not based on network distance). This article takes a slightly different approach to modeling bikesharing: instead of modeling all cities simultaneously, this paper builds models in one city and attempts to transfer them to other cities, to evaluate their efficacy as planning tools. This article also focuses on spatial autocorrelation as an indicator of model specification. Nevertheless, Rixey's article provides a valuable starting point for the discussion that follows.

Data Sources

Seven variables were calculated to use as input to the model: the number of jobs within 60 and 30 minutes by transit of each station (shortened to jobs60 and jobs30 in code and some figures), the resident population within 60 and 30 minutes by transit (population60 and population30), the resident population and number of jobs within 10 minutes by walking (population10 and jobs10), and the number of bikeshare stations within 30 minutes by cycling (bike30). Bikeshare is often combined with transit in multimodal trips (Capital Bikeshare 2013, 29), so it makes sense to use an accessibility measure that incorporates transit. The cumulative measure of bikeshare accessibility is calculated for 30-minute bike trips because all of the considered systems charge users an additional fee for trips longer than 30 minutes (Capital Bikeshare 2014; Bay Area Bike Share 2013; Nice Ride Minnesota

2014). Accessibility by transit was calculated at 8am on a weekday. It would be interesting to use accessibility measures at different times of day to evaluate the effect of off-peak accessibility on bikeshare use.

Data for the project come from many different sources. Population data is taken from the 2010 US Census TIGER/Line combined demographic and geography files, using the total population within each census block. Block-level employment data is taken from the US Census Longitudinal Employer-Household Dynamics Origin-Destination Employment Statistics (LODES).

Station popularity data come from the bikeshare system operators themselves. In Washington, DC and Minneapolis/St. Paul, the system operators provide data files with information about each trip, including the origin and destination. To calculate station popularity, I summed the number of trips that originated or terminated at a particular station and divided by the number of days that station has been or was operational.

In San Francisco, unfortunately, such fine-grained trip-level data was not available to the public at the time the analysis was conducted.[2] However, there is a real-time station information feed. By fetching this feed frequently for a long period of time, one can infer station popularity by seeing how many bikes have been taken from or deposited at a station. For this project, data from the real-time feed was fetched every minute, from August 29, 2013 until January 28, 2014, a period of five months. A script then tabulated the data, calculating popularities based on how many bike movements had occurred. There are a few problems with this method. If bikes both arrived and departed during a given minute, the net movement rather than the total movement was counted. Thus, popularity could be understated (especially for popular stations, where multiple bike movements per minute are more likely). This method also does not account for rebalancing; system operators use vans to move bikes

2 This data has since been released.

from crowded stations to less crowded ones. The trip-level data excludes rebalancing trips from station popularity measurements, but the approach taken in San Francisco counts rebalancing movements as part of the popularity. Bias from this will be most pronounced at stations where bike movements are very unbalanced (perhaps at particular times of day). This method has been used by researchers in the past, although they corrected for rebalancing movements (Dempsey et al., l. 81–156).

Accessibility indicators were computed using the open-source OpenTripPlanner multimodal trip planning framework (OpenTripPlanner Team 2014). This is still beta software, so some bugs may remain,[3] but results are believed to be sufficiently correct for this analysis. Street network data from OpenStreetMap was used to calculate walking distances. Transit schedule data was received in General Transit Feed Specification format from various transit agencies in each analysis area (see Table 1) and was used to calculate transit accessibility. OpenTripPlanner output data in CSV files, which were easily loaded into R (R Core Team 2013) for data analysis.

Region	Agencies
Washington, DC	Washington Metropolitan Area Transportation Authority Fairfax Connector Virginia Railway Express Maryland Transit Administration/MARC Arlington Transit
San Francisco	AC Transit Caltrain Bay Area Rapid Transit (BART) SamTrans San Francisco Muni Valley Transportation Authority (VTA)
Minneapolis/St. Paul	Metro Transit

Table 1: Transit agencies used for analysis of accessibility by transit

Modeling Methodology

The goal of the project is to develop a model in Washington, DC, that not only predicts the

3 https://groups.google.com/d/msg/opentripplanner-users/bWi2XyegAvA/333Q3_q-tv0J

popularities of stations in Washington, but also can be transferred to other cities. With that in mind, the model was trained only on data from Washington, in order to get a better picture of how well it would transfer to a new city. Another goal was to try to explain the autocorrelation in the data; nearby stations tend to have similar popularities (Conway 2013, 5). Autocorrelation, it is hypothesized, results from wanting to be near some urban amenity. Ergo, accessibility should be able to encapsulate the autocorrelation, leaving us with residuals that are not spatially autocorrelated.

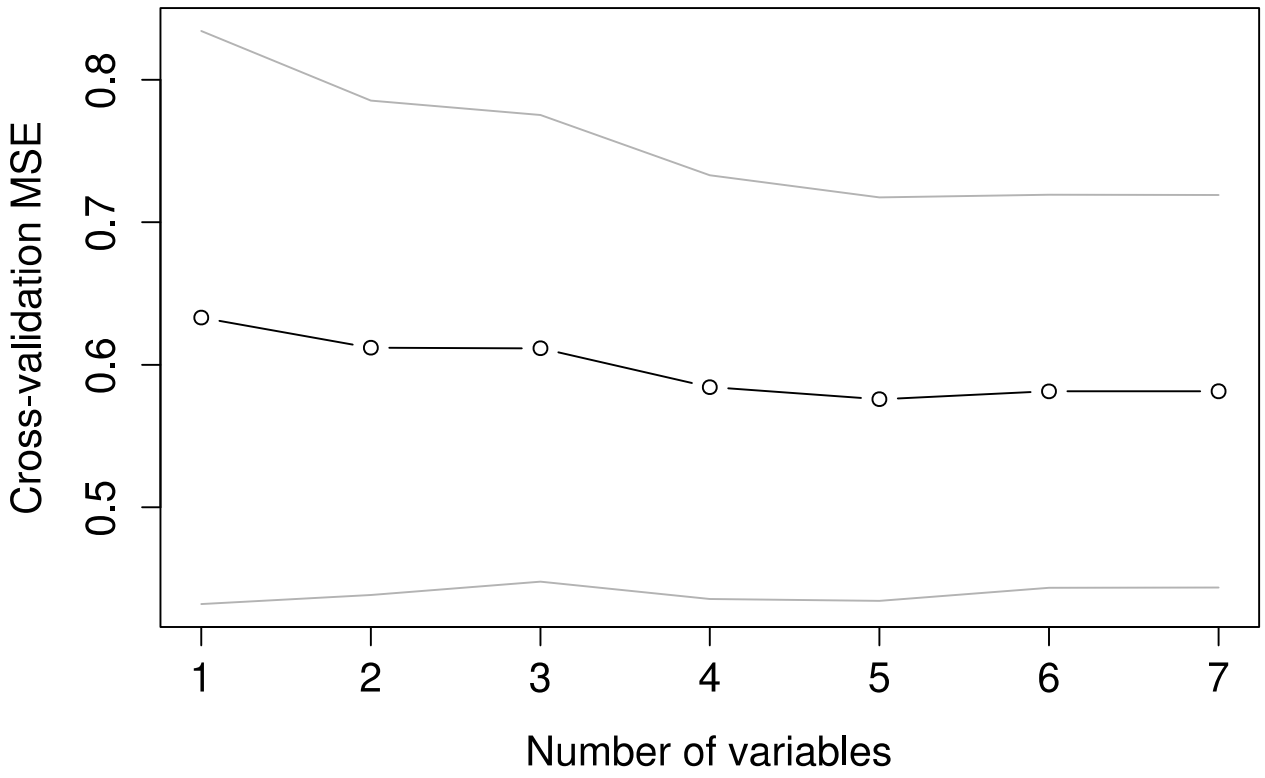

Figure 1: Cross-validation MSE for linear model with different numbers of predictors and one standard deviation lines

