Check for updates

DOI: 10.1111/tgis.12933

RESEARCH ARTICLE

Transactions in GIS WILEY

Comparing spatial interaction models and flow interpolation techniques for predicting "cold start" bike-share trip demand

Zheng Liu | **Taylor Oshan**

Department of Geographical Sciences, University of Maryland, College Park, Maryland, USA

Correspondence
Zheng Liu, Department of Geographical Sciences, University of Maryland, 4600 River Road, Riverdale Park, MD 20737, USA.
Email: zliu1208@umd.edu

Abstract

Bike-sharing systems are expanding rapidly in metropolitan areas all over the world and individual systems are updated frequently over space and time to dynamically meet demand. Usage trends are important for understanding bike demand, but an overlooked issue is that of "cold starts" or the prediction of demand at a new station with no previous usage history. In this article, we explore a methodology for predicting the bike trips from and to a cold start station in the NYC Citi Bike system. Specifically, gravity-type spatial interaction model and spatial interpolation models, including natural neighbor interpolation and kriging, are employed. The overall results come from experiments of a real-world bike-sharing system in NYC and indicate that the regression kriging model outperforms the other models by taking advantage of the robustness and interpretability of gravity-type spatial interaction regression models and the capability of ordinary kriging to capture spatial dependence.

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 (United Nations, Department of Economic and Social Affairs, Population Division, 2019). As a result, it has become increasingly important to understand and anticipate human mobility within and

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2022 The Authors. *Transactions in GIS* published by John Wiley & Sons Ltd.

across cities. In particular, public transportation systems, historically composed of buses, trains, and taxicabs, form the backbone of cities, facilitating interaction and providing an affordable and reliable alternative to carry out routine activities compared to private transportation. The development and optimization of public transportation infrastructure has therefore remained an important aspect of urban planning and community development.

A more recent trend is the emergence of bike-sharing as an alternative means of transportation and the establishment of bike-sharing systems in cities around the world. 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, Nieuwenhuijsen, & Rojas-Rueda, 2018; Wang & Zhou, 2017; Zhang & 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, Hampshire, Marla, & Eluru, 2017). This has led to the proliferation of bike-sharing systems, with nearly 2000 bike-sharing systems now in operation around the world.[1]

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. This dynamic development of bike-sharing systems requires a strong understanding of mobility behavior and flexible methods for predicting spatial-temporal demand. In particular, travel demand models may focus on predicting overall system demand, the total demand at each location or station within a system, or for demand between individual locations. It is this later scenario that is the primary focus of this research as it typically receives less attention (e.g., Li & Shuai, 2020; Zhou, Chen, et al., 2019). This is perhaps because it is more challenging to model individual origin-destination (OD) flows due to a lack of detailed data and because they usually contain more noise and higher levels of variability compared to the overall system usage or the total inflows or outflows for a set of locations. However, models of OD flows, often referred to as spatial interaction models, are becoming increasingly possible as detailed data are generated by GPS-enabled sensors and the internet-of-things (Shaw & Sui, 2018). For example, Calafiore, Palmer, Comber, Arribas-Bel, and Singleton (2021) leveraged user OD flows based on social media check-ins to study the characteristics of cities' neighborhoods. Furthermore, recent machine learning methods for incorporating various dimensions of spatial and temporal dependence hold promise for building models with enhanced predictive capabilities (Chu, Lam, & Li, 2020; Ke et al., 2019; Liu et al., 2019; Wang et al., 2019). Consequently, this work investigates a methodology for predicting OD flows of bike-share systems given that the system evolves over time and demand is constantly changing. 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. Therefore, Section 2 of this article provides some additional background and formalizes the challenges that need to be overcome. Then, Section 3 presents a novel methodological approach for predicting the spatial interaction demand associated with new bike-share stations and Section 4 briefs the data used in the case study and experiment settings. Section 5 describes the experimental results used to benchmark the proposed approaches. Overall, the results indicate that regression kriging models are slightly more performant than the other techniques considered. Finally, Section 6 concludes with a discussion of the contributions of this research, limitations of the proposed approach, and suggestions for future work in this area.

14679671, 2022, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tgis.12933 by University Of Maryland, Wiley Online Library on [21/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

2 | BACKGROUND

2.1 | Bike trip demand modeling

The widespread adoption of information and communication technology has produced massive amounts of transportation and mobility data, sparking a wave of research into how to best model human movement. Meanwhile, the popularity of bike-sharing systems has inspired much recent research. Si, Shi, Wu, Chen, and Zhao (2019) provide a review of bike-sharing papers published from 2010 to 2018 and group them based on the following categories: (1) demand factors; (2) rider behavior; (3) system optimization; and (4) impact on other modes. The factors associated with variation in bike-sharing ridership demand can largely be summarized as being related to the weather, built environment, sociodemographic, or temporal dimensions (Barbour, Zhang, & Mannering, 2019; El-Assi, Salah Mahmoud, & Nurul Habib, 2017; Eren & Uz, 2020; Guo, Zhou, Wu, & Li, 2017; Yang, Zhang, Zhong, Zhang, & Ling, 2020). Knowledge of the distribution of trips is useful for developing strategies to optimize systems through bike rebalancing or station repositioning, allowing the system to satisfy higher demand (Chen et al., 2020; Faghih-Imani et al., 2017; Pan, Cai, Fang, Tang, & Huang, 2019; Zhu, Zhang, Kondor, Santi, & Ratti, 2020). The introduction of a bike-sharing system may impact other transport modes, leading researchers to compare bike share usage to taxi and bus ridership (Campbell & Brakewood, 2017; Zhou, Wang, & Li, 2019). Another trend focuses on bike-share trip prediction using travel demand models and OD flow models, the latter of which is the focus of this article and will be used to consider a range of factors influencing bike-share trip demand to predict OD flows.

Trip demand at cold start stations, however, catches less attention and is only explicitly discussed in Noland, Smart, and Guo (2016) and Zhang, Thomas, Brussel, and van Maarseveen (2017). And both papers used linear regression models on station-level demand to examine the generalization of the models. Research by Wang, Cheng, Trépanier, and Sun (2021) discussed the use of a regression model for making predictions at new stations but only performed out-of-sample predictions by randomly selecting stations as a test set, which is not an exhaustive validation for trip demand at potential new stations. Furthermore, there is a lack of effort dedicated to *OD flow* predictions for new stations.

