Analyzing Bay Area Bikeshare Usage in Space and Time

Adam Wilkinson Davis (corresponding author)
Graduate Student Researcher, Department of Geography and GeoTrans Lab
University of California, Santa Barbara
Santa Barbara, CA 93106
Phone: 650-534-6597
Email: awdavis@geog.ucsb.edu

Jae Hyun Lee
Graduate Student Researcher, Department of Geography and GeoTrans Lab
University of California, Santa Barbara
Santa Barbara, CA 93106
Phone: 805-284-7835
Email: lee@geog.ucsb.edu

Konstadinos G. Goulias
Professor, Department of Geography and GeoTrans Lab
University of California, Santa Barbara
Santa Barbara, CA 93106
Phone: 805-284-1597
E-mail: goulias@geog.ucsb.edu

Number of Words (including refs) = 7,743
Number of References = 25
Number of Tables and Figures = 5*250 = 1,250
Total equivalent = 8,993

Paper submitted for 94th Annual Transportation Research Board Meeting, Washington, D.C.,
January 11-15, 2015
ABSTRACT

Bikeshare programs are an increasingly popular means for cities to provide alternative transportation for their residents and visitors. In Spring 2014, numerous American bikeshare systems released comprehensive rental records, generally spanning several months and hundreds of thousands of trips, which allow researchers to study system use in more detail than is often possible with urban bicycling. Bay Area Bikeshare’s data release covers the first six months of the system’s operation and formed the basis of this research. This paper seeks to differentiate between the different general use patterns that emerged from the bikeshare data by analyzing rental start times, aggregated trip azimuths, and logit regression. The results suggest that most San Francisco bikeshare usage falls into one of two categories: weekday commute travel focused on transit stations and office buildings, and mid-day and weekend leisure travel that was much less restricted by time and focused on areas with denser and more diverse opportunities.
LITERATURE AND INTRODUCTION

Bikeshare

Due to bikeshare programs’ relative newness and the limited data available with which to study them, there has been fairly little scholarly work on the subject. Multiple studies have compared system design and history, and this work indicates San Francisco’s system is very typical of contemporary systems (1, 2). Most quantitative studies have used changes in station-by-station bike availability (which many programs make available in real time online) as a proxy for system usage that limits the ability of analysis to provide meaningful results. One review article on the subject found the following general results (3):

- Bikeshare trips may not replace car trips, since many users would have brought their own bike or taken transit to make the same trip in the absence of the system (though other studies have shown the opposite result (4)); bikeshare users appear to value the convenience of stations most when choosing to use the system; proximity to other bikeshare stations and to densely populated neighborhoods and transit are generally the best predictors of a station’s usage (5); and different systems exhibit widely varying spatial/temporal patterns that depend on the geographic extent of the system and its users.

Another study sought to create a typology of bikeshare systems based on their temporal use patterns, pay structure, and subscriber-customer balance. The biggest differences in use among systems corresponded to their coverage, with stations covering only the city center tending to be used very differently than systems with stations that extended into peripheral areas (6). Bay Area Bikeshare could be categorized as belonging to either of these categories, since it is centered on downtown San Francisco and San Jose, but has numerous stations in suburban areas. In addition to system extent, access restrictions appear to play a significant role in system use. Both Barcelona and London initially limited who could use the systems, with Barcelona only permitting city residents (to avoid competing with existing touristic bike rental companies) (7) and London initially requiring all users to pay for an annual subscription.

O’Brien et al (6) suggest London’s reason for easing this restriction may have been to balance system usage and redistribute bicycles, since subscribers tend to leave some stations with no available bikes and others with no empty slots during commute hours. A study that focused on the effects of adding short-term users to the London system indicates they may not have had the desired effect (8). Though Lathia, et al. (8) are subject to the same data constraints that most early bikeshare studies face, their work provides valuable information on the specific effects of short-term users. They found that new users reinforce the overall temporal peaking of system use (which is clearly not the case in San Francisco) but substantially change the use patterns of some stations. Some stations switched from having generally high availability (suggesting low use) to having severe bicycle shortages for most of the day on weekdays. Stations near tourist-heavy locations like Westminster Abbey seemed to be particularly likely to see increases in use (8). Since their study predated the London system’s release of origin-destination data for individual rentals, their conclusions were limited by lack of data, but they do show some differences in behavior between subscribers and short-term customers. San Francisco’s detailed origin-destination data indicates that the differences between user types are more clearly defined, at least in San Francisco.