Initially, linear regression was used to fit a model to the Washington dataset. Best-subset selection was used to determine what the best combination of variables was, and cross-validation was

used to pick a number of variables that created an accurate yet parsimonious model. Generally, when using cross-validation to choose a number of predictors, the most parsimonious model with MSE less than one standard error more than the minimum is chosen (Hastie, Tibshirani, and Friedman 2009, 244). In this case, the model with only one variable is chosen (see Figure 1). The constructed linear model is predicting the natural log of station popularity, rather than the raw popularity, to reduce heteroskedasticity. When the model is fit with the raw popularity, there is a noticeable funnel shape in the residuals, with larger residuals for more popular stations. This makes sense. A 10% error at a station with 10 bike movements per day is only 1 bike movement, while that same error at a station with 100 bike movements per day is 10 bike movements. Taking a log of the response solves the problem (Figure 2).[4] Rixey (2013, 4) also used the natural log of popularity in his analysis.

4 In a previous project, I used a Box-Cox transformation to normalize the station popularities (Conway 2013, 5). A log was used instead here because Box-Cox requires fitting a parameter, making the method more flexible. Since this model is intended for transfer, excess flexibility is not desirable.

	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
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.

Table 2: Summary of all models fit for evaluating bikeshare use

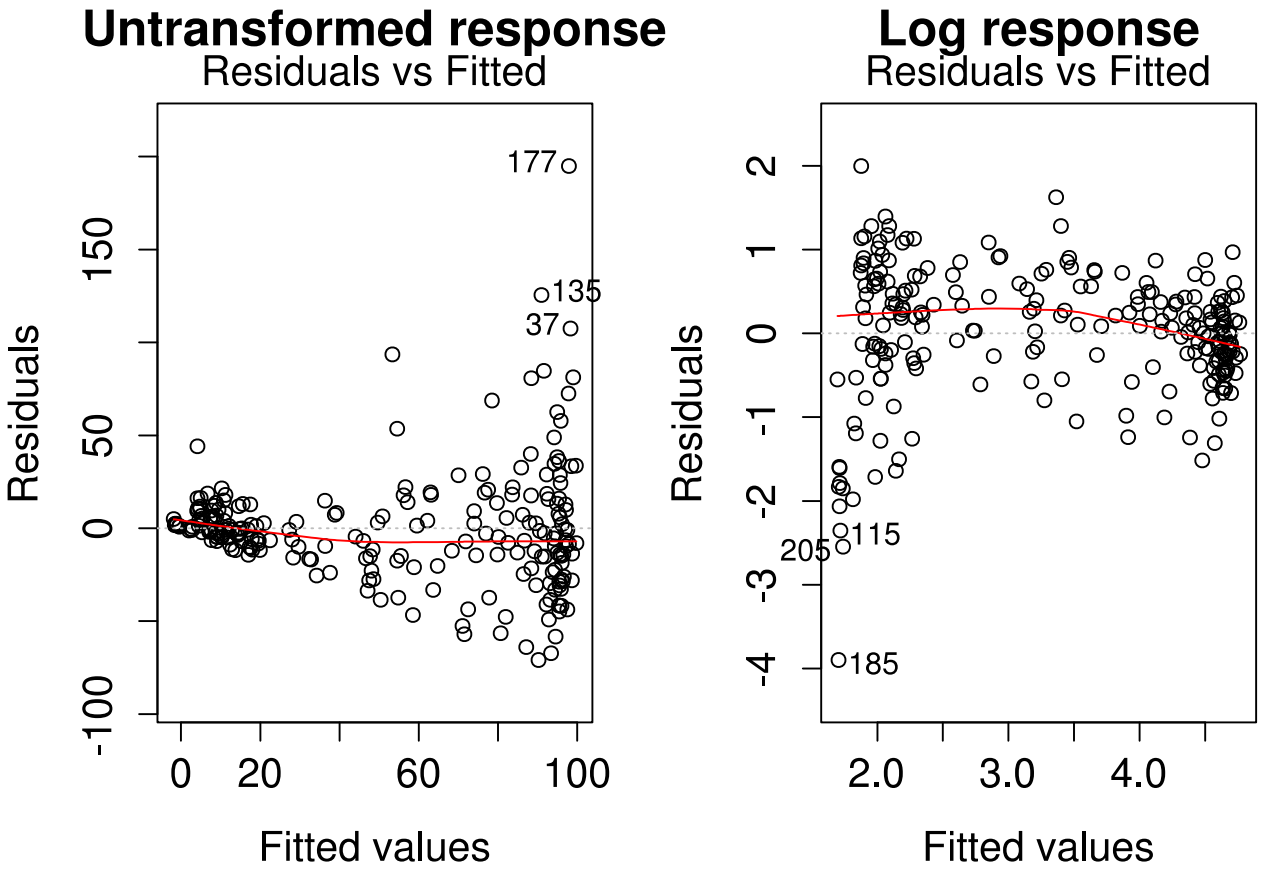

Figure 2: Log-transforming the response. Note the funnel shape in the left-hand plot, indicating heteroskedasticity, and the more uniform shape of the right-hand plot.

Initially, a semilog linear model was fit regressing log(Popularity) onto jobs within 60 minutes by walking and transit, as suggested by best subset selection. The model does a fairly good job at explaining the variation in the station popularities in Washington, DC, with cross-validation MSE of 0.63 (in log-transformed units) and a training R^2 of 0.68. It does show a significant connection between accessibility and station popularity. The intercept is 1.64 and the coefficient 0.06 (all of the accessibility measures are expressed in units of 10,000 except for bikeshare stations within 30 minutes, so jobs within 60 minutes is actually tens of thousands of jobs within 60 minutes). Both the intercept and the coefficient are significant.[5]

5 Though it may initially seem that this is not sampling and thus these types of things do not apply, it is important to

Unfortunately, the residuals are still strongly autocorrelated, with Moran's *I* equal to 0.50. Moran's *I* is 0.79 for log(Popularity), so it has been reduced but not as much as one would hope. *I* values greater than 0.3 suggest strong spatial autocorrelation (O'Sullivan and Unwin 2010, 206). Parameters and statistics for all fitted models are shown in Table 2.

This model is not completely satisfactory for a number of reasons. For one, the autocorrelation of the residuals is problematic. It also would make intuitive sense for popularity to depend on the other accessibility measures. One reason potentially important variables may be excluded from by best-subset selection is that all of the predictors are highly correlated (Figure 3). Random forests are an alternate regression method that can be useful for data with a number of highly correlated predictors (James et al. 2013, 320).

remember that, while data from all of the stations was used, this model is intended for prediction, and thus the true population is all the stations that could ever exist; the stations that currently exist are a sample from this (and not necessarily a random sample).

Figure 3: Correlations between predictors and response, Washington, DC.

Random forests are built by fitting hundreds of decision trees; however, at each split of each tree, a random sample of the predictors are drawn and used to make the split. The results of all of the fitted trees are then averaged to reach a final prediction. This means that the model works particularly well when variables are highly correlated; since most variables are excluded from any one particular split, the effects of slightly weaker predictors are not drowned out by the effects of very strong predictors (James et al. 2013, 320).