2.2 | Spatial interaction

Spatial interaction (SI) describes aggregate movements of individuals, commodities, capital, or information over geographic space, resulting from some requisite decision-making process (Farmer & Oshan, 2017). A typical quantitative representation of SI is the origin-destination (OD) matrix:

$$\mathbf{T} = \begin{bmatrix} T_{11} & \cdots & T_{1j} & \cdots & T_{1m} \\ \vdots & & \vdots & & \vdots \\ T_{i1} & \cdots & T_{ij} & \cdots & T_{im} \\ \vdots & & \vdots & & \vdots \\ T_{n1} & \cdots & T_{nj} & \cdots & T_{nm} \end{bmatrix} \begin{matrix} O_1 \\ \vdots \\ O_i \\ \vdots \\ O_n \end{matrix} \\ \begin{matrix} D_1 & \ldots & D_j & \ldots & D_m \end{matrix} \quad T \tag{1}$$

where T is the OD flows matrix from a set of origins $(O_1, O_2, \ldots, O_n)$ to a set of destinations $(D_1, D_2, \ldots, D_m)$, and T_{ij} is the magnitude of flows from origin i to destination j. Inspired by Newton's law of gravity, early models of spatial interaction theorized OD flows to be proportional to the product of potential at origins and destinations and inversely proportional to the squared distance between them, yielding the following formulation:

14679671, 2022, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tgis.12933 by University Of Maryland, Wiley Online Library on [21/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

$$T_{ij} = k\frac{P_i P_j}{d_{ij}^2} \tag{2}$$

where T_{ij} still denotes the flows between origin i and destination j, P_i and P_j represent the potential at i and j, d_{ij} describes the distance between i and j, and k is a scaling factor that ensures the number of flows predicted by the model matches the number of observed flows. In this context, potential is often defined as location population or the number of job opportunities (Gao, Liu, Wang, & Ma, 2013; Krings, Calabrese, Ratti, & Blondel, 2009; Lenormand, Bassolas, & Ramasco, 2016), but can be expanded to consider a series of origin and destination attributes that may contribute toward the flow generation, each with their own parameter (Fotheringham & O'Kelly, 1989; Oshan, 2020). According to Kar, Le, and Miller (2021), two groups of variables can be identified besides population or jobs: socioeconomic and built environment attributes. The former group usually explains trip generation, such as labor force (Pourebrahim, Sultana, Thill, & Mohanty, 2018; Signorino et al., 2011); GDP (Zhang, Cheng, & Jin, 2019) and human activity density (Marrocu & Paci, 2013). The latter are common destination determinants of travel demand, such as amenities (e.g., school, hospitals, markets) (Botella, Gora, Sosnowska, Karsznia, & Querol, 2021; Kar et al., 2021), tourism attractions (e.g., hotel rooms) (Khadaroo & Seetanah, 2008), and land-use types (Liu, Kang, Gong, & Liu, 2016).

2.3 | Dynamic demand and spatial-temporal dependence

Recent work seeks to exploit temporal dependence within SI flows to increase predictive performance. The simplest instance involves taking the average of historical observations, which can yield accurate predictions for future flows when there is limited variation in the historical observations. Typical methods using temporal dependence include historical averaging (HA); autoregressive integrated moving average (ARIMA); deep neural network structures including recurrent neural networks (RNN) and long short-term memory (LSTM) architectures (Cheng, Trepanier, & Sun, 2021; Chu et al., 2020; Ke et al., 2019). The common factor amongst all these methods is that they require sufficient previous OD flows and become invalid to predict future flows when there are no historical observations to draw upon.

Meanwhile, bike-share systems are also spatially dynamic in that stations are frequently updated or perhaps more importantly new stations are added to extend the system, though this is often overlooked in much related work. One exception is Lu, Hsu, Chen, and Lee (2018) who explored the extension of system infrastructure by deploying agent-based methods to simulate the result of adding new stations toward the usage of different transportation modes. In contrast, most previous research focuses on predicting bike-sharing trip flows at existing locations (i.e., in-sample spatial prediction), often using historical data as an important input feature. This is problematic when the focus is instead on predicting flows associated with a new station that is being added to the system (i.e., out-of-sample spatial prediction) because there are no historical flow data that can be used to learn temporal dependencies. A similar issue often occurs in recommender systems where new users or new items will have no historical record to be used for preference analysis (Su & Khoshgoftaar, 2009; Volkovs, Yu, & Poutanen, 2017) and is often called the "cold start" issue. This term is also used in the literature associated with detecting stops and trips from GPS trajectories (Schuessler & Axhausen, 2009; Stopher, Jiang, & FitzGerald, 2005). Overcoming this limitation is the primary contribution of this research in order to develop a robust methodology for predicting historical OD flows at a new bike-share station.

It is reasonable to leverage spatial dependence to predict values at unobserved locations based on values from nearby observed locations. The First Law of Geography states that things that are closer together are typically more similar (Tobler, 1970) and is the basis for popular spatial statistics, such as the Moran's I measure of spatial autocorrelation and spatial interpolation methods, including natural neighbor interpolation, inverse distance weighting, Kriging, and more (Mitas & Mitasova, 1999). In particular, kriging has become a core spatial interpolation tool and is now used in many topic areas such as air quality analysis (Bayraktar & Turalioglu, 2005), natural resource

14679671, 2022, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tgis.12933 by University Of Maryland, Wiley Online Library on [21/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

analysis (Emery, 2005), water studies (Zimmerman et al., 1998) and traffic (Eom, Park, Heo, & Huntsinger, 2006; Wang & Kockelman, 2009). Recently, spatial interpolation tasks have been adapted into lattice or graph structures and integrated into generative adversarial networks (Zhu, Cheng, et al., 2020) or graph neural networks (Wu, Zhuang, Labbe, & Sun, 2020), but these extensions do not yet apply to the case of OD flows. For the interpolation of OD flows, Jang and Yao (2011) proposed an areal weighting method, which was further employed by Šimbera and Aasa (2019). As far as the authors are aware there are no studies extending interpolation methods to SI flows. However, the spatial dependence between SI flows have previously been leveraged for community detection (Gao et al., 2013; Yao et al., 2018) and land use identification (Liu et al., 2016), indicating that kriging, which also relies on the presence of spatial dependence, may be promising for flow interpolation. Therefore, this research explores flow-based kriging and compares it to natural neighbor interpolation and gravity-type spatial interaction models.

In the next section, the central issues and challenges are described more formally. Next, a methodology is outlined to compare flow prediction strategies, which is then applied to the case study of bike-share trips in New York City.

3 | METHODOLOGY

3.1 | 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 modeling procedures. A cold start refers to 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/or future inflows $T_{t+1,i,x}$. The primary issue that arises is that there are no previous flows to use in any of the methods that leverage historical data. A methodology is proposed below to overcome this limitation by classifying stations and using a combination of regression modeling and interpolation techniques depending on the station class.

3.2 | Station classification

Predicting the flows associated with the addition of new stations is the primary task of this research. However, depending on the location of the new station and its proximity to currently existing stations, different techniques and information are available to carry out the predictions. At least three different scenarios can be distinguished. The first is referred to here as interpolation, which entails borrowing information directly from the existing stations. Interpolation typically only applies in the situation where the newly added station is sufficiently proximal to existing stations (i.e., within current system coverage). The second scenario is referred to here as margin interpolation and is the case where a new station is added to the margin of the current system. As such, there are some nearby stations with the previous flow information that can be borrowed. The final scenario is referred to here as an extrapolation and is concerned with the addition of stations that are essentially outside the coverage of the current system. Predictions for this scenario are hypothesized to be only possible through the extrapolation of modeled relationships, since there are no existing nearby stations to borrow information from. A method for identifying empirical instances of system expansion and classifying new candidate stations across the three scenarios is proposed below and then subsequently deployed and evaluated.