Bicycle use in the environment

Though weather and the built environment impact all modes to some degree, bicyclists are particularly vulnerable to environmental conditions. Numerous articles have studied the effects of severe weather and found that it has a substantial impact on bike use, though most research on bicycle routes has focused specifically on bicycle facilities and the built environment. In a study of bike commuters in Vermont,
Sears et al. (9) found considerable bicycle use variability due to temperature and precipitation but no weather-independent patterns relating to hours of daylight (despite many respondents indicating that darkness would dissuade them from riding) or time of year. This work indicated that occasional bike commuters were 3% more likely to use their bike on a given day for each additional degree Fahrenheit (9), but given that temperatures in San Francisco rarely drop below the mean temperature reported by this study, it is doubtful whether this effect would extend to San Francisco.

A comparative study analyzing bike use in five North American cities (including San Francisco) classified bicycle facilities and by primary use type (utilitarian-recreational). The researchers found strong seasonal variability in recreational bike use but much less among commuters. All seasonal effects were weaker in San Francisco than in the other cities studied, all of which experience much more severe winters. “Utilitarian” networks tend to experience much higher ridership during the workweek and typically see a peak during the morning commute and a larger one during the evening commute, while “recreational” systems have a single broad peak that runs from late morning to midafternoon and have more traffic on weekends (10).

In addition to their vulnerability to weather, bicycle users’ mobility is strongly affected by local topography, with steeper sections of road rendered nearly impassable. Bicyclists strongly disprefer routes with steep terrain, and particularly seek flat routes in cities (11). A study in the Netherlands indicated that hilly topography can greatly decrease local bicycle use even in areas where it is an otherwise popular mode (12).

The built environment has been the focus of many policy-oriented papers on urban cycling (12), with considerable attention to the effects of bike-friendly infrastructure on bike use. San Francisco has relatively few dedicated bike paths, but traffic calming and painted bike lanes are nearly ubiquitous in the study area (13).

Land use density, diversity, and travel

Previous work incorporating diversity/entropy measures into transportation studies found that dense and diverse mixes of land use had a substantial impact on transportation, specifically increasing the use of non-motorized modes (14). One of the reasons for this effect is that while bicycles travel slower than cars, cars are much more time consuming to park and are much less maneuverable in dense areas, giving bicycles a substantial “effective speed” advantage (15). Additionally, denser areas provide a wide range of opportunities within a small area, encouraging short trips on foot or by bike. By providing dedicated, self-locking parking in dense areas, bikeshare programs may compound these advantage.

Studies of land use diversity based either on four-way classified land use entropy – residential, public facilities, industry, and offices (16) – or a similar set of categories for “mix of land use” and density (17) found that areas of higher diversity corresponded to more frequent use of alternative modes. Research classifying urban land use classified at a finer scale also found diversity to particularly favor slower “active” modes (18). Previous work in the Bay Area found that general land use entropy indices that included multiple types of work land-uses were much more strongly linked to travel behavior than indices based either on highly-specific land use classes or on more general classification schemes that grouped all work locations (19).

BAY AREA BIKESHARE

Although a variety of designs have been used since the first systems debuted in the 1990s, most modern bikeshare systems operate a network of stations, each with numerous docks at which bikes can be checked out and checked in at will (1). Some systems are concentrated in the core of an urban area, some
cover a broader area of a city, and some extend into the suburbs. Most systems require users to buy a subscription either for a full year or a shorter-term pass covering 1-3 days and offer free trips for up to 30 minutes. Trips beyond 30 minutes incur fees, which vary considerably among systems. In the Bay Area Bikeshare system, an annual membership costs $88, a three-day pass costs $22, and a one-day pass costs $9; overage charges are $4 for trips between 30 and 60 minutes and $7 for each additional 30 minutes (20). The goal of this fee structure may be to keep as many bikes available at once as possible, but because a user can check out a bicycle shortly after checking one in, experienced users may be able to avoid incurring overage fees.

Bay Area Bikeshare includes 35 stations in downtown San Francisco and 34 in southern San Mateo and northern Santa Clara Counties (the South Bay). The area served by the South Bay portion of the network is considerably more suburban, which has a major impact on overall system use (6) and San Francisco stations served as the start and end points for roughly 90% of all rentals. To keep the analysis tractable, in this paper we exclude the South Bay portion of the system. In San Francisco, no pair of stations is more than two miles apart, so all direct trips should be possible within the 30 minute time frame, though numerous rentals are longer.

Many bikeshare systems made data public through Open Data Challenges. The general goal of these programs seems to have been to raise awareness of the systems and promote bikeshare scholarship. Bay Area Bikeshare provides checkout and return times and locations (OD Pairs) and user type (annual subscriber or 1/3-day customer) and home zip code for all 144,015 rentals made between the system’s opening in late August 2013 and February 2014. The database also includes geographic coordinates and opening dates for each station as well as daily weather data for each city the system serves. Though San Francisco has diverse microclimates, the weather data matches a station in Downtown and reports conditions very similar to the weather experienced throughout the study area.