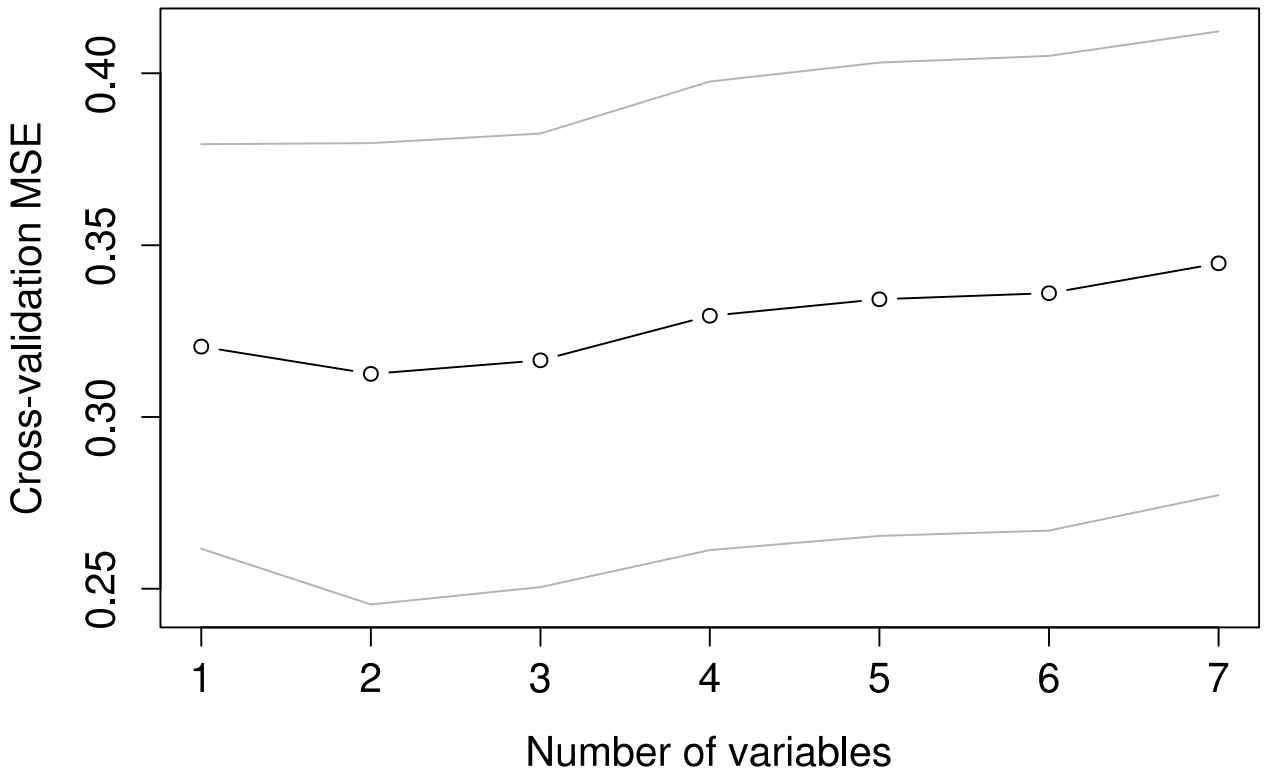

Figure 4: Cross-validation MSE for random forest explaining log(Popularity) based on accessibility measures, with one standard deviation lines

Fitting a random forest requires determining how many variables should be used at each split. This can be considered a tuning parameter (Hastie, Tibshirani, and Friedman 2009, 592), so we fit it using cross-validation (5-fold, in this case). From the graph of the values of cross-validation MSE with different numbers of variables, we see that they are all roughly equivalent (Figure 4). The creators of the random forest method suggest selecting the number of predictors divided by 3 (Hastie, Tibshirani, and Friedman 2009, 592); as there is no evidence to suggest this is not a good choice, we use 2 variables at each split. We continue to use log(Popularity) as the response variable. Though random forests can handle nonlinearity, there is still the problem of heteroskedasticity; it's simply possible to be more wrong (in absolute terms) when popularities are high. The random forest has cross-validation

MSE of 0.31 (considerably lower than that of the linear model) and training R^2 of 0.84. All variables contributed to the model, although jobs within 60 minutes by transit contributed the most (Figure 5). This is consistent with the results of best-subset selection on the linear model. The spatial autocorrelation in the residuals has also disappeared. This model is much more satisfactory than the linear model, presumably because it is able to capture more of the structure in the accessibility data.

Figure 5: A plot of the relative importance of different variables to the random forest fit. Longer bars indicate that variables contributed more to the final random forest.

Transferring the Models

The models do a fair job of predicting the popularity of bike-sharing stations in Washington, DC. However, the model would be far more useful as a planning tool if it worked in other cities as well. Bikeshare station data is available for Minneapolis's Nice Ride system and the San Francisco Bay Area's newly-launched Bay Area Bike Share system. These data allow us to test transferring the model.

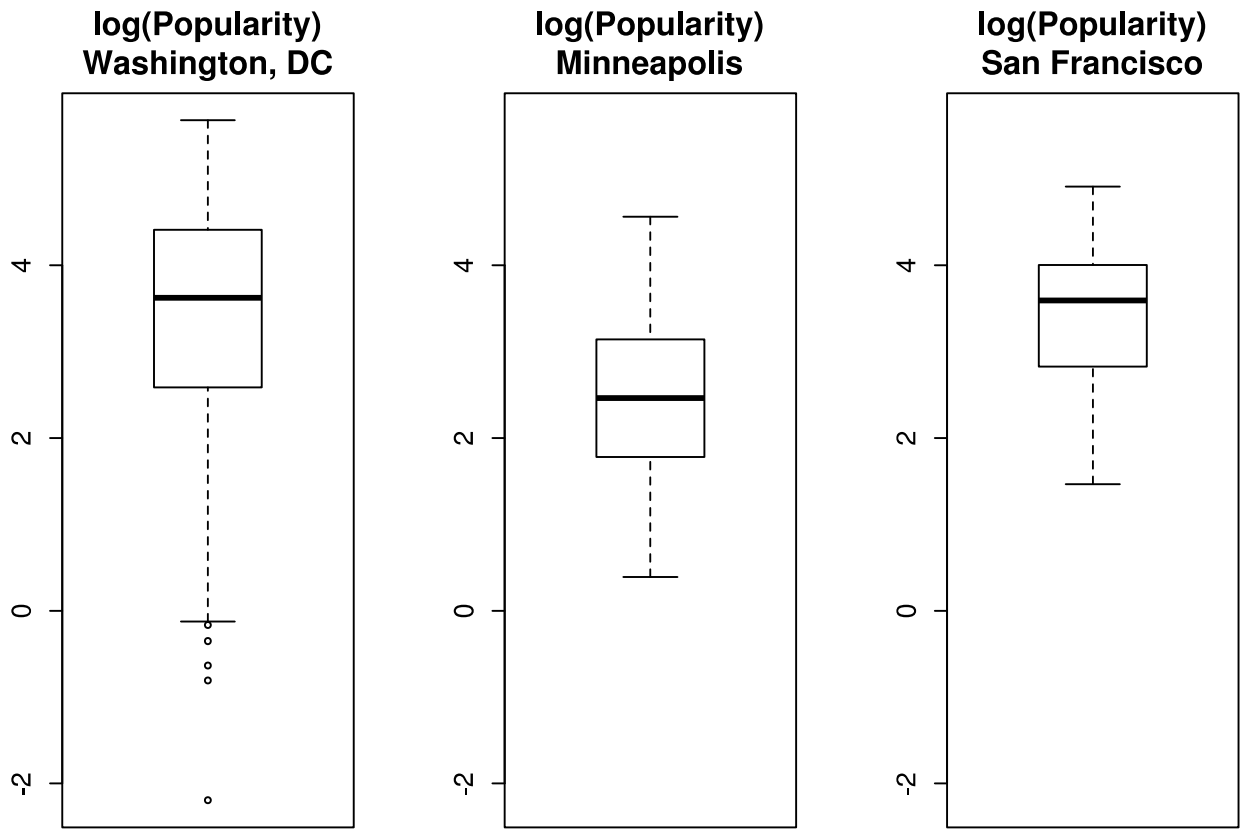

Figure 6: Boxplots of the log-transformed popularities of stations in each of the project areas