14679671, 2022, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tgis.12933 by University Of Maryland, Wiley Online Library on [21/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Newly added stations are identified by examining the time series of trip data for the bike share system. For each station present in the system before July 2020, an array of station inflows (or station outflows) for each temporal subset (weekly in this case study) is used to track its operation status. Intuitively, the first non-zero entry of the array, $t_{in} > 0$, is recorded as a preliminary inferred date that a station was initially put into operation. However, these preliminary dates are then refined by manually checking for consistent station service as in practice there sometimes are test trips prior to the official launch of a new station. In addition, the system record indicates that all new stations added after the initial system rollout in 2013 did not occur until 2015. Therefore, the original stations existing before 2015 (Table 1) are not included in the set of classified stations nor are they used for prediction validation.

Next, these new stations, S_x, are classified between interpolation, margin, and extrapolation. This is done based on their relationship to the existing system coverage, which is represented here by computing the convex hull of all the stations in operation at the time S_x is introduced. The convex hull of all the station points provides a simplified representation of the system service area in continuous Euclidean space, though the urban environment does not exist separately from the natural environment (i.e., rivers). To ensure these restricted areas are not included in the representation of system coverage, a few extra steps were taken to prepare a modified convex hull for the bike share system. First, stations on Governor's Island and a single isolated station in the southern part of Brooklyn were removed from the analysis because these stations are essentially disconnected from the larger system. Then, multiple convex hulls were created separately for stations in Manhattan and stations in Brooklyn/Queens, ensuring that neither of the convex hulls cross into major waterways.

For every temporal subset t, only the system coverage associated with stations running at time $t - 3$ is used to classify the new station S_x initialized at time t where each temporal subset pertains to a week within the period from January 2015 through July 2020. This time lag of up to 3 weeks ensures that stations rolled out consecutively over a relatively short time period are not considered as previously existing to each other and is necessary because there is usually a short period of time before a new station becomes fully integrated within the system. If S_x falls outside this current system coverage, it will be regarded as an extrapolation case. In contrast, if S_x is within this current system coverage, it will be regarded as an interpolation case. It is then necessary to further distinguish the partial interpolation cases, which correspond to stations added to the margin of the current system coverage. The three cases are formally classified by the percentage of overlap between their Voronoi tessellation cell and the current system coverage. Stations overlapping < 5% are classified as an extrapolation station while

TABLE 1 Number of stations added over time by month and year

Month and year	2013	2015	2016	2017	2018	2019	2020
January	–	–	2	2	9	6	8
February	–	–	2	8	2	3	5
March	–	–	2	4	6	6	1
April	–	–	2	4	3	17	2
May	–	–	3	5	7	11	36
June	336	–	5	4	5	5	35
July	1	1	8	4	4	3	–
August	–	85	98	4	5	4	–
September	–	39	43	64	1	15	–
October	–	16	2	74	3	38	–
November	–	5	2	6	3	34	–
December	–	2	2	4	4	15	–

14679671, 2022, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tgis.12933 by University Of Maryland, Wiley Online Library on [21/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

those overlapping > 95% are classified as an interpolation station. The remaining stations are classified as margin stations.

Applying the above classification to all newly added stations identified in the previous step. Using these three different categories, it is possible to apply and evaluate unique mechanisms to borrow flow information based on data and/or modeled relationships for each category. The hypothesis employed here is that different mechanisms will be more efficient for each category because different types and levels of information are available. In particular, it is anticipated that better predictive performance will be attained for stations classified as extrapolation cases using a SI modeling approach because borrowing information for data from relatively far stations could introduce a disproportionately large amount of misinformation. In contrast, the interpolation and margin cases are anticipated to achieve higher predictive performance using methods that incorporate data-borrowing through interpolation techniques. That is, flows for stations classified as interpolation and margin are expected to be best predicted using interpolation methods whereas those that are not must rely on extrapolation through modeled relationships (i.e., SI model). However, it is less clear whether or not the same method will be the most effective for both interpolation and margin stations, since different levels of information are available. To investigate these issues, several interpolation methods are explored in the following sections and then subsequently compared to each other and to predictions from SI models.

3.3 | Modeling strategies

3.3.1 | Gravity-type spatial interaction model

The unconstrained gravity-type spatial interaction model (Gravity SI) described in Section 2.3 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). Following Eren and Uz (2020), these attributes include points of interest for urban amenities and services, transportation infrastructure, and sociodemographic factors. For transportation infrastructure, accessibility to other bike stations is computed based on a distance-weighted sum (i.e., spatial lag) of nearby bike station capacity.[2] POI data were obtained from SafeGraph (2020) and contain a hierarchical schema. Eleven categories were formed to aggregate POIs from this schema and include *care*, *education*, *finance*, *food*, *housing*, *recreation*, *religious*, *shopping*, *travel*, *professional*, and *other services*. Further descriptions of data sources and variables are summarized in Table 2. The selection process of distance parameters for station accessibility and POIs is based on trial and error, however, sensitivity results show there are no significant differences in gravity SI results when using other distance bands (e.g., 500 m or 1000 m) to compute these variables.

3.3.2 | Natural neighbor interpolation

Spatial interpolation methods predict unknown values for an unsampled point in space based on its relationship to a set of sampled points. Given values of a random field $Z(\cdot)$ measured at locations $s_1, \ldots, s_n$ yielding $Z(s_i)$ for all i, the objective is to estimate the value $Z(s_x)$ for one or more unmeasured locations s_x. One well-established spatial interpolation method is the natural neighbor 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). Based on the assumption that nearby stations typically experience similar levels of demand, it is possible to borrow data for a new station from neighboring stations using this method. The natural neighbor interpolation method for points/polygons was extended to spatial interaction flows by applying a modified areal-based weighting scheme (Jang

14679671, 2022, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tgis.12933 by University Of Maryland, Wiley Online Library on [21/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 2 Attributes used in the spatial interaction models

Attributes	Source	Feature description
Population	Cenpy Census.gov	Population (2017 ACS5) of each census tract
Employment	Longitudinal Employer-Household Dynamics (LEHD)[a]	Workplace Area Characteristics—census block
Station accessibility	Station Info[b]	Weighted sum of nearby station capacity with inverse distance weights with a distance band of 1500 m
Access to subway	NYC Open Data[c]	Euclidean distance to the nearest subway station
Points of interest	SafeGraph[d]	Weighted sum of nearby POIs with inverse distance weights with a distance band of 1500 m

[a]https://lehd.ces.census.gov/data/.
[b]https://gbfs.citibikenyc.com/gbfs/en/station_information.json.
[c]https://data.cityofnewyork.us/Transportation/Subway-Stations/arq3-7z49.
[d]https://www.safegraph.com/covid-19-data-consortium.