About 85,000 bicycle trips were made per day in San Francisco in 2013, a mode share of about 3.5% (21, 22). Over its first six months of existence, Bay Area Bikeshare averaged 705 trips per day, or nearly 1% of San Francisco’s total.

After the system’s opening, Bay Area Bikeshare usage increased gradually, with subscribers eventually coming to provide the bulk of the system’s use, then maintained a relatively steady state from October, 2013 on, with minimal seasonal effects independent of rainfall. As figure 1 shows, midweek rental totals are extremely consistent, with the few major decreases in usage occurring during days with substantial rainfall, when all use decreases dramatically, and holidays (notably Thanksgiving, Christmas, and New Year’s Day), when subscribers make far fewer trips than usual and short-term users make far more.
FIGURE 1 Time distribution of San Francisco bikeshare use

Citywide counts indicate that bike trips are fairly consistent during the week, with Saturdays typically seeing about 90% as many trips and Sundays around 60% of typical weekday volumes (21). Bikeshare use exhibits a considerably steeper decline on weekends, when less than half as many trips are made either day as on a typical weekday. This pattern presumably results from the system’s concentration in the downtown area and commute-dominated “utilitarian” bike use patterns during the week (21), a pattern that often corresponds to much heavier use on weekdays as well (10). From Monday to Thursday, the system resembles what Miranda-Moreno et al. (10) describe as “mixed utilitarian,” with strong peaks around commute hours, and fairly consistent, lower use between peaks (figure 2) (10). A small peak in use around lunchtime is also apparent on weekdays.

Bay Area Bikeshare use as a whole is generally consistent with that found in other utilitarian systems, but short-term customers and weekend users exhibit very different temporal preferences. Weekend usage is much more in line with a “recreational” system (10). Recreational use has a single, broad peak in late morning and early afternoon, with use gradually increasing before that and decreasing afterward (figure 2). Non-subscribers closely match the weekend use patterns, suggesting they can generally be categorized as recreational users. Based on the home zip codes of users who provided them, short-term customers are also much more likely to live outside the Bay Area, which may partly explain why they do not use the system like the commuting subscribers do.
FIGURE 2 Temporal usage patterns differ by day and user type.

METHODS AND ANALYSIS
To study Bay Area Bikeshare use in time and space, we performed two main sets of analysis. While several previous studies used the availability of bikes as a proxy for usage of individual stations, the available dataset makes it possible to bring the characteristics of specific rentals into the analysis. Because the South Bay cities served by the system are much flatter and less dense than downtown San Francisco, all trips that included a station in the South Bay are excluded.

Bikeshare Use in Space – Methods
Because this dataset includes over 100,000 bike rental records and nearly 2,000 possible station pairings, aggregation and simplification were necessary in order to make sense of the spatial patterns of trips.
Though we lack detailed information about the specific routes individual bikers took, we can get a general sense of direction from the frequencies of individual station pairings. The aggregate direction of bikeshare travel varies substantially among the stations, by time of day, and by type of user, and this paper seeks to determine the extent to which these differences are systematic and explainable.

To study the changes in the spatial use of the bikeshare system between stations and in time, we calculated the average direction of travel from or to each station. These average travel vectors were created by summing the vectors of all individual trips that either started or ended at the station. The few trips made from or to station 82 (Broadway St. at Battery St.) were excluded because the station was added to the system late in the study period. Trip azimuth (in degrees clockwise from north) is equal to the arctangent of the ratio between the total station-to-station offset in x and y, using UTM coordinates (equation 1). This was calculated using the Python programming language as math.atan2(difX,difY). This method is also shown in figure 3. The total azimuth from multiple trips can be calculated in this manner using the total displacement in each direction among all trips. Mean azimuths can also be calculated from a set of azimuths, using equation 2, which weights all angles equally by converting them to distances in x and y along the unit circle. The former method was used to calculate the total azimuth for each station and the latter was used to calculate the mean of simulated azimuths. Average azimuths were calculated independently for trips starting from and ending at each station, with the following subsets: weekday subscribers, weekend subscribers, and customers; average azimuths were also calculated by time of day for each station using time intervals that corresponded to generally consistent use patterns: AM Rush (7-9:30), Mid-Day (9:30-16:30), PM Rush (16:30-18:00), Evening (18:00-23:00), and Late Night (23:00-7:00).

\[\text{Azimuth} = \text{atan2}(\Delta X, \Delta Y)\]

**Equation 1:** Azimuth (angle clockwise from north) is calculated from the displacements between the stations in X and Y. This can also be used to calculate the overall azimuth from multiple trips.

\[\text{Mean Azimuth} = \text{atan2} \left( \frac{1}{N} \sum_{n=1}^{N} \sin a_n, \frac{1}{N} \sum_{n=1}^{N} \cos a_n \right)\]

**Equation 2:** Mean azimuth is calculated by converting azimuths into coordinates on the unit circle and taking the arctangent of the ratio between the sums of each.
FIGURE 3  How to calculate an overall azimuth from two trips.