First, the models developed in Washington, DC were directly transferred to the other cities (that is, both the form of the models and the fitted parameters and coefficients were transferred). The performance of the linear model in Minneapolis is comparable to its performance in Washington, DC; both have MSE of approximately 0.6.[6] The R^2 in Minneapolis is much lower than the R^2 in Washington; there are two reasons for this discrepancy. One is that the R^2 in Washington is a training R^2, while the R^2 in Minneapolis is a test R^2; training R^2 tends to be larger because the model is specifically fitted to reproduce that data (although this effect should be controlled by using cross-validation; the model should not be overfit). The other issue comes from the definition of R^2: it is the ratio of the explained

6 It should be noted that this is 5-fold cross-validation MSE in Washington, DC and test MSE in Minneapolis and San Francisco; the 5-fold cross-validation is subject to bias and may thus be slightly overestimated.

variation to the total variation. There is less variation in the popularity of stations in Minneapolis (Figure 6), so the R^2 will be lower even when the MSE is similar. It is also a bit misleading to use test R^2 to evaluate the effectiveness of a transferred model, as it is not directly comparable to the training R^2 used to evaluate the other models. In San Francisco, test MSE is higher than it is in either Washington or Minneapolis; test R^2 is actually *negative*, indicating that the model predicts worse than simply predicting the mean popularity for all the stations. That is, there is more variation in the residuals than there was in the original data.

It is perhaps overly optimistic to expect the fitted parameters as well as the model form to transfer easily from one city to another city in a different context. Instead, we can take the model forms fit in DC and refit them using data from another city. For the linear models, we refit a linear regression of log(Popularity) onto jobs within 60 minutes by walking and transit. In Minneapolis, this results in no noticeable change in MSE over the directly transferred model (the coefficients are also very similar). In San Francisco, however, this reduces MSE by approximately 0.3.[7] The intercept is larger and the coefficient on jobs smaller. This is because there are fewer low-popularity stations in the Bay Area Bikeshare network (Figure 6), so the entire model is shifted up and flattened. In the refit models, MSE is similar to the MSE of the model fit in Washington, DC, but R^2 is much lower. Again, this is because there is less variation to explain, in both San Francisco and Minneapolis. The coefficients and the intercepts remain significant.

We can transfer the random forest model, preserving the structure and relationships between the

7 It should be noted that cross-validation MSE is computed slightly differently for the refit linear model compared to the original Washington, DC model. In the Washington model, variable selection was performed inside the cross-validation, so each fold could potentially have a different set of best variables; otherwise the result will be biased as the predictors were chosen on the basis of the test data (Hastie, Tibshirani, and Friedman 2009, 245–247). For the refit linear models, the subset selection was performed on the basis of an entirely different dataset (the Washington dataset), so subset selection need not be performed in the cross-validation, as the test data is already separate from the training data used to screen predictors.

variables but allowing for variations in the magnitude of the predictions, by fitting a linear model of the observed popularities against the popularities predicted by the transferred random forest model. This is effectively a log-log model, with the log of popularity on both sides. In both cities, the coefficient on the random forest prediction is less than one, indicating that there is less variation in the popularity. These models again don't predict particularly well. It would be more interpretable to regress the observed popularities against predicted popularities, without the log transformations, but this creates heteroskedasticity. There exist other solutions to this problem, such as weighted least squares, but they are beyond the scope of this paper. The residuals of these scaled models are still spatially autocorrelated.

The log-scaled random forest models allow for variation in the magnitude of the effect of accessibility measure on bikeshare use, but they constrain the *relative* importance and the interactions of each of the accessibility measures to be the same as in Washington, DC. We can also refit the random forest entirely, allowing the relationships between the variables to change. We are now transferring only the (relatively successful) modeling technique, and allowing everything else to be refit. When we do this, cross-validation MSE is down somewhat from the semilog-scaled models, but not a huge amount. This is unsurprising; when we fit to the data, as opposed to fitting to other data from a different city and transferring the model, MSE should go down. What is most interesting is that residual spatial autocorrelation is down; it is no longer significantly different from 0 (not autocorrelated) in San Francisco, and is 0.3, right on the edge of being considered strong, in Minneapolis. Fitting a random forest to local data yields the lowest residual spatial autocorrelation in all cities.

