Research Paper: Predicting ‘cold start’ spatial interaction demand in a dock-based bike-sharing system

Zheng Liu[1], Taylor Oshan[1]

[1]Department of Geographical Sciences, University of Maryland, College Park, US
Email: {zliu1208; toshan}@umd.edu

Abstract

A rapid deployment of bike sharing system in recent years makes it imperative to study bike-sharing mobility. The study focuses on the unique challenge that is not applicable or less discussed in previous studies, the cold start of a dock station. Cold start is predicting demand for a new station where there is no knowledge of previous flows, which is necessary when the goal is to progressively update transport infrastructure in order to maintain efficiency and grow the ridership of bike-share systems. Consequently, this work investigates a methodology for predicting Origin-Destination (OD) flows of bike-share systems given that the system evolves over time and demand is constantly changing.

1. Introduction

The past century has witnessed a significant increase in urbanization with more than half of the world's population currently living in urban areas (DESA 2018). As a result, it has become increasingly important to understand and anticipate human mobility within and across cities. Transportation data is frequently used in urban mobility modelling. Among the modes of transportation, bike-sharing are becoming more popular, as millions of trips happen every month in the NYC Citi Bike system. Compared to other modes of mobility, cycling provides health and environmental benefits in addition to offering a more efficient means of navigating the urban environment (Oja *et al.* 2011; Otero *et al.* 2018; Wang and Zhou, 2017; Zhang and Mi, 2018). For example, the New York City (NYC) Mobility Report indicates that trips made using the NYC Citi bike-share system are over a minute faster than taxi trips across all distance categories within the Midtown area of Manhattan, and cost less than 25% for taxi trips for all trip length categories except those less than half a mile (NYC Department of Transportation, 2019). Such advantages are further highlighted during rush hour (Faghih-Imani *et al.*, 2017). This has led to the proliferation of bike-sharing systems, with more than 2000 bike-sharing systems now in operation around the world.

Individual bike-sharing programs are often piloted for a limited number of zones within an urban area that are known to have relatively high overall activity and transport demand. Systems are then expanded and updated according to the evolution of demand over time and across space, with new dock stations being added into a bike-sharing system. This dynamic development of bike-sharing systems requires a strong understanding of mobility behaviour and flexible methods for predicting spatial-temporal demand. More specifically, this work focuses on predicting demand for a new station where there may be no knowledge of previous flows, also known as a ‘cold start’, which is necessary when the goal is to progressively update transport infrastructure in order to maintain efficiency and grow the ridership of bike-share systems. Previous work has not examined in detail the task of predicting spatial interaction demand for new stations, likely because more traditional public transportation infrastructure evolves much more slowly compared to the relatively inexpensive and flexible bike-share infrastructure. That is, much of the related state-of-the-art research assumes that the infrastructure is persistent across time and there is prior knowledge of station activity, which is not always the case for bike-share systems. Meanwhile, it is reasonable to leverage spatial dependence to predict values at unobserved locations based on values from nearby observed locations using the First Law of Geography (Tobler 1970). This inspired us to develop spatial interpolation methods for new station flow estimation in addition to spatial interaction models.

2. Methods

2.1 Problem statement

For a station-based bike-share system, S_n, there are n docking stations serving as both origins S_i and destinations S_j and information is available for each trip in the system regarding its origin station, destination station, start time, and end time. Trips are also sorted into discrete temporal subsets, t , based on their starting time (e.g., hour, day, week, etc.). Therefore, each trip in system S can be denoted using a 3-tuple (t, S_i, S_j) and the corresponding OD flow matrix T is comprised of entries denoting aggregate trips between stations at time t (i.e., $T_{t,i,j}$). The diagonal elements of T(i.e., $i = j$) are filtered out and set to zero to remove their undue influence on any subsequent modelling procedures. A station 'cold start' problem refers the scenario where there is no information available about previous flows for a newly added station S_x and the goal is to predict future outflows $T_{t+1,x,j}$ and future inflows $T_{t+1,i,x}$.