To provide a basis for comparison, we estimate the probability distribution function for the angular mean vector at each station by simulating 10,000 sets of overall azimuths using three different random trip assignment methods, one fully random and two weighted. To simulate station departure azimuths, for each simulation run, a number of OD pairs was generated equal to the total number of times a bike was checked out from that station and returned to a different station. For random trip assignment, all destination stations had an equal probability of being selected for any given trip, with the displacement in x and y between that station and origin station added to running totals. For departure-weighted assignment, destination stations were selected using a weighted random process with weights equal to the total number of bicycles arriving at that station. To simulate the arrival azimuths, station arrival totals set the number of trips simulated with a particular station as its destination, and departure totals were used for weights. Once sufficient random trips were generated, the resulting mean angle was extracted using equation 1, and an overall angular mean of means was calculated from all simulated mean azimuths using equation 2.

Distance-weighted simulation is based on the average rental OD pairs per possible pair within each 500 meter distance band. Renting a bike is not much faster or more convenient than walking for fairly short distances, so bicycles were rarely checked out from one station and returned to one very nearby. Weights were highest in an intermediate distance range (500 to 3,000 meters), and lowest for trips beyond 3,000 meters, likely because of the 30 minute free rental limit. For each station, network distance to all other stations was extracted (using the Google Maps API) classified into a 500 meter range, and assigned a weight equal to the total number of trips between all pairs of stations within that distance range divided by the number of possible pairings. The rest of the simulation worked the same as the use-weighted trips, using the distance-based weights instead of the usage-based ones.

Though circular distributions do not have meaningful positive or negative directions, approximations of one-tailed or two-tailed tests are still possible. In both cases, these were performed by comparing a station’s observed azimuth against the distribution of simulated azimuths. Two-tailed tests were run by counting the simulated azimuths at a greater absolute distance from the mean of simulated
total azimuths than the observed azimuth was. One-tailed tests were performed by counting only those simulated azimuths were farther in the same direction. Because each station had a different number of trips per simulation, azimuths vary more at less popular stations.

Because they are generated through a consistent process that necessarily favors a certain direction, simulated azimuths strongly resemble the von Mises (circular normal) distribution, which approximates a Gaussian distribution wrapped infinitely around a circle (23). Because it is relatively straightforward to estimate a probability distribution function through simulation, it is unnecessary for this research to estimate the shape parameter of the von Mises distribution.

A set of regression analyses were performed on the set of time-specified station azimuths (n = 34 stations * 5 time periods), using indicator variables for each station and time period as the only independent variables (the late-night time period indicator and one station indicator were excluded). To escape the problems caused by attempting to model cyclical angles using linear regression, azimuths were converted to coordinates along the unit circle and regressions were performed on the resulting northing (equal to the cosine of the azimuth) and easting (equal to the sine of the azimuth). Because Market St. (which divides the street grid) generally runs SW-NE and many other key features of the study area are not in a cardinal direction from the center (notably Civic Center in the southwest and Caltrain in the Southeast), this analysis was also rotated, with northeasting and southeasting calculated using, respectively, the cosine and sine of the azimuth minus 45°.

Stations with azimuths that are consistent across all time periods should have coefficients that match their overall direction. Stations with coefficients that do not fit this pattern presumably have temporal variability that may be explained by the coefficients on the time period indicators. The directional analyses can either be used jointly (which will cover both perpendicular directions) or independently (in which case any value close to 0 implies the azimuth is close to perpendicular to the direction of analysis).