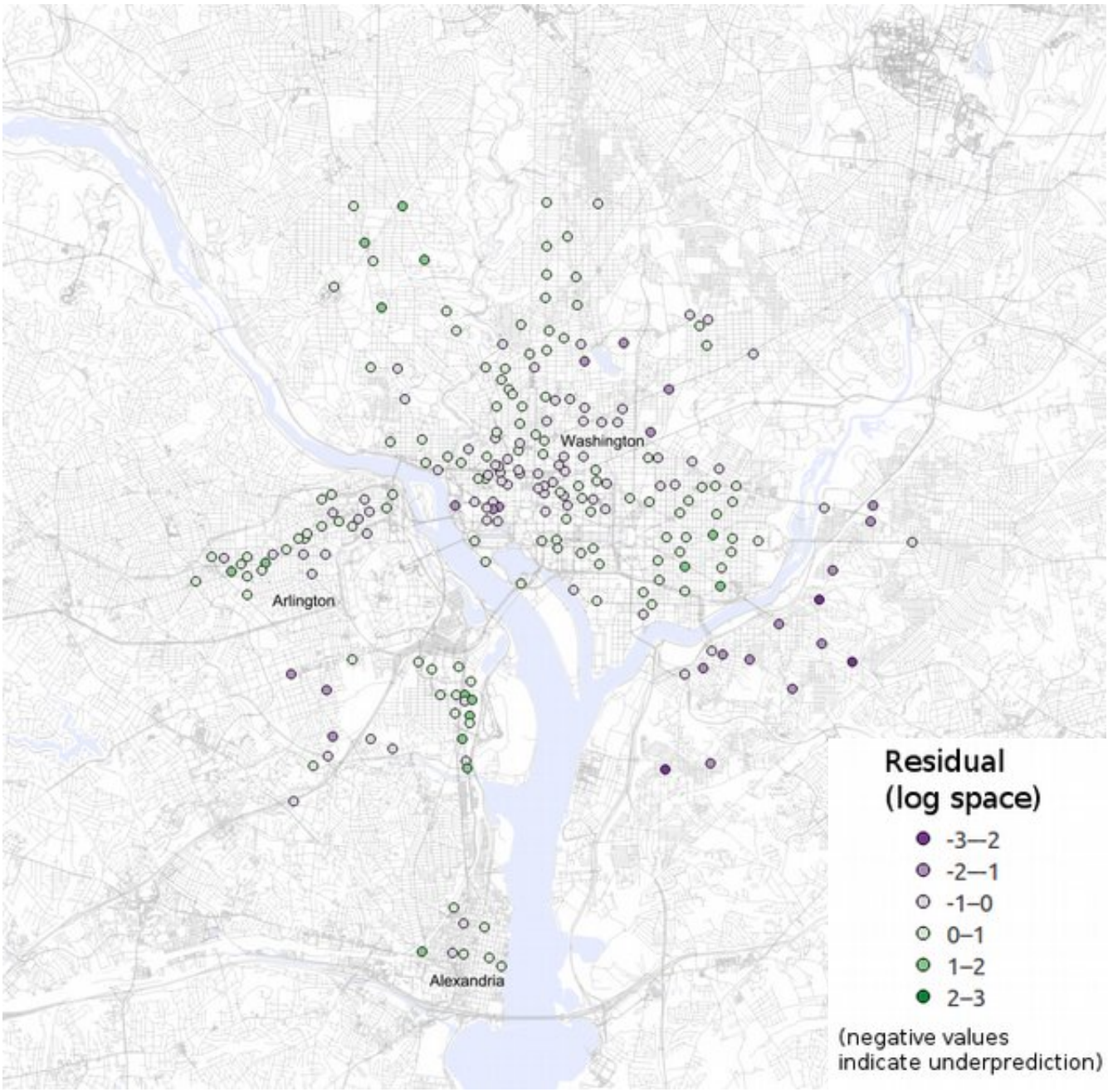

Figure 7: Residuals from the linear model fit in Washington, DC

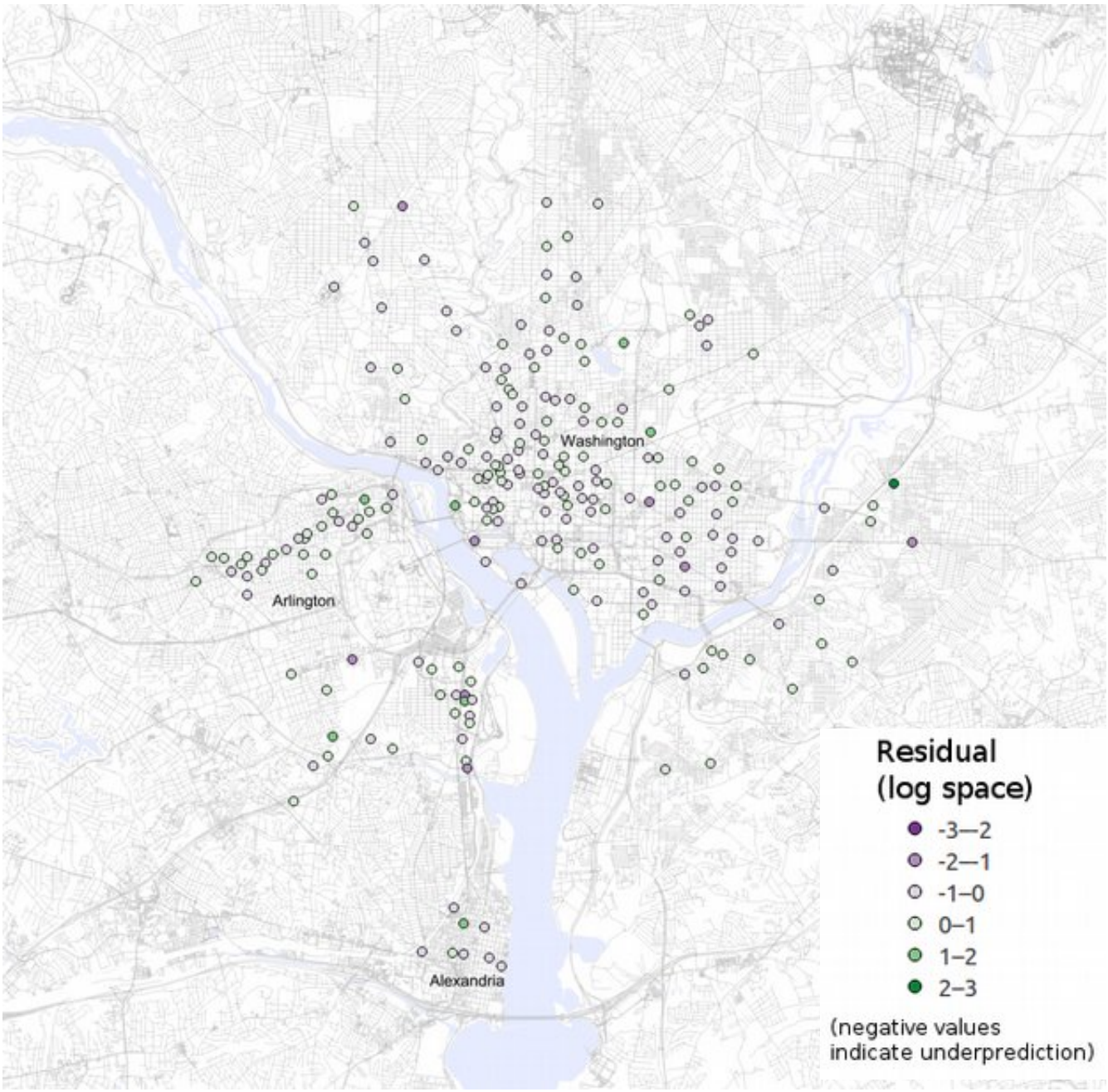

Figure 8: Residuals from the random forest model fit in Washington, DC

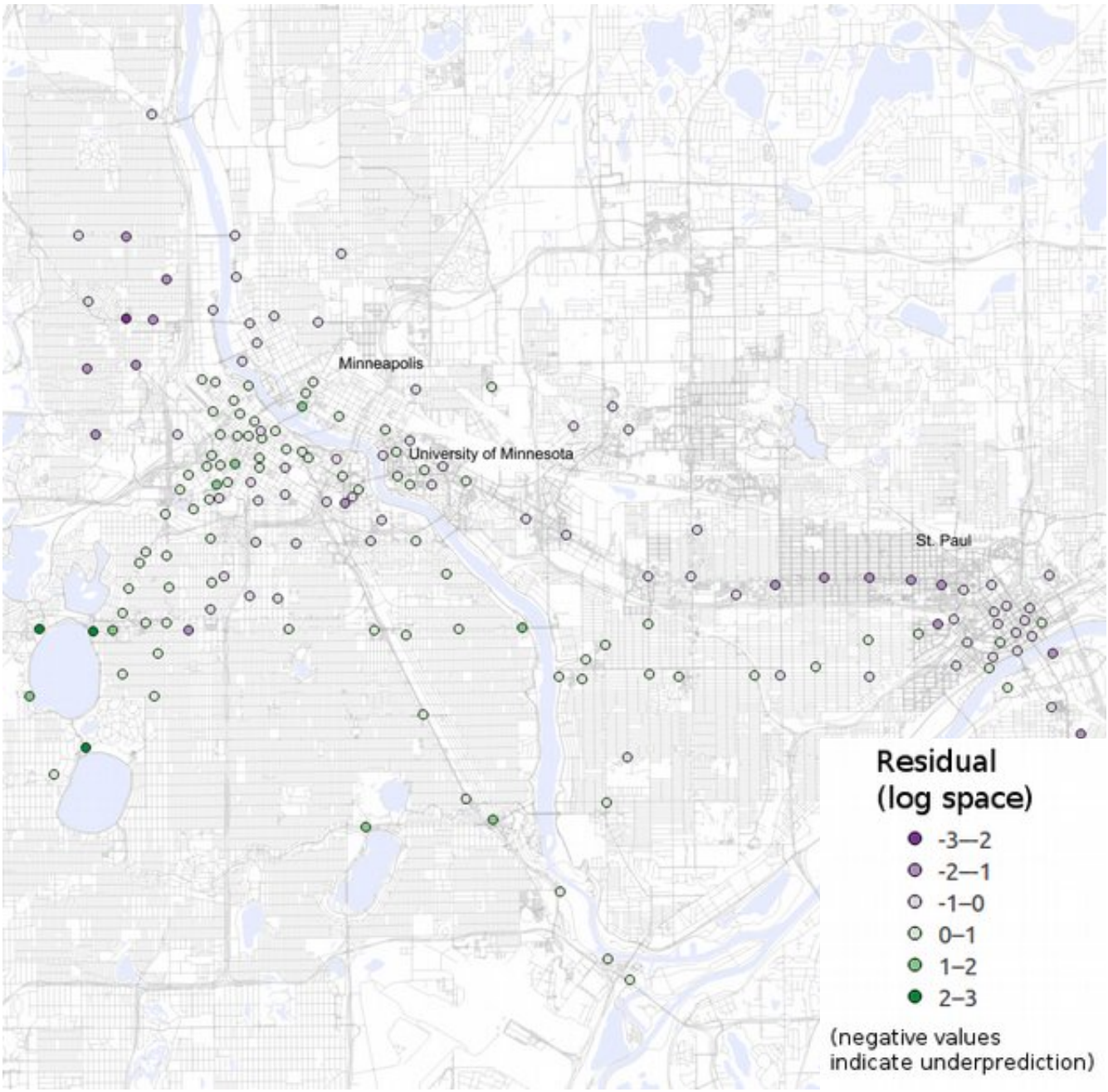

Figure 9: Residuals from the refit linear model in Minneapolis

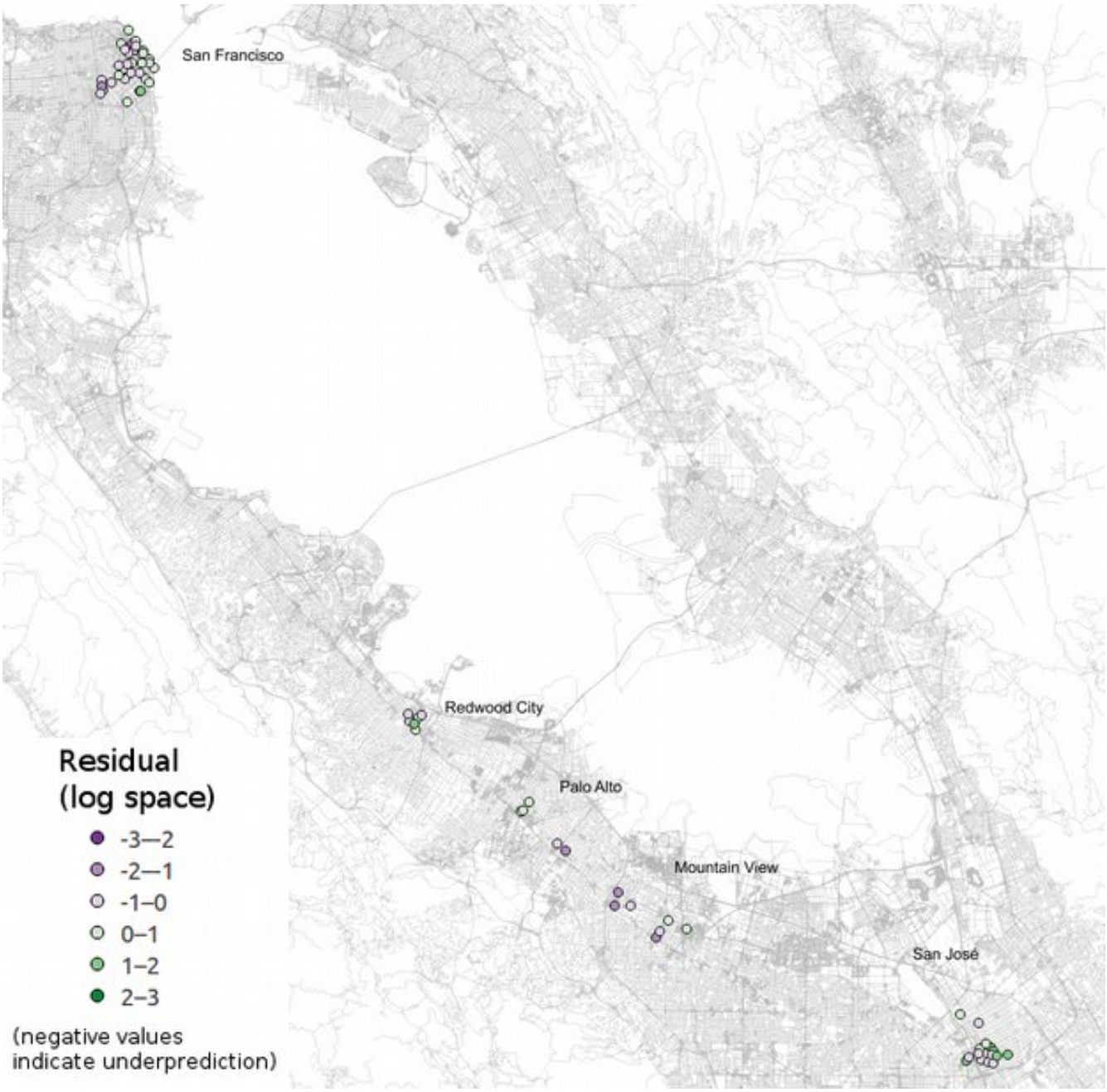

Figure 10: Residuals from the refit linear model in San Francisco

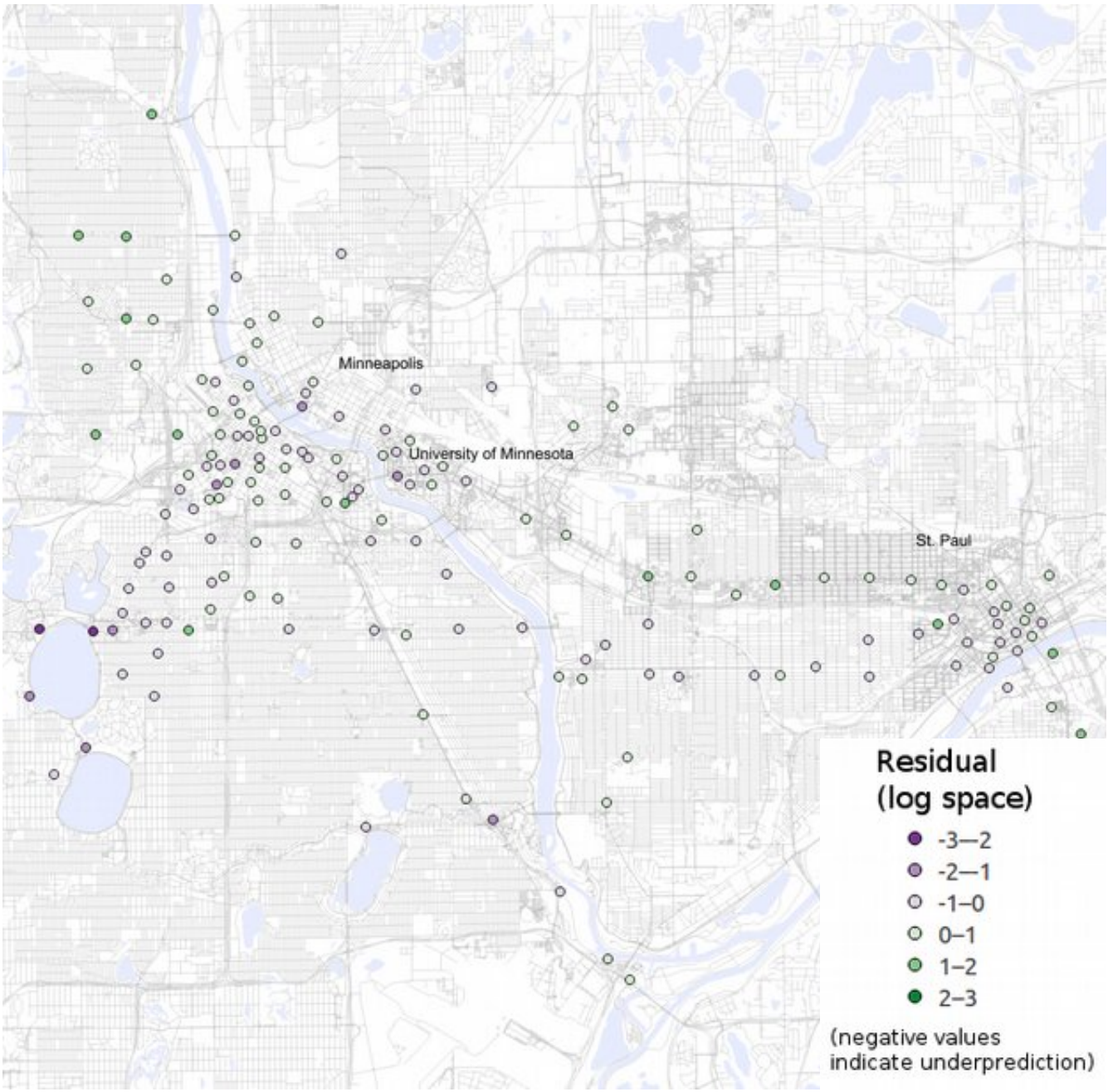

Figure 11: Residuals from the refit random forest model, Minneapolis—St. Paul

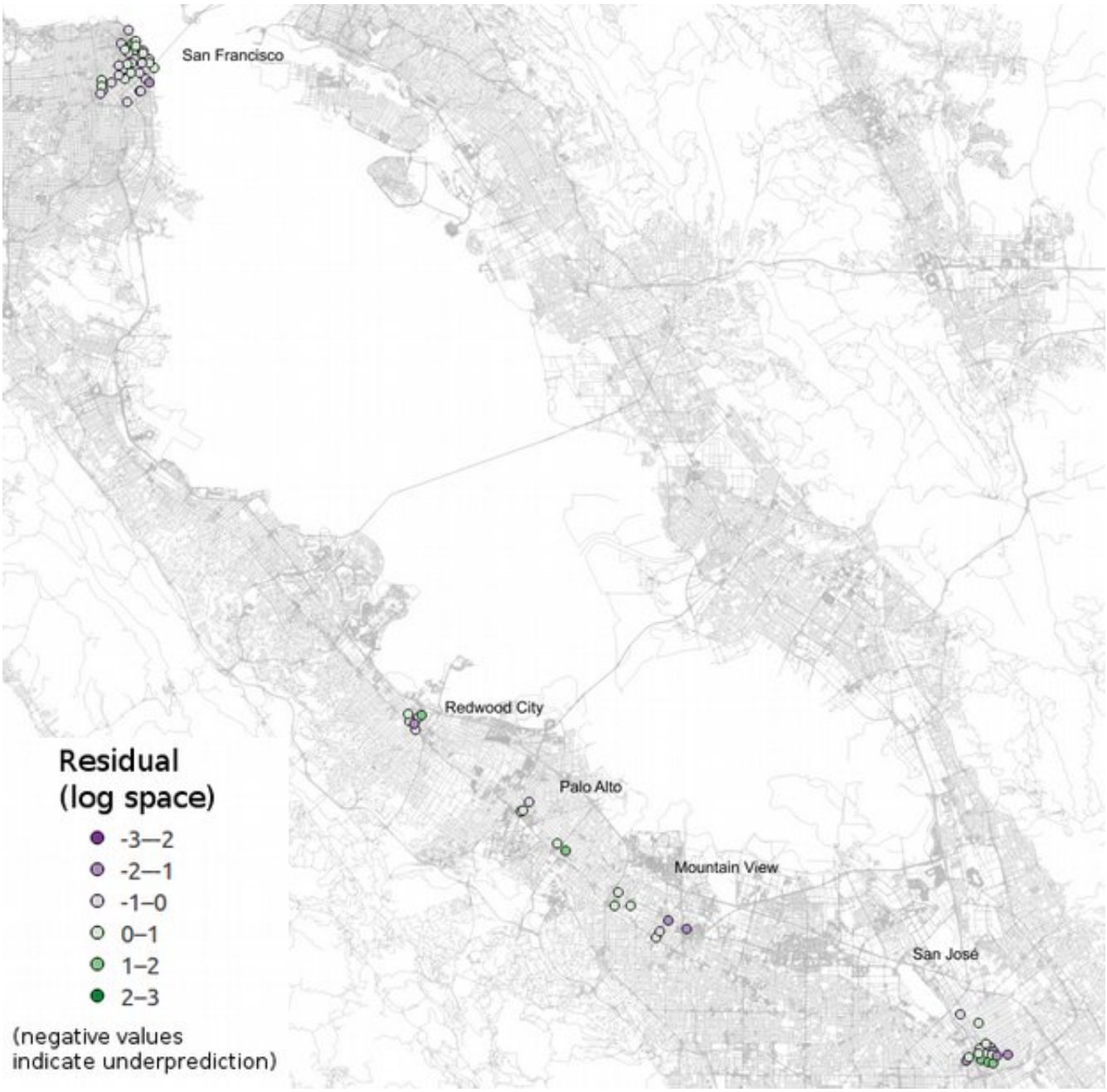

Figure 12: Residuals from the refit random forest in San Francisco

Discussion and Further Research

The results of this study are decidedly mixed. There is a significant link between accessibility and bikeshare station popularity. However, while the models predict defensibly well in in the city in which they are fit (even under cross-validation, suggesting that they should predict new stations with similar efficacy), the models do not transfer very well at all. The models transfer fairly well when refit, but if the models are to be used as planning tools to assess potential new bikesharing systems, they need to transfer well without refitting parameters.