2.2 Station classification

Cold start stations are first identified from original stations that were added at the beginning weeks of the system. In the case study of NYC, stations first added after 2015 can be practically classified as cold start stations. Next, one of three different classes is labelled to a cold start station depending on its relative location with the other dock stations. In the classification process, the convex hull of all the other stations in operation are first calculated as the current system coverage. Some extra processing steps are deployed to ensure geographical restrictions like rivers are not included in the convex hull. Then the Voronoi shapes are calculated from the set of all stations including the newly added one. The first class is *interpolation*, with full overlap between the new station's Voronoi shape and the system coverage. Second class is *extrapolation* with no overlap of the Voronoi shape and system coverage. And the last class is *margin* with partial overlap. Applying the above classification to all newly added stations identified in the previous step, the distribution of station types includes 262 (32.8%) interpolation cases, 68 (8.6%) partial interpolation cases, and 469 (58.6%) extrapolation cases, which are mapped in Figure 1 along with the original (pre-2015) set of stations. Using these three different categories, it is possible to apply and evaluate unique mechanisms to borrow flow information based on data and modelled relationships for each category.

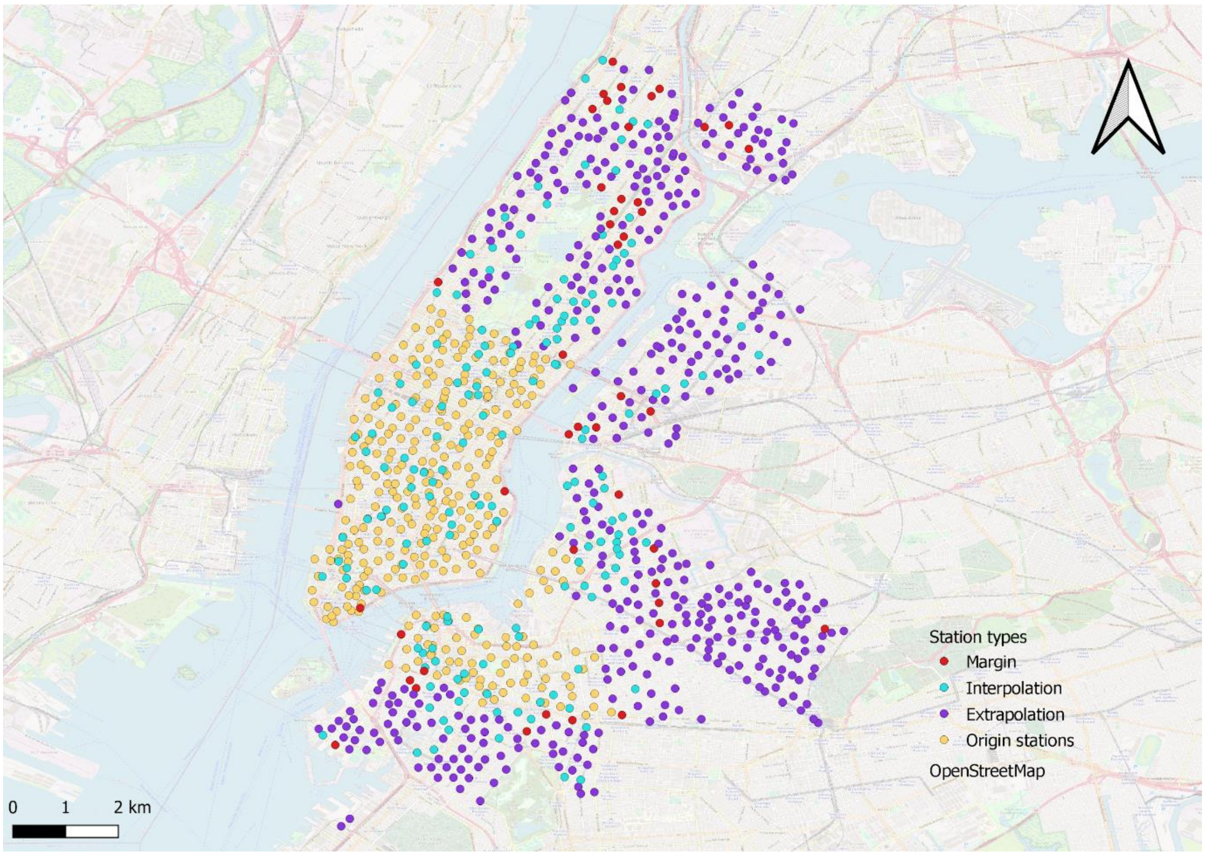

Figure 1. Map of distribution of bike stations color-coordinated by the three derived classes for newly added stations and the original stations.