Use Vectors

Since it would be impossible to study all rental events or even all station pairs individually, some degree of abstraction is necessary. Stations form the start and end points for all bikeshare rentals, so they are a reasonable target for analysis. Station-specific temporal patterns have been studied by numerous researchers, as detailed above, but few of those studies were able to investigate the spatial effects of station pairings due to data limitations. Figure 5a shows the mean rental vector for bicycles checked into (red arrows with heads at stations) and checked out from (black arrows with tails at station) each station, scaled by total number of trips to or from, respectively. The From vectors appear mostly to be diverging and the To vectors appear to be converging, which an inevitable consequence of the arrangement of stations, but local patterns are more varied.
FIGURE 4  Station azimuths a) Overall (scaled by number of trips), b) In different time periods, c) between subscribers and customers, and d) for subscribers between weekdays and weekends.
In most cases, the size and direction of from and to vectors match, but some stations show
definite differences between the two. Station 73 on Telegraph Hill has half again as many departures as
arrivals, requiring Bay Area Bikeshare to refresh its supply frequently. While many different factors could
cause this sort of imbalance, local geography suggests a simple explanation: this station has the highest
elevation in the system by a considerable margin and is in an area of very steep slopes. Bicycles are more
sensitive to steepness than any other mode, so cyclists are much more interested in riding down from
Telegraph Hill than up it. The two popular stations at the San Francisco Caltrain station (69 and 70) have
oddly divergent trajectories, with riders departing from station 70 heading in a more northern direction
than those departing from station 69. Station 70 also has more trips generally, possibly because different
stations on different sides of a street provide different degrees of access to each direction of traffic. In
addition, the mean vector for rentals heading to these stations match each other but are substantially
different from either of the from vectors, suggesting that users of these stations users may pick which
station to leave from based on where they are headed, but do not make the same choice when returning to
the station.

Stations in the interior have even more widely varied vectors, partly because they are less
constrained by station geometry than those around the exterior. To see if station directions can be
predicted using any simple rules, observed vectors were tested against the simulated vector means and
distributions. In almost all cases, station azimuths differed significantly from the simulated means.
Because the simulated distributions generally seemed symmetric around the mean, roughly twice as many
simulated azimuths were beyond the observed azimuth in two-tailed tests as in the one-tailed test. None of
the simulation models consistently matched actual azimuths, but the weighted models generally did
better.

Simulated Azimuths
For randomly simulated azimuths From, only five stations (47, 62, 64, 71, and 75) had any simulated
values farther from the mean in either direction than the observed azimuths, while 29 stations did not;
stations 47 and 62 were closest to the simulated overall mean than 82.0% and 76.5% of simulated
azimuths, respectively; the other three stations were farther from the mean than more than 98% of
simulations. There were similar results for randomly simulated azimuths for trips To each station, as only
five (47, 49, 56, 71, and 75) were within the range of any simulated azimuths, and only two were not
significantly different from the distribution: station 49 was closer to the mean than 50.6% of simulations
and station 56 was closer than 29.9%. Weighted random simulation led to more accurate simulation,
though for most stations, the difference was minor. For trips From stations, 9 had observed azimuths that
were closer than at least one simulated azimuth, and 4 (41, 62, 64, and 67) were closer than at least 5% of
simulated azimuths. For trips To, 6 stations that were closer than at least one weighted simulation run and
5 (42, 46, 58, 64, and 67) that were closer than more than 5%. Somewhat surprisingly, distance-weighted
simulations performed worse than trips-weighted simulations. For the From azimuths, only 6 distributions
included the actual distribution and only station 75 was closer to the simulated mean than more than 10%
of the simulations. Distance-weighting performed substantially better for To azimuths, with 8 falling
within the range of the distributions and 3 falling closer than 10% of simulations. Clearly bikeshare
station pairings follow a structure that is not simply a matter of overall popularity or distance.

The spatial distribution of bikeshare trips depends greatly on who is making the trips and when.
Much as the weekend start times were much more consistent with short-term customers and recreational
use than the commute-dominated weekdays, weekend riders and customers seem to prefer more
recreational destinations as well. Figure 5c shows the overall azimuths for weekday vs weekend trips. The
apparent center of convergence for the weekday trips is considerably to the south of the apparent centers,
likely because they are much more likely to use bikeshare to get to Caltrain or less commercial areas south of Market St. Weekend users and customers have a much more varied pattern, partly due to the smaller numbers, but they appear to make more trips further north (likely to the stations on Embarcadero along the bay) and appear to converge between the Union Square and Chinatown shopping districts. Figure 5d shows that subscriber weekend behavior is fairly similar to customer behavior Figure 5b