& Yao, 2011). Equations (3.1) and (3.2) present the natural neighbor formula for calculating the flow T_{xj} between new station S_x and a destination station S_j. First, Voronoi tessellations are computed for S_n with and without S_x, denoted V_{n+x} and V_n, respectively, with each individual Voronoi polygon approximating the service area for each station. In Equation (3.1), m is the number of stations overlapping with S_x, p_k is the proportion of the overlapping area between Voronoi shape of S_x in V_{n+x} and S_k in V_n in relation to the total area of the Voronoi shape of S_x in V_{n+x} (Equation 3). The resulting $[T_{x1}, T_{x2}, \ldots, T_{xn}]^\top$ are the "borrowed" outflows of S_x. Thus, natural neighbor interpolation provides an areal-based method that can be used on flow data and will be one of the interpolation methods applied here.

$$T_{xj} = \sum_{k=1}^{m} [p_k \times T_{kj}] \tag{3.1}$$

$$p_k = \frac{A_{kx}^{V_{n+x} \cap V_n}}{A_x^{V_{n+x}}} \tag{3.2}$$

3.3.3 | Kriging

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 $i \in (1, 2, \ldots, n)$. It is an optimal linear estimator of the form:

$$Z(s_x) = \sum_{i=1}^{n} \alpha_i Z(s_i) \tag{4}$$

where the weights α_i are chosen to make the estimator unbiased and of minimal prediction error. Kriging is a two-step process. First, the spatial covariance structure of the sampled points is determined by fitting a variogram (a spherical model is used here) that captures the variability between all data points as a function of distance. Then, weights derived from this covariance structure are used to interpolate values for unsampled observations across the spatial field. Ordinary kriging assumes the random field $Z(\cdot)$ is intrinsically stationary; that is for any location s:

$$E[Z(s)] = \mu \tag{5.1}$$

14679671, 2022, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tgis.12933 by University Of Maryland, Wiley Online Library on [21/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

$$Var\left[Z(s) - Z(s+h)\right] = 2\gamma(|h|) \tag{5.2}$$

where $\gamma(|h|)$ is the semi-variogram and a function of distance h that separates two locations. Further mathematical details are available in Cressie (1993).

Ordinary kriging typically uses a k-dimension (usually $k \leq 3$) points as an input location and a 1-dimension target value Z. However, a spatial interaction flow between stations is a spatial process involving two points in a two-dimension euclidean space ($k = 4$). To accommodate this higher dimensionality, a flow-based extension inspired by Zhou, Chen, et al. (2019) is proposed that aggregates all the outflows to the m possible destination stations or inflows from m stations to the single-origin station s_i as $Z(s_i)$ and the definition of T_{ij} remains the same as in Equation (1). The coordinates of the origin station remain in 2-D space, but the target value Z_i is an m-length vector representing the flows to each destination. According to Liu et al. (2016), this station-centric view, which aggregates the OD flows vector coming in and going out from one single station, is also referred to as the OD flow signature of a station. In this station-centric view, the adoption of the vectorized Z is needed for handling higher-dimensional Z within kriging techniques. The inflow predictions Z_{in} and outflow predictions Z_{out} are carried out separately and then concatenated as the overall OD predictions for a target station.

Alternatively, regression kriging leverages information from exogenous variables in addition to the spatial dependence in a sample by first fitting a regression model and then kriging the residuals of the regression model. Therefore, a gravity-type SI model is first fit and then a flow-based ordinary kriging (OK) model as described above is used to interpolate the residuals of the gravity-type SI model. Specifically, when using the station-centric view in the regression kriging model (RK), the kriging step uses 2-dim input with k-dim targets, while the regression step uses location data in OD flow form with 4-dim features.

To better highlight the difference between the kriging interpolation and natural neighbor interpolation, an illustration with pseudo data is provided in Figure 1 to demonstrate the two interpolation processes. Both methods take identical data structure as input data, for example, existing locations S_n, a cold start location S_x and flow matrix among existing stations T. The bottom-left section shows the steps used for areal-based natural neighbor interpolation. First, the service area of each station defined by the Voronoi shapes is calculated before and after a cold start station is added. Then the area overlapping the cold start station is calculated at each existing station, consisting of the portion (p) of new stations which is the final proportional weights to calculate the interpolated outflows and inflows at the new station shown in the last row. In contrast to the natural neighbor that exploits locations only, ordinary kriging interpolation leverages both locations and flow similarity. Under the assumption of ordinary kriging, variations between stations at a fixed range of distance rely on the distance h. Thus, distances between each pair of origins and destinations are grouped into bins (Right-bottom section, Figure 1 left) and the variance in each group is calculated and plotted (Right-bottom section, Figure 1 right). The optimization equation can be solved using the semivariogram only, as shown in geostatistical textbooks such as Le and Zidek (2006). So, it is essential to interpolate the flow variance between new stations and any existing stations by fitting a covariance function (black solid line in right-bottom section, Figure 1 right). Finally, the optimized alpha that minimizes the prediction error is calculated and used for the flow interpolation results just as the natural neighbor interpolation method.

4 | DATA

4.1 | Study area

The Citi Bike system in New York City started operation in 2013 and is now the largest bike-share system in the U.S. Among the 809 stations added to the system since 2015, 55 of them had too few trips to produce reliable output in any one of the methods, specifically producing a null or negative correlation index value. Thus, the

14679671, 2022, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tgis.12933 by University Of Maryland, Wiley Online Library on [21/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