2.3 Modelling strategies

There are two types of models to use. One is a generative model, such as the spatial interaction (SI) model, a regression model with non-flow attributes as input. Another type is dependence model, such as spatial interpolation model. It is a model borrowing data from nearby stations using spatial dependency.

Specifically, this study applied four sub-models under the two types. First is the unconstrained gravity-type spatial interaction model (Gravity SI). It is perhaps the most widely used model for diverse types of aggregate transport flows (for several recent examples see Kar *et al.* 2021; Lenormand *et al.* 2016; Oshan 2020; Zhou *et al.* 2020). It is calibrated here using a log-linear Poisson regression with a power distance-decay function and a set of origin/destination attributes using the *spint* module of the Python Spatial Analysis Library (PySAL) (Oshan 2016). Eleven categories were formed to aggregate POIs from SafeGraph (SafeGraph, 2020) and include *care*, *education*, *finance*, *food*, *housing*, *recreation*, *shopping*, *travel*, *professional*, and *other services*.

One well-established spatial interpolation method is the natural neighbor (NN) technique that finds the closest subset of input samples to a query point and weights them proportionally based on areal overlap to estimate a value (Sibson 1981). The natural neighbor interpolation method for points/polygons was extended to spatial interaction flows by applying a modified areal-based weighting scheme (Jang and Yao 2011).

While natural neighbor interpolation derives weights based only on location, kriging techniques derive weights based on both the location of S_x and value of each sample point $Z(s_i)$ with $i \in (1,2,\dots,n)$. It is an optimal linear estimator of the form,

$$Z(s_x) = \sum_{i=1}^{n} \alpha_i Z(s_i) \qquad (1)$$

where the weights α_i are chosen to make the estimator unbiased and of minimal prediction error. In this study, ordinary kriging (OK) and regression kriging (RK) based on Gravity SI are used as candidates of spatial interpolation methods.

The proposed methodology is then applied to the case of bike-sharing trips in New York City. Trip duration in seconds is used as the distance factor of SI models and trips with duration more than 3 hours are excluded as noise in the study. Individual trip records are aggregated with week level as trip OD flows with time label t. Training data consisted of flows one week before a new station was added at time t, (i.e., $t-1$) while the evaluation data was set to flows from two weeks after the roll out of a new station (i.e., $t+2$). To evaluate the performance of each prediction method for all S_x, Pearson's R correlation coefficient is employed.

3. Results

3.1 Prediction results

The prediction results based on the Pearson's R were recorded in Table 2 for the four methods (natural neighbor, ordinary kriging, regression kriging, gravity-type SI model) using data for each classification category (interpolation, margin, extrapolation), as well as the entire dataset. Overall, the results demonstrate that regression kriging is the most accurate model to capture the correlation in all three station types. The results also provide evidence in support of the primary hypothesis that different mechanisms in spatial interaction and spatial interpolation will be different depending on the location, as prediction performance of OK drops significantly from interpolation to extrapolation while Gravity SI has a smaller decrease in the correlation coefficient of the two classes of cold start stations.

Table 1. A summary of the prediction results based on three metrics for each method using data for each classification category and using all data.

Pearson's R	Interpolation	Margin	Extrapolation	All
Natural neighbor	0.51	0.31	-	-
Ordinary kriging	0.52	0.32	0.22	0.33
Regression kriging	0.55	0.40	0.37	0.43
Gravity SI	0.44	0.28	0.36	0.38

3.2 Spatial and temporal trend

Figure 2 first reveals the evolution of stations under the interval of a year. It shows the spatial and temporal trend of the two best methods, e.g., where the new station was added in each year and how the two methods work in the station flow estimation. Choropleth maps in Figure 2 are presented to show the prediction results of the best overall method, regression kriging and they have an uneven distribution. Specifically, interpolation stations in the middle of stations tend to have darker blue than the peripheral stations. Stations starting at Brooklyn (east NYC) in the 2020 snapshot are apparently less predictable than other stations added to the edge of the systems (margin and extrapolation). These stations have lower Pearson's R probably because they are further away from the core of the system than other stations.