Regression on Direction

Regression analysis of the directional components of station azimuths yielded some compelling results. Most individual stations (20 of the 33 included) have coefficients that are significant at a 0.05 level in both directions; 26 and 27 are significant in the north and east directions, respectively. Temporal indicators have some effect: for the north/east regressions, the mid-day time period had the single most significant coefficient (p=0.009), tending to increase the northern component of station azimuths. The northern part of the study generally has a higher business density and particularly numerous restaurants and tourist attractions, which generally will be more popular in midday than during the commute hours. Evening rush hour had a significant coefficient (p=0.019) in the eastern direction, possibly because so many people ride bikes to Caltrain (stations 69 and 70). Rotating the axes being compared by 45° shows somewhat similar results. 16 of the stations have significant coefficients in both directions, with the main difference being that only 19 stations are significant in the southeast-northwest direction. This indicates that bikeshare travel parallel to Market St. is generally more significant than travel perpendicular to Market. Individual time periods are more likely to have a significant impact when the analysis is rotated. The mid-day time period is significant (at the 0.1 level) in both directions, generally favoring travel towards the northeast and northwest. Afternoon commute and evening travel both tend to push bikeshare trips southeast, likely because of the strong attraction created by the Caltrain station.

Because rebalancing bikes among stations is a major expense for bikeshare systems, and unbalanced use can lead individual stations to have no free slots or no bikes at times (3), it is worth considering whether promoting a wide variety of users might improve the overall balance of the system. In the London study, short-term customers did not improve the overall balance of the system – instead they shifted the imbalances around, mainly affecting areas popular with tourists (8), but in San Francisco, recreational use clearly maintains steadier use rates throughout the day, and may have the same effect spatially. The differences between trip directions at different times of day indicate that different user groups can help rebalance bikes in the system to some extent. Further analysis of the use and bicycle/slot availability of specific stations would be required to confirm this.

BIKESHARE USE IN TIME – PREDICTING OVERAGES

Bikeshare use in time

For this analysis, we focus on individual rental events. Several users reported having issues with the bike check-out and check-in process, which contributed to the large number of OD Pairs that started and ended at the same station within a brief time interval; all such entries with duration under three minutes are excluded. Similar errors were found in other systems, and these reported trips were eliminated in other studies (24). After the cleaning process, 129,068 OD Pairs remained.

Though San Francisco bikeshare users usually returned their bike less than 30 minutes after checking it out, the data included 7,664 OD Pairs with durations over 30 minutes. Given the available data, it is impossible to tell precisely why people sometimes kept their bikes past the free period, but given the overwhelming share that were made by non-subscribers (who accounted for only 21% of checkouts, but nearly 90% of overages), it seems reasonable to suggest at least some of these trips were recreational in form. Earlier bikeshare systems restricted membership in order to discourage touristic use
– in Barcelona, this was explicitly with the goal of protecting the city’s existing bike rental system (25) – though London has eased these restrictions, possibly to combat spatio-temporal demand imbalances caused by commuter-heavy use (6).

To attempt to understand overages (checkouts over 30 minutes in length) we use binary logit regression. In initial runs that included all OD pairs, many variables with very small effects were rated significant and the model seemed focused on overwhelming majority of the trips that were under 30 minutes. To diminish these effects, a new dataset was constructed that contained all 7,644 overage trips and a random sample of 7,645 non-overage trips. The resulting model took into account local opportunities, time of day, subscriber status, day of week, weather, and other factors that seem to influence bikeshare program use patterns. The following data was added to each OD pair for use in the model, specific methods are described below (all source data provided in Open Data Challenge files unless noted):

- User home zip codes classified by region of origin (zip centroid within 1 mile of a station, San Francisco City/County, nine county Bay Area, and elsewhere) using ArcGIS and data available through ArcGIS online.
- Start times classified into five ranges that correspond to similar use patterns: AM Rush (7-9:30), Mid-Day (9:30-16:30), PM Rush (16:30-18:00), Evening (18:00-23:00), and Late Night (23:00-7:00).
- Civil daylight hours from timeanddate.com, which turned out not to be significant at all, confirming that bicyclists may not worry about daylight as much as they say they do (9).
- Bicycle network distance and predicted travel time between the origin and destination station, collected using the Google Maps API.
- Presence of a BART or Caltrain station within 200 feet of origin or destination station.
- Number of docks at start and end stations (min: 11, max: 27).
- Opportunity-based accessibility: half-mile and one-mile (network distance) counts of business establishments and employments at the start and end stations and the midpoint between stations. Business establishment counts were divided by 1,000 and employment counts were divided by 10,000 for ease of use. Method discussed below.
- Shannon and Simpson diversity indices of business establishments (within half a mile of the start and end station and midpoint. Method and categories discussed below.