There is still spatial autocorrelation in the residuals of most models, indicating that the models are not fully explaining systematic spatial effects. The random forest models have less autocorrelation and use more of the variables. This is promising; adding additional accessibility measures explains more of the systematic variation in the data.

Plotting the residuals on a map is a valuable way to assess the autocorrelation. In Washington, DC, we see that the linear model underpredicts in the southeast part of the city, as well as the western part of the core (Figure 7). Recall that the residuals of the random forest model show no significant autocorrelation (Figure 8). The obvious explanation of the reduced autocorrelation is simply that the random forest uses more predictors; if, as hypothesized, multiple accessibility measures do feed into driving station popularity, the residual autocorrelation in the linear model could be because it is using only one variable. The remaining autocorrelation can be explained using more variables. It would be interesting to fit a linear regression with more variables and see how it performs in comparison to the random forests. It is also possible that the unique combination of different accessibility measures in these parts of the city allows the random forest to overfit, by having trees with leaves specific to a given area.

In Minneapolis—St. Paul, the refit linear model underpredicts in St. Paul and overpredicts in Minneapolis (Figure 9). The refit random forest does better (Figure 11), although residuals are still autocorrelated. It has been argued that lower-than-expected bikeshare use in St. Paul is due to a lack of bicycle infrastructure (Lindeke 2014). This could explain the residual spatial autocorrelation; bicycle infrastructure is not a variable that was considered in this analysis, although it was considered and found to be significant in Rixey 2013 (10–11). Also illustrating the volatility of the models are the two stations on the western side of the city, which are highly underpredicted by the random forest model, and highly overpredicted by the linear model.