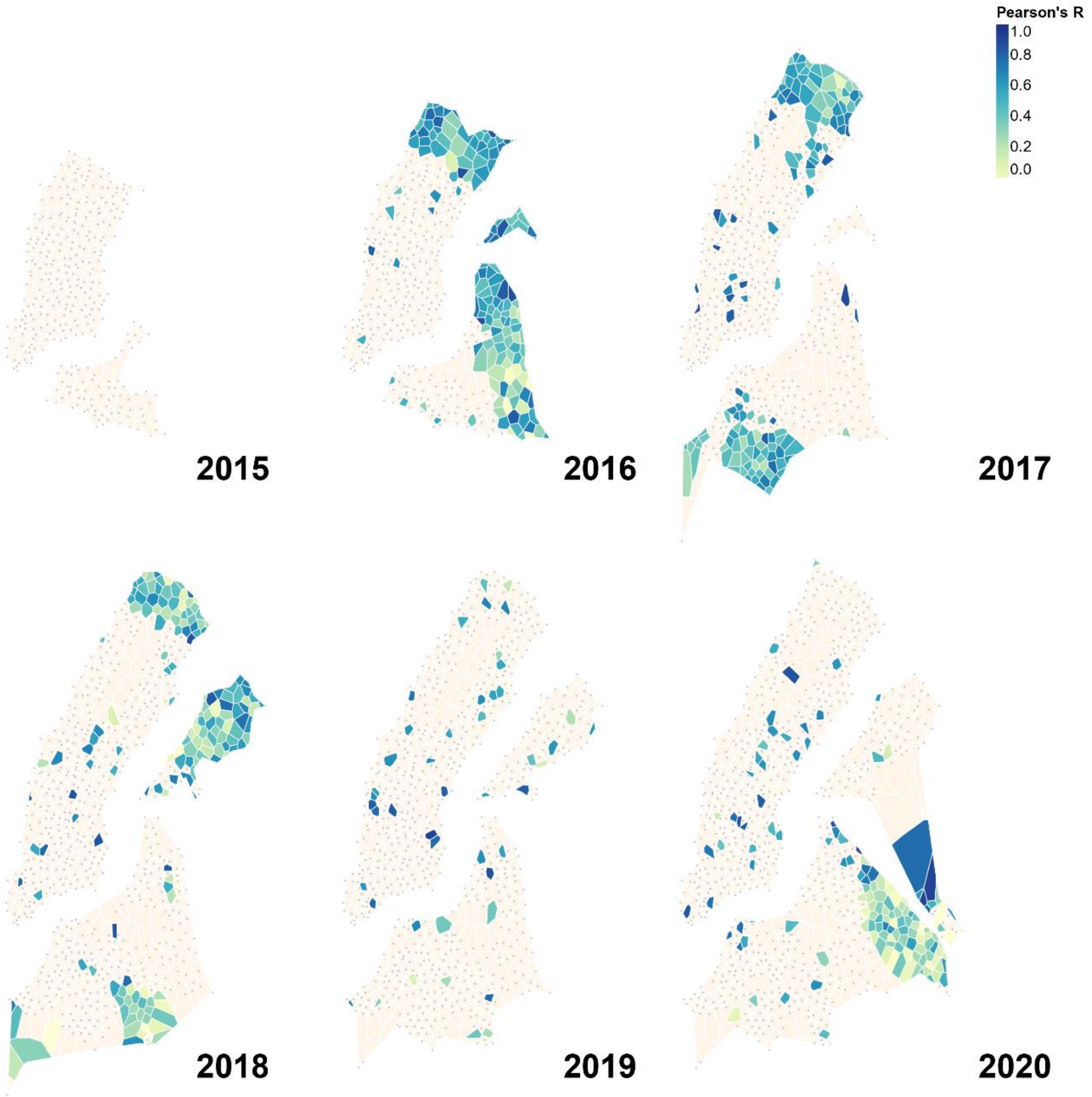

Figure 2. Pearson's R maps of Citi Bike stations snapshot at first week of the year from 2015 to 2020 using Regression kriging. Points are the actual location of dock stations, while the polygon is the Voronoi shape calculated from all the stations running at the time of snapshot. Stations that already exist before the first or previous snapshot are regarded as existing stations filled with light orange, while stations added between current and previous snapshot (except for the first snapshot 2015) will be regarded as added stations and rendered by the Pearson's R of corresponding model.