Opportunity-based accessibility, a stand-in for density, was computed for each station and midpoint. This measure was calculated by counting all businesses and all employees that fell within one or one-half mile of the station, based on the distance along bike-accessible roadways extracted from the 2010 US Census Tiger Line networks. A total of 72,735 business establishments and their respective employment totals were drawn from the NETS 2010 dataset (a database that tracks the location, employees, and industry type of every business establishment in the United States). The data row for each OD Pair (rental) thus included 12 opportunity-based accessibility metrics: employees and business establishments at the start and end stations and the midpoint between them at both a ½ and 1 mile radius.

Considerable work has gone into quantifying diversity in the fields of ecology, where species diversity and richness are important, and information science, where uncertainty is a major focus. Economists have developed several diversity indices (notably Ogive Diversification Index, Krugman Specialization Index, and Gini), but these seem mainly to be used in macroeconomic research to address the overall makeup of a country or region's labor force or production. Species richness diversity measures seem to be a better way to understand the range of opportunities people experience when traveling.
through an urban area. Because they are simple to calculate and understand and requires only class
proportions as an input, the Shannon (equation 3, left) and Simpson (equation 3, right) diversity indices
are used here. Simpson diversity is equal to the probability that two randomly selected members of a
population (in this case San Francisco businesses) will belong to the same class. Shannon entropy
quantifies the uncertainty in the class membership of a randomly selected entity (26).

\[ \text{Shannon Entropy} = - \sum_{i=1}^{R} p_i \ln p_i \; ; \; \text{Simpson Diversity} = \sum_{i=1}^{R} p_i^2 \]

Equation 3 Shannon Entropy and Simpson Diversity Indices for R groups, assuming large samples
of each.

To calculate class counts for each station or route midpoint, a Python script iterated over all
business establishments, classified them by type, and added them to the appropriate category count of all
stations within half a mile (in cartesian space). The data used for this study included a comprehensive list
of San Francisco business establishments with SIC classifications. Businesses were classified into the
following 11 categories based on their 2-digit SIC codes: Construction, Entertainment, Finance/Insurance,
Food Service, Government, Grocery, Industrial, Retail, Service, Transportation, and Other. The periphery
of downtown scored highest by this metric, likely because more residential areas require a wider range of
businesses to serve the local community, whereas a few categories (namely entertainment, finance, and
food service) cluster downtown. Each OD Pair was assigned the Shannon entropy value for its start and
end stations and the midpoint.

Subscribers and short-term customers use Bay Area Bikeshare very differently from each other,
and this difference is the single clearest predictor of whether a user keeps a bike longer than 30 minutes,
both in terms of the magnitude of the effect and the reasonableness of the explanation. Therefore the
process of developing a useful model necessarily included this variable in all steps. Variables tested but
rejected include several time categories (on the assumption that commute-hour users are more likely to be
in a hurry), daylight (which was assumed to relate to tourist activity or general willingness to travel for
long periods by bike), and temperature (which was shown to be a significant predictor of bike use in other
cities, but appeared to have no impact in San Francisco, presumably in part because the city has mild
weather year-round). Because they match the scale of travel in downtown San Francisco and were
relatively more significant in the models, half-mile accessibility and diversity measures were used instead
of the full-mile ones. Coefficients were converted to odds ratios for ease of interpretation.

The final model kept subscriber status; an indicator for trips starting and ending at the same
station (AtoA); a Bay Area home zip code indicator variable and an interaction term with subscriber
status; one time-of-day indicator for the mid-day time period (9:30 to 16:30); precipitation in inches;
indicators for Saturday, Sunday, and Federal Holiday; an indicator for trips ending near a BART/Caltrain
station; the number of docks at start and end stations; Shannon entropy for business establishments
around the end station; and number of business establishments around the midpoint (divided by 1,000).

(Table 1)

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Coefficient</th>
<th>Odds Ratio</th>
<th>Std. Error</th>
<th>P</th>
<th>Mean of X</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constant</td>
<td>-0.6789</td>
<td>0.5072</td>
<td>0.4340</td>
<td>0.1177</td>
<td></td>
</tr>
<tr>
<td>Subscriber (dummy)</td>
<td>-2.0743</td>
<td>0.1256</td>
<td>0.3091</td>
<td>0.0000</td>
<td>0.4584</td>
</tr>
<tr>
<td>AToA (dummy)</td>
<td>2.6493</td>
<td>14.1436</td>
<td>0.1049</td>
<td>0.0000</td>
<td>0.1750</td>
</tr>
<tr>
<td>Bay Area Zip Code (dummy)</td>
<td>-0.4589</td>
<td>0.6320</td>
<td>0.0667</td>
<td>0.0000</td>
<td>0.6284</td>
</tr>
</tbody>
</table>
Subscriber status and AtoA had the strongest effect, corresponding respectively to an 88% decrease and a 1310% increase in the odds of a trip going over 30 minutes. Residents of Bay Area zip codes were also much less likely to take trips over 30 minutes (and the interaction term between Bay Area and subscriber status had an even stronger effect). The only time-of-day indicator that mattered was midday, when trips were roughly twice as likely to run overtime. Holidays and both weekend days corresponded to an increased likelihood over over-time trips. In addition to decreasing overall system usage, precipitation decreased the likelihood that a given trip ran over-time. These effects combine to indicate that over-time trips generally correspond to touristic behavior rather than regular users accidentally keeping a bike too long.

Station attributes also had substantial predictive power: trips that ended near a BART or Caltrain station were 24% less likely to run over, since those bikeshare stations were generally much more commute-dominated. Attributes of the start station matter less, though the number of docks had a nearly significant effect. In general, more popular stations have more docks, and trips that start at stations with many docks appear slightly more likely to run over 30 minutes. End docks had a significant negative effect, as larger stations attracted more people using bikes solely for transportation. Business establishment diversity also predicted more overage trips, suggesting that people travelling through areas with diverse land uses may be more likely to make stops during their trips. However, business establishment diversity was generally highest around the city’s periphery, where stations are rarer, making accidental overages more likely. Establishment density around the midpoint also has a strong effect, suggesting that bikeshare users travelling through areas of dense development are more likely to take their time, perhaps to window shop or explore the area.

The model is able to accurately differentiate rentals of under 30 minutes from those that went over, with a likelihood ratio of 11,167 (associated with a p-value of <1*10^-7 given the difference of 14 degrees of freedom between this and the restricted model). When model probabilities were converted to predictions using a cutoff of 50%, the model correctly predicted duration above or below 30 minutes for 87.95% of the 15,289 OD pairs. The model’s sensitivity (trips over 30 minutes that the model also predicted were over 30 minutes) was 92.75% and its specificity (trips under 30 minutes that it predicted were under) was 83.15%.
CONCLUSIONS

While bicycles are a relatively minor mode in most US cities, their rising popularity, lack of emissions, and support of an active lifestyle make them a valuable target for study. The availability of comprehensive bikeshare Origin-Destination datasets provides a unique opportunity to study urban cycling and urban transportation as a whole. Limiting trip start and endpoints to available stations is quite unlike normal bike use – since bicycles can be parked in a wide range of locations or brought into buildings if necessary – but in other ways, bikeshare seems sufficiently similar to normal bicycle usage that it can be treated as fairly representative in the area covered by the system. Unlike much of the “sharing economy,” such as carshare and rideshare, which have little or no upfront cost and charge by use, bikeshare programs put all typical charges up front and use overage pricing as a punitive measure. Though this fee structure discourages long trips, it also enables the system to bring in substantial revenue from them. Based on the stated pricing structures and the number of short-term passes and annual memberships sold (27), it appears that overage charges accounted for between 25% and 35% of the system’s total revenue during the study period (depending on the mix of 1- and 3-day passes).

The main finding of this research is that San Francisco bikeshare use can be split into two general categories: commute and leisure. Weekday morning and evening rentals tend to be briefer and are much more likely to start or end near a transit station. Their use exhibits two strong temporal peaks during commute hours, and generally centers on the business-heavy South of Market area, likely in part due to the pull of the numerous trips to/from the Caltrain station in the southeast. Leisure users likely have a broader range of reasons for renting bikes (urban exploration, getting lunch, etc.), but are generally much more likely to take bikes on longer trips and travel through more business-dense areas. As shown and shown by their directional azimuths, non-subscribers and all users around mid-day and on weekends tend to ride further north in the city, focusing on the retail- and dining-dense areas along Market St. and to the north. The differences between these groups are visible in the start time histograms, station azimuth maps, and overage-prediction logit regression model.

While local land uses clearly have a strong effect on Bay Area Bikeshare usage, these effects may be very different in other cities or less dense parts of San Francisco. Work comparing different bikeshare systems would be needed to determine whether networks of stations that extend beyond the core of downtown into less-dense areas create notably different system use patterns. Analysis that specifically investigated the effects of bicycle/parking availability within a system could determine the overall impact of system inefficiencies on the way people use bikeshare. Additionally, while the release of data with persistent rider identifiers raised privacy concerns in other cities, it would be extremely valuable to be able to tie trip records directly to specific individuals. This is needed to establish patterns that correspond to actual people and to distinguish between persons that took trips of a single purpose and activity (e.g., go to work) from bicyclists that take trips of combined purposes and possibly more complex activity patterns. see whether the use two use categories encompassed to distinct groups of people or whether many users took trips of both sorts.
References