In San Francisco, station popularities are less spatially autocorrelated to begin with (I=0.49 for log(Popularity)). The refit linear model leaves autocorrelated residuals, while the refit random forest leaves residuals that are not significantly autocorrelated. Examining the residuals of the linear model (Figure 10), one can again hypothesize that autocorrelation is due to the exclusion of some variables (especially since the random forest does not exhibit significant residual autocorrelation). Using more variables captures more of the structure of the data, although overfitting is still a possibility (see paragraph about DC, above).

The models are almost certainly misspecified. These models are based only on accessibility to jobs, residents, and other bikeshare stations. The most straightforward theoretical reason for positive correlations between accessibility measures and station popularities is that bikeshare users are accessing the amenities represented by the accessibility measures. With that assumption, a properly-specified model would contain accessibility measures to all of the things that bikeshare is frequently used to access. A study of Capital Bikeshare users found that they did use bikeshare to go to work, but also to access shopping, social events, restaurants, and so on (Capital Bikeshare 2013, 31). It would thus make sense to add measures of accessibility to things such as shops. Many bikeshare trips

are used to access transit (Capital Bikeshare 2013, iv). This analysis included transit stops only implicitly (through the use of transit-based accessibility measures). It might make sense to include measures of accessibility to transit stops. Others have used additional variables and found them to be significantly correlated with bikeshare use. For instance, Rixey found that accessibility to alternative commuters (bicycle/walk/transit) and educated individuals had positive effects on station popularity (2013, 4, 10).

Adding these additional measures, however, will make models difficult to interpret and will decrease the significance of the coefficients (by increasing their variances).[8] The accessibility measures calculated thus far are highly correlated, and there is no reason to believe that additional measures will not be. Specialized statistical techniques could be used to deal with this issue. Random forests are one of these techniques, however they are also a fairly flexible method; flexibility is undesired because the models are intended for transfer. Fitting a model very close to the training data may hurt transferability.

Two alternate methods for highly-correlated values are ridge regression and principal components regression. Ridge regression shrinks the coefficient estimates from their least-squares values, decreasing the flexibility because the model cannot fit as closely to the data (James et al. 2013, 215–17). This introduces bias but reduces variance, which is desirable when many correlated predictors are used. Principal components regression is also an inflexible method, and should work well with correlated predictors. It first constructs principal components along which the data vary significantly, then uses those as predictors in the a regression model (James et al. 2013, 230–36). If data are strongly correlated, low-numbered principal components should capture much of the variation in the data and reduce the variance of the fits. The bootstrap could be used to estimate the variance of the coefficients produced by principal components regression.

8 Adding additional correlated variables increases the variance of the coefficient estimates, which in turn decreases the *t*-statistic and the significance of the variable (James et al. 2013, 101).

In sum, modeling bikeshare use in one city worked fairly well, but transferring the models did not. A significant connection between accessibility and bikeshare use was found, however. Adding additional accessibility measures should help with model specification issues, and ridge regression or principal components regression could be used to constrain the flexibility of the models and address the effects of high correlation of the predictors.

Appendix: Software Used

Accessibility measures were calculated using OpenTripPlanner (OpenTripPlanner Team 2014). Data were processed and maps were made using QGIS (QGIS Team 2014). Data were loaded into the R Statistical Programming Environment for analysis (R Core Team 2013). The R package plyr (Wickham 2011) was used to manipulate data, ggplot2 (Wickham 2009) and scales (Wickham 2012) were used for data graphics (specifically correlation matrices), leaps was used for best-subset selection (Lumley 2009), spdep was used for calculating spatial statistics (Bivand 2013), and the randomForest library was used to fit random forests (Liaw and Wiener 2002). Code adapted from from *An Introduction to Statistical Learning* was used to perform cross-validation in conjunction with best-subset selection (James et al. 2013, 249).

Acknowledgements

The author wishes to thank Kostas Goulias in the UCSB Department of Geography for his assistance with this project, and Eric Fischer for his assistance with San Francisco system data. Any errors that remain are, of course, my own.

GTFS data courtesy:

- WMATA: WMATA Transit information provided on this site is subject to change without notice. For the most current information, please visit http://www.wmata.com.
- Fairfax Connector: Visit fairfaxconnector.com for more information

- Maryland Transit Administration
- Arlington Transit
- AC Transit
- Caltrain
- BART
- SamTrans
- San Francisco Muni: Reproduced with permission granted by the City and County of San Francisco. The information has been provided by means of a nonexclusive, limited, and revocable license granted by the City and County of San Francisco.

  The City and County of San Francisco does not guarantee the accuracy, adequacy, completeness or usefulness of any information. The City and County of San Francisco provides this information "as is," without warranty of any kind, express or implied, including but not limited to warranties of merchantability or fitness for a particular purpose, and assumes no responsibility for anyone's use of the information.
- Valley Transportation Authority
- Metro Transit

  Street network data © OpenStreetMap contributors, available under the Open Database License. See http://www.openstreetmap.org and http://www.opendatacommons.org.

References

Bay Area Bike Share. 2013. "Pricing." http://bayareabikeshare.com/pricing.

Bivand, Roger. 2013. "Spdep: Spatial Dependence: Weighting Schemes, Statistics and Models." http://cran.r-project.org/package=spdep.

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

———. 2014. "Pricing." http://capitalbikeshare.com/pricing.

Conway, Matthew Wigginton. 2013. "Analyzing the Effects of Space and Time on Bikeshare Use: A Case Study in Washington, DC". Unpublished. http://www.indicatrix.org/publications/2013/Conway-Bikeshare-SpaceTime.pdf.

Dempsey, Walter, Juan-Pablo Velez, Adam Fishman, Jette Henderson, Breanna Miller, and Vidhur Vohra. "Poisson_data_extract.py." https://github.com/dssg/bikeshare/blob/38068a4ba8f99ba3b2ad32987eee3ed6629df8ab/model/poisson_model/poisson_data_extract.py.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. New York: Springer.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. *An Introduction to Statistical Learning, with Applications in R*. New York: Springer.

Liaw, Andy, and Matthew Wiener. 2002. "Classification and Regression by randomForest." *R News* 2 (3): 18–22. http://cran.r-project.org/doc/Rnews/.

Lindeke, Bill. 2014. "Updating the Downtown Nice Ride Numbers: Saint Paul Still Lags." *Streets.mn*,

January 17. http://streets.mn/2014/01/17/updating-the-downtown-nice-ride-numbers-saint-paul-still-lags/.

Lumley, Thomas; using Fortran code by Alan Miller. 2009. "Leaps: Regression Subset Selection." http://cran.r-project.org/package=leaps.

Nice Ride Minnesota. 2014. "Subscriptions." https://www.niceridemn.org/subscriptions/.

O'Sullivan, David, and David Unwin. 2010. *Geographic Information Analysis*. 2nd ed. Hoboken, NJ: John Wiley & Sons.

OpenTripPlanner Team. 2014. "OpenTripPlanner." http://www.opentripplanner.org.

QGIS Team. 2014. "QGIS." http://qgis.org.

R Core Team. 2013. "R: A Language and Environment for Statistical Computing". Vienna, Austria. http://www.r-project.org/.

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.

Wickham, Hadley. 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer New York. http://had.co.nz/ggplot2/book.

———. 2011. "The Split-Apply-Combine Strategy for Data Analysis." *Journal of Statistical Software* 40 (1): 1–29. http://www.jstatsoft.org/v40/i01/.

———. 2012. "Scales: Scale Functions for Graphics." http://cran.r-project.org/package=scales.

Copyright © 2014 Matthew Wigginton Conway. All rights reserved.