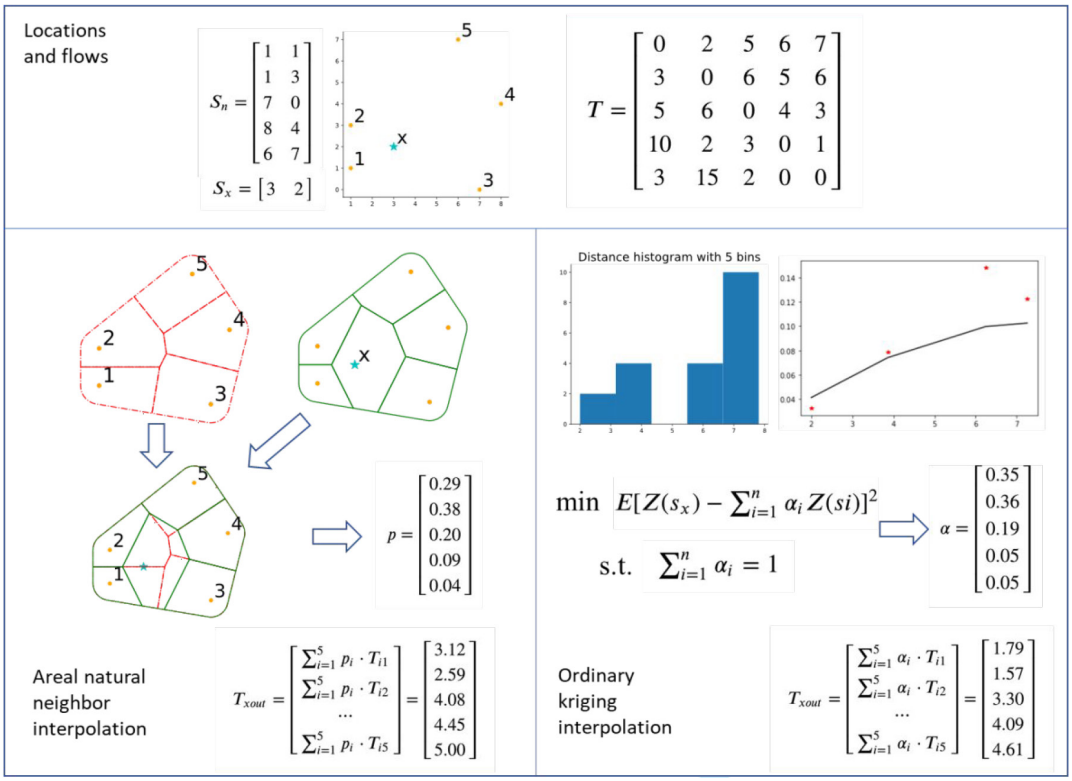

FIGURE 1 Illustration of areal natural neighbor interpolation and ordinary kriging interpolation with pseudo data

following results come from the remaining sample of stations ($n = 754$) with 313 interpolation stations, 123 margin stations, and 318 extrapolation stations using the (0.05, 0.95) classification threshold. The spatial distribution of the stations is mapped in Figure 2 along with the original (pre-2015) set of stations. All the trips are aggregated into unique spatial and time tuples by their original stations, destination stations and the week of the trip starting from January 1, 2015. In the regression-based models, the following expression is used: *"flows ~ cost + origin attributes (15) + destination attributes (15)"* Origin and destination attributes include 15 variables, each are listed in Table 2, where POI includes 11 categories, for example, *care*, *education*, *finance*, *food*, *housing*, *recreation*, *religious*, *shopping*, *travel*, *professional*, and *other services*. In the interpolation-based models, the OD-matrix is formatted as shown in Figure 1.

4.2 | Study design

Each new station is associated with a unique time frame for training and testing. So iterating station-wise, SI models were trained for each added station S_x, with a training set including all of the OD flows available in the time period before S_x is added, and with a prediction test set from the time period after S_x is rolled out. Training data consisted of flows 1 week before a new station was added at time t, (i.e., $t - 1$), while the evaluation data were set to flows from 1 week after the rollout of a new station (i.e., $t + 1$). This 1-unit period provides a chance for demand to stabilize after a new station is added to the system with a complete week. Bike trips with a duration of more than 3 h are excluded as abnormal trips. To evaluate the performance of each prediction method for all S_x, Pearson's R correlation coefficient is employed.[3]

14679671, 2022, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tgis.12933 by University Of Maryland, Wiley Online Library on [21/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

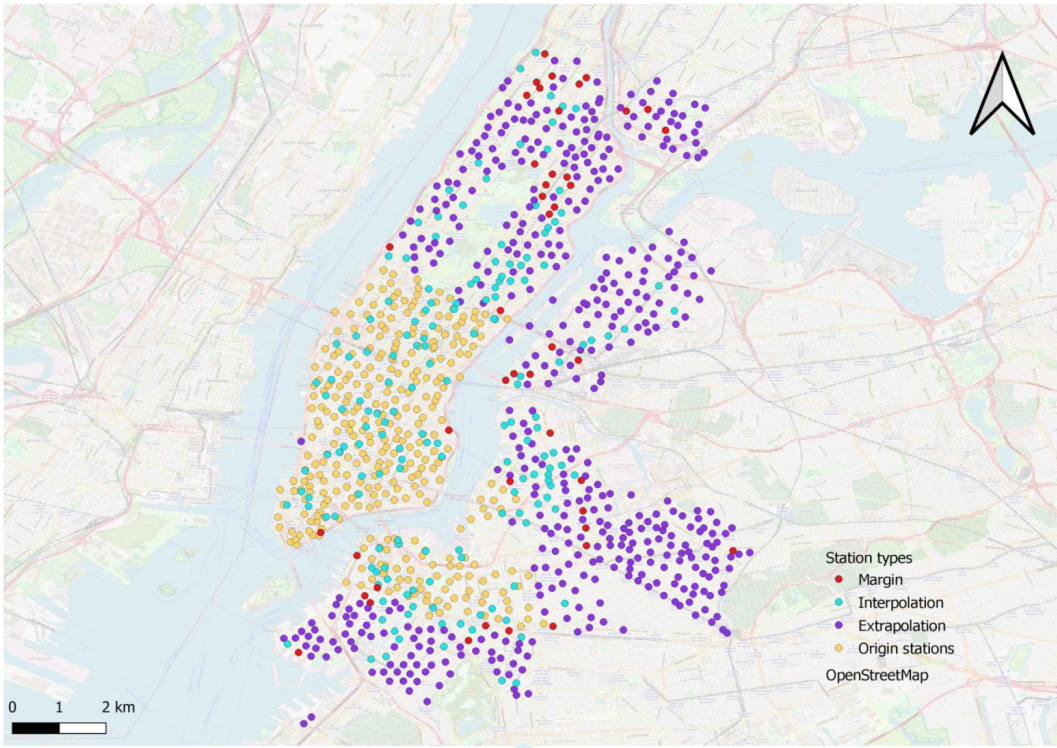

FIGURE 2 Spatial distribution of bike stations color-coordinated by the three derived classes for newly added stations and the original stations

5 | RESULTS

5.1 | Gravity-type spatial interaction model calibration

Before producing the prediction results, it is necessary to calibrate the gravity-type SI models.

An advantage of the SI model over the interpolation methods is the ability to elucidate how independent variables affect bike trip flows based on the parameter coefficients. To highlight the weight of each independent variable, the top of Figure 3 shows the average parameter magnitude and 95% confidence intervals based on the models for each station. The bottom of Figure 3 captures the top 11 most important features over time. Generally, parameter estimates associated with a variable have the same direction, either negative or positive, and similar magnitudes for the origin stations and the destination stations. The cost factor, which is the average trip completion time of the OD pair and typically called the distance-decay effect, is the most important factor, attesting to the First Law of Geography. The next most important factors included *recreation* and *other services* POIs. The former POI group suggested the purpose of the bike-sharing trips for leisure activities. The latter group, *other services*, includes many parking facilities. It could be interpreted as a potential common commuting pattern: people park the car and then take the bike trip. Surprisingly, population was not a top factor, suggesting the residential population around stations is not necessarily a strong indicator of the size of the potential bike-share user base. In terms of the parameter estimate stability over time, the top 11 important factors typically have consistently positive or negative signs. Some general trends can also be observed over time. First, the distance-decay factor associated with trip duration became more negative (i.e., stronger) over time as the system expanded. There is also a periodic fluctuation that captures the seasonal trend where distance-decay is less negative in the winter (the spikes around 50th, 110th, 210th weeks). Second, the impact of the COVID-19 lockdown was a decrease in the

14679671, 2022, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tgis.12933 by University Of Maryland, Wiley Online Library on [21/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

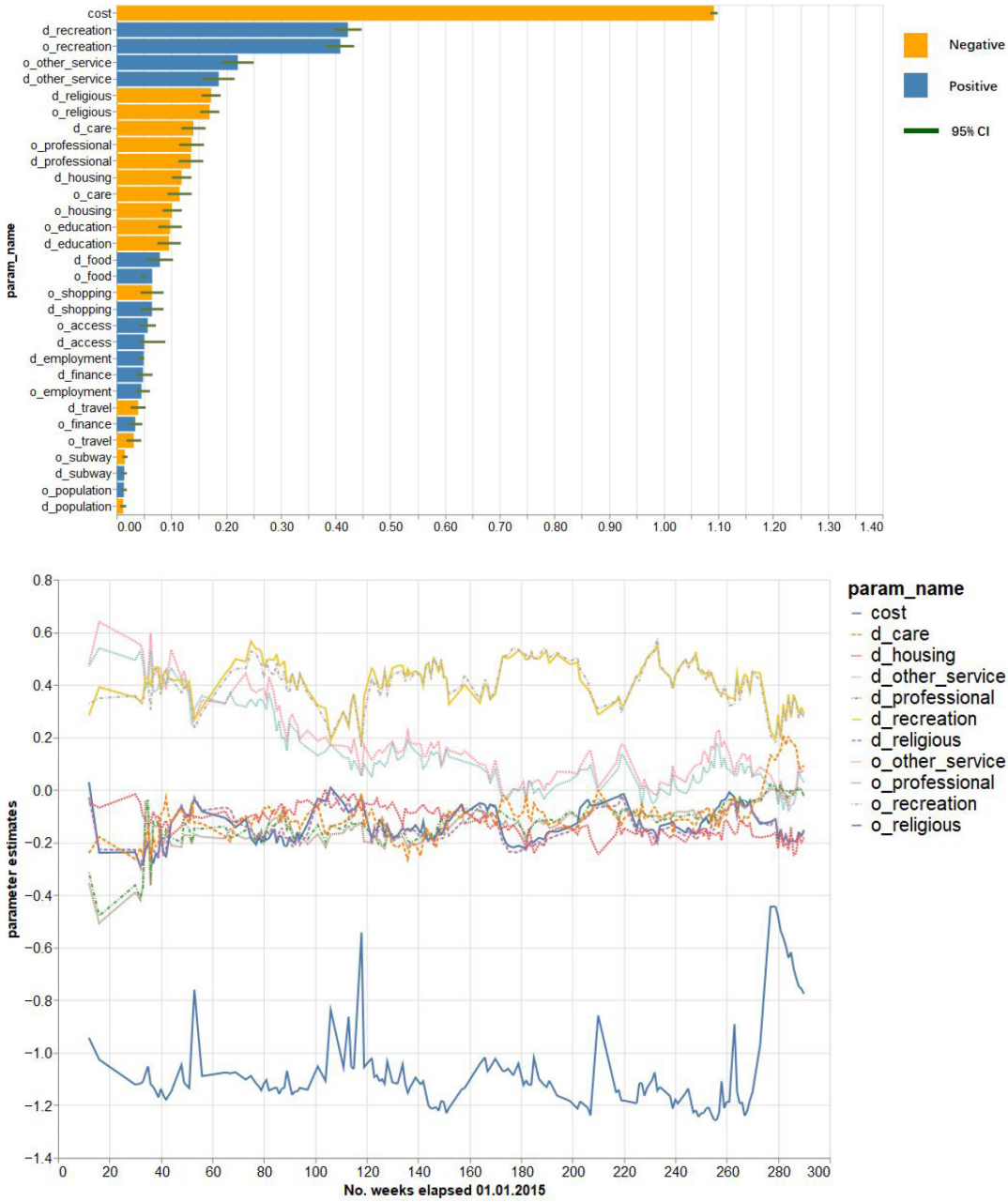

FIGURE 3 (Top) Feature weight from the SI model results averaged over the time periods. The *X* axis represents the absolute value of the parameter estimates and the color represents the sign of the average magnitude. The dark green strip shows the average range of 95% confidence interval. (Bottom) Time series of top 11 largest feature weights. The *X* axis is the number of weeks elapsed since January 1, 2015

magnitude for the parameter estimates of most factors, likely because the regular trips of public were essentially halted during this time and the alternative behavior was either less associated with these factors or generally more random. Besides, the gravity SI model also showed robustness in the parameter importance when changing the distance band from 1500 to 1000 m to compute spatially lagged explanatory variables (i.e., POIs and accessibility), providing that the direction and magnitude of coefficients remain the same over the two distances.

14679671, 2022, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tgis.12933 by University Of Maryland, Wiley Online Library on [21/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

5.2 | Prediction results

The prediction results based on the correlation index Pearson's R between predicted and actual trip counts were recorded in Table 3 for the five methods (natural neighbor, ordinary kriging, regression kriging, gravity-type SI model, and negative binomial model) using data for each classification category (interpolation, margin, extrapolation), as well as the entire dataset. Overall, the results demonstrate that regression kriging is probably the most well-rounded 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 and the regression kriging can outperform the single methods of ordinary kriging or SI model.

Meanwhile, the standard deviations around 0.2 indicate the correlation varies much across stations. It is hard to say regression kriging can always outperform other methods.

5.3 | Spatial trends

The relative comparison of the candidate methods within one location is more meaningful than the quantitative comparison of metrics across stations. Figure 4 reveals the spatial distribution of cold start stations rendered with the best method in the individual station. Warm colors are associated with interpolation methods (Red: Natural Neighbor; Orange: Ordinary Kriging) and cold colors stand for regression methods (Cyan: Gravity SI; Blue: Negative Binomial). Green dots represent the compound model: regression kriging method. Overall, there is a substantial portion (518 of 754) of interpolation methods winning the best method over the two pure regression methods, which supports flow interpolation as flow prediction tools.

Then we focus on the lower Manhattan Island where stations are mostly added as interpolation types. Interpolation methods are commonly rated the best method there which complies with the intuition that nearby flow patterns facilitate interpolation methods. Two regression methods as best methods could hardly be spotted in lower Manhattan, but can be seen in North Manhattan, Queens, and Brooklyn (north, east, and south of the system, respectively).

On the edge of the bike share system, where added stations are usually of the extrapolation type, a mixture of best methods can be seen. Even though extrapolation stations are out of the coverage of the existing system with limited nearby spatial dependence, regression models are not necessarily superior to interpolation methods in all cases, which is unexpected. It implies the necessity of considering the spatial dependence or spatial structure of the BSS at all times regardless of when locations are added. Another explanation could be that fewer trips at extrapolation stations introduces more noise than signal, having a negative impact on all methods. When comparing two gravity models, the Negative Binomial (NegBin) model seems to do better in the outskirts of the system where there are probably more stations with a few number of rides and therefore more overdispersion.

TABLE 3 A summary of the prediction results based on Pearson's R for each method using data for each classification category and using all data. Standard deviations are in parentheses after the mean values

Pearson's R (standard deviation)	Interpolation	Margin	Extrapolation	All
Natural neighbor	0.53 (0.22)	0.47 (0.22)	–	–
Ordinary Kriging	0.52 (0.20)	0.45 (0.22)	0.41 (0.20)	0.46 (0.21)
Gravity SI (Poisson)	0.44 (0.19)	0.41 (0.24)	0.40 (0.24)	0.42 (0.22)
Gravity SI (NegBin)	0.42 (0.18)	0.37 (0.22)	0.39 (0.22)	0.40 (0.20)
Regression Kriging (Poisson)	0.54 (0.21)	0.49 (0.24)	0.41 (0.22)	0.47 (0.23)

14679671, 2022, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tgis.12933 by University Of Maryland, Wiley Online Library on [21/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

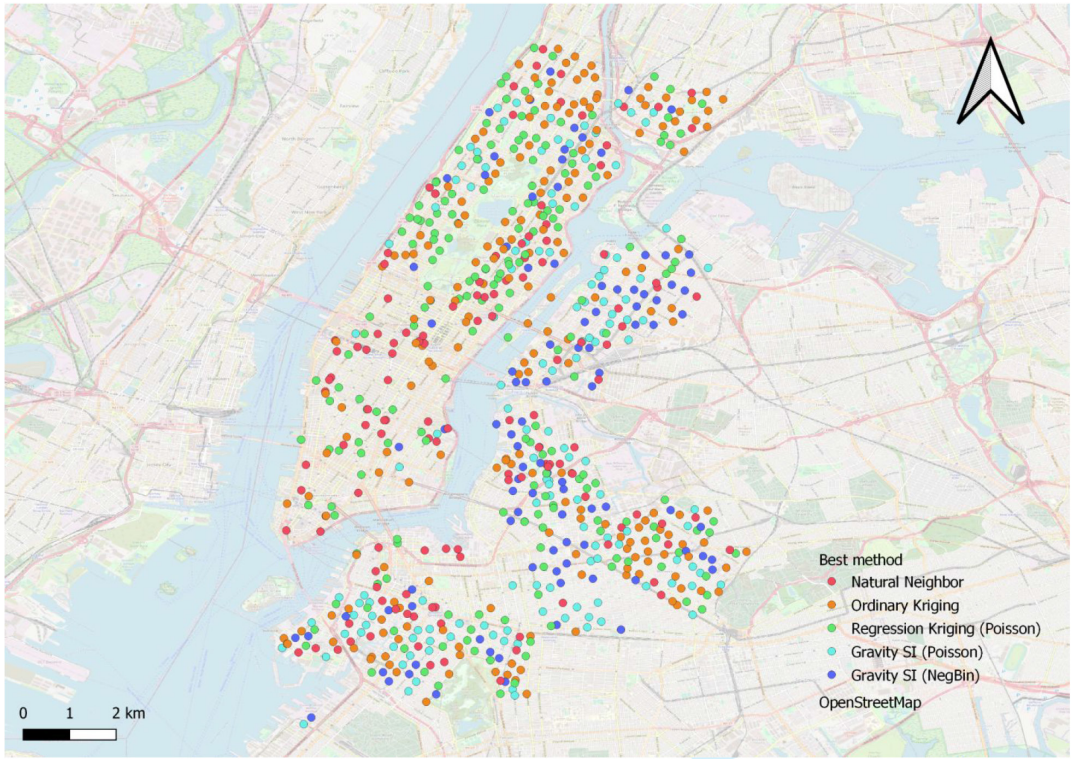

FIGURE 4 Spatial distribution of the best method for flow prediction at each cold start station

6 | DISCUSSION AND CONCLUSIONS

Spatial interaction demand for bike-sharing trips has caught the attention of many researchers in recent years as new bike-sharing systems are installed at an astonishing speed around the world. Yet the prediction of demand at cold start stations is an overlooked issue. Previous studies work around the issue either by filtering out the new stations or using a different spatial aggregation, such as grids. This article leveraged spatial interaction models and flow interpolation techniques to propose an approach to compare different models through a case study using the NYC Citi Bike system. Specifically, the proposed approach attempts to predict OD flows of newly added stations after the BSS infrastructure changes.

After comparing all candidate methods on each identified cold start station, results show that the gravity SI model comes with the advantage of the interpretability of the parameter estimates. For example, the interpretation of gravity SI models shows that the leading trip generation factor is the number of recreation sites nearby, which provides useful additional information to advise the planning of BSSs. Alternatively, interpolation methods, including natural neighbor and ordinary kriging, show a better predictability for stations classified as interpolation but are less performant for stations classified as extrapolation. Regression kriging combines both gravity SI and ordinary kriging and yields the best performance. Finally, spatial trends reveal some heterogeneity of the prediction performance of different methods. Specifically, the downtown Manhattan region shows a higher predictability for interpolation methods, but stations added in Brooklyn are less performant for regression methods only. Interpolation methods unexpectedly outperform regression models in many extrapolation cases, implying the presence of long-range spatial dependence and suggesting the use of spatial interaction models that incorporate spatial dependence (Oshan, 2021) in future work.

14679671, 2022, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tgis.12933 by University Of Maryland, Wiley Online Library on [21/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

One significance of this work is that it begins to fill the gap of the missing historical data that is needed for many popular methods. For example, the methodology could be used to generate pseudo-historical flows so that methods dependent upon them can be used to make predictions over time even for cold start stations. Therefore, the proposed approach facilitates a more robust demand modeling framework.

However, trip prediction for stations classified as extrapolation is still limited using the proposed approach because neither the gravity-type spatial interaction model, nor the interpolation techniques work as well as they do for the interpolation stations. For future work, the idea of geostatistical transfer learning may help overcome this issue (Hoffimann, Zortea, de Carvalho, & Zadrozny, 2021). Another limitation that needs to be more fully explored is the weak predictability of 2020 stations, which is possibly due to the low trip counts or drastic behavioral changes during the pandemic or both. When there are low trip counts, the methods used here tend to underestimate the few stations with higher activity and more effort is needed to develop methods that can handle this scenario.

The methods included here only provide an initial investigation of the cold start issue. Future studies could also further investigate and leverage the temporal dependencies of the OD flows. In this work, a weekly time interval was used, though other intervals could also be explored, as well as the temporal windows used to characterize demand before and after stations are added to the system. Future work could therefore explore different time parameters or even combine them to increase predictive accuracy. These are just a few suggestions that could be on the line of inquiry initiated here to grapple with the issues caused by cold start stations in dynamic transportation systems.

ACKNOWLEDGMENTS

We thank two anonymous reviewers for their insightful comments.

CONFLICT OF INTEREST

The authors declared no conflict of interests with respect to the authorship and/or publication of this original research article.

DATA AVAILABILITY STATEMENT

Bike-share trip data that support the findings of this study are openly available in Citi Bike System Data at https://ride.citibikenyc.com/system-data. POI data can be privately accessed from SafeGraph through https://www.safegraph.com/academics upon request.

ORCID


Zheng Liu https://orcid.org/0000-0002-3256-2070

ENDNOTES

[1] According to bikesharingworldmap.com and including both docked and dockless systems (last accessed on July 30, 2021).

[2] Since capacity values from the most recent station information are representative of all previous system states, capacity values were based on system snapshots of the station information from September 2018, May 2019 and February 2020. Of the 1103 stations that existed from 2015–2020 this provided valid capacity information for all but 135 stations, which either were not running during the snapshots or had zero capacity values. These 135 stations were not included in the computation of the accessibility metric used here.

[3] Zero flows were not included when calculating the performance metrics.

REFERENCES

Barbour, N., Zhang, Y., & Mannering, F. (2019). A statistical analysis of bike sharing usage and its potential as an auto-trip substitute. *Journal of Transport & Health*, *12*, 253–262. https://doi.org/10.1016/j.jth.2019.02.004

Bayraktar, H., & Turalioglu, F. S. (2005). A Kriging-based approach for locating a sampling site: In the assessment of air quality. *Stochastic Environmental Research and Risk Assessment*, *19*(4), 301–305. https://doi.org/10.1007/s00477-005-0234-8

14679671, 2022, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tgis.12933 by University Of Maryland, Wiley Online Library on [21/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Botella, P., Gora, P., Sosnowska, M., Karsznia, I., & Querol, S. C. (2021). *Modelling mobility and visualizing people's flow patterns in rural areas for future infrastructure development as a good transnational land-governance practice*. Preprint, arXiv:2103.01777.

Calafiore, A., Palmer, G., Comber, S., Arribas-Bel, D., & Singleton, A. (2021). A geographic data science framework for the functional and contextual analysis of human dynamics within global cities. *Computers, Environment and Urban Systems*, *85*, 101539. https://doi.org/10.1016/j.compenvurbsys.2020.101539

Campbell, K. B., & Brakewood, C. (2017). Sharing riders: How bikesharing impacts bus ridership in New York City. *Transportation Research Part A: Policy and Practice*, *100*, 264–282. https://doi.org/10.1016/j.tra.2017.04.017

Chen, P.-C., Hsieh, H.-Y., Su, K.-W., Sigalingging, X. K., Chen, Y.-R., & Leu, J.-S. (2020). Predicting station level demand in a bike-sharing system using recurrent neural networks. *IET Intelligent Transport Systems*, *14*(6), 554–561. https://doi.org/10.1049/ietits.2019.0007

Cheng, Z., Trepanier, M., & Sun, L. (2021). *Real-time forecasting of metro origin-destination matrices with high-order weighted dynamic mode decomposition*. Preprint, ArXiv:2101.00466.

Chu, K.-F., Lam, A. Y. S., & Li, V. O. K. (2020). Deep multi-scale convolutional LSTM network for travel demand and origin-destination predictions. *IEEE Transactions on Intelligent Transportation Systems*, *21*(8), 3219–3232. https://doi.org/10.1109/TITS.2019.2924971

Cressie, N. (1993). *Statistics for spatial data*. New York, NY: John Wiley & Sons.

El-Assi, W., Salah Mahmoud, M., & Nurul Habib, K. (2017). Effects of built environment and weather on bike sharing demand: A station level analysis of commercial bike sharing in Toronto. *Transportation*, *44*(3), 589–613. https://doi.org/10.1007/s11116-015-9669-z

Emery, X. (2005). Simple and ordinary multi-Gaussian kriging for estimating recoverable reserves. *Mathematical Geology*, *37*(3), 295–319. https://doi.org/10.1007/s11004-005-1560-6

Eom, J. K., Park, M. S., Heo, T.-Y., & Huntsinger, L. F. (2006). Improving the prediction of annual average daily traffic for non-freeway facilities by applying a spatial statistical method. *Transportation Research Record*, *1968*(1), 20–29. https://doi.org/10.1177/0361198106196800103

Eren, E., & Uz, V. E. (2020). A review on bike-sharing: The factors affecting bike-sharing demand. *Sustainable Cities and Society*, *54*, 101882. https://doi.org/10.1016/j.scs.2019.101882

Faghih-Imani, A., Hampshire, R., Marla, L., & 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

Farmer, C., & Oshan, T. (2017). Spatial interactions. In J. P. Wilson (Ed.), *Geographic information science & technology body of knowledge* (2017, Quarter 4 ed.). University Consortium for Geographic Information Science. Retrieved from https://gistbok.ucgis.org/bok-topics/spatial-interaction

Fotheringham, A. S., & O'Kelly, M. E. (1989). *Spatial interaction models: Formulations and applications*. Boston, MA: Kluwer.

Gao, S., Liu, Y., Wang, Y., & Ma, X. (2013). Discovering spatial interaction communities from mobile phone data. *Transactions in GIS*, *17*(3), 463–481. https://doi.org/10.1111/tgis.12042

Guo, Y., Zhou, J., Wu, Y., & Li, Z. (2017). Identifying the factors affecting bike-sharing usage and degree of satisfaction in Ningbo, China. *PLoS One*, *12*(9), e0185100. https://doi.org/10.1371/journal.pone.0185100

Hoffimann, J., Zortea, M., de Carvalho, B., & Zadrozny, B. (2021). Geostatistical learning: Challenges and opportunities. *Frontiers in Applied Mathematics and Statistics*, *7*, 689393. https://doi.org/10.3389/fams.2021.689393

Jang, W., & 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, H. T. K., & Miller, H. J. (2021). What is essential travel? Socioeconomic differences in travel demand in Columbus, Ohio, during the COVID-19 lockdown. *Annals of the American Association of Geographers*, 1–24. https://doi.org/10.1080/24694452.2021.1956876

Ke, J., Qin, X., Yang, H., Zheng, Z., Zhu, Z., & Ye, J. (2019). *Predicting origin-destination ridesourcing demand with a spatio-temporal encoder-decoder residual multi-graph convolutional network*. Preprint, arXiv:1910.09103.

Khadaroo, J., & Seetanah, B. (2008). The role of transport infrastructure in international tourism development: A gravity model approach. *Tourism Management*, *29*(5), 831–840. https://doi.org/10.1016/j.tourman.2007.09.005

Krings, G., Calabrese, F., Ratti, C., & Blondel, V. D. (2009). Urban gravity: A model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment*, *2009*(7), L07003. https://doi.org/10.1088/1742-5468/2009/07/L07003

Le, N. D., & Zidek, J. V. (2006). *Statistical analysis of environmental space-time processes*. New York, NY: Springer. https://doi.org/10.1007/0-387-35429-8

Lenormand, M., Bassolas, A., & Ramasco, J. J. (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

14679671, 2022, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tgis.12933 by University Of Maryland, Wiley Online Library on [21/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Li, Y., & Shuai, B. (2020). Origin and destination forecasting on dockless shared bicycle in a hybrid deep-learning algorithms. *Multimedia Tools and Applications*, *79*(7), 5269–5280. https://doi.org/10.1007/s11042-018-6374-x

Liu, L., Qiu, Z., Li