4. Discussion

Missing previous flows at new stations can be estimated from spatial interpolation and regression models from attributes. Furthermore, results indicated regression kriging works better than other candidate models. But modelling performance varies much across space and time in the same station classification. The methods included here only provide an initial investigation of the cold start issue. Additional methods as well as more data processing techniques that may increase the predictability, are anticipated to predict cold start station demand more adequately and ubiquitously.

References

DESA, 2018, *Revision of world urbanization prospects*. UN Department of Economic and Social Affairs, 16.

Faghih-Imani A, Hampshire R, Marla L and Eluru N, 2017, An empirical analysis of bike sharing usage and rebalancing: Evidence from Barcelona and Seville. *Transportation Research Part A: Policy and Practice*, 97, 177–191. https://doi.org/10.1016/j.tra.2016.12.007

Jang W and Yao X, 2011, Interpolating Spatial Interaction Data. *Transactions in GIS*, 15(4), 541–555. https://doi.org/10.1111/j.1467-9671.2011.01273.x

Kar A, Le HTK and Miller HJ, 2021, What is essential travel? Socio-economic differences in travel demand during the COVID-19 lockdown. *OSF Preprints*. https://doi.org/10.31219/osf.io/qtkhb

Lenormand M, Bassolas A and Ramasco JJ, 2016, Systematic comparison of trip distribution laws and models. *Journal of Transport Geography*, 51, 158–169. https://doi.org/10.1016/j.jtrangeo.2015.12.008

NYC Department of Transportation, 2019, *New York City Mobility Report*. https://www1.nyc.gov/html/dot/downloads/pdf/mobility-report-singlepage-2019.pdf

Oja P, Titze S, Bauman A, Geus BD, Krenn P, Reger-Nash B and Kohlberger T, 2011, Health benefits of cycling: A systematic review. *Scandinavian Journal of Medicine & Science in Sports*, 21(4), 496–509. https://doi.org/10.1111/j.1600-0838.2011.01299.x

Oshan TM, 2016, A primer for working with the Spatial Interaction modeling (SpInt) module in the python spatial analysis library (PySAL). *REGION*, 3(2), 11. https://doi.org/10.18335/region.v3i2.175

Oshan TM, 2020, Potential and Pitfalls of Big Transport Data for Spatial Interaction Models of Urban Mobility. *The Professional Geographer*, 72(4), 468–480. https://doi.org/10.1080/00330124.2020.1787180

Otero I, Nieuwenhuijsen MJ and Rojas-Rueda D, 2018, Health impacts of bike sharing systems in Europe. *Environment International*, 115, 387–394. https://doi.org/10.1016/j.envint.2018.04.014

SafeGraph (2020) The Impact of Coronavirus (COVID-19) on Foot Traffic. *U.S. Consumer Activity During COVID-19 Pandemic*.

Sibson R, 1981, A Brief Description of Natural Neighbor Interpolation. *Interpreting Multivariate Data* (p. pp 21-36). Chichester: John Wiley.

Tobler WR, 1970, A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, 46(sup1), 234–240. https://doi.org/10.2307/143141

Wang M and Zhou X, 2017, Bike-sharing systems and congestion: Evidence from US cities. *Journal of Transport Geography*, 65, 147–154. https://doi.org/10.1016/j.jtrangeo.2017.10.022

Zhang Y and Mi Z, 2018, Environmental benefits of bike sharing: A big data-based analysis. *Applied Energy*, 220, 296–301. https://doi.org/10.1016/j.apenergy.2018.03.101

Zhou T, Huang B, Liu X, He G, Gou Q, Huang Z and Xie C, 2020, Spatiotemporal Exploration of Chinese Spring Festival Population Flow Patterns and Their Determinants Based on Spatial Interaction Model. *ISPRS International Journal of Geo-Information*, 9(11), 670. https://doi.org/10.3390/ijgi9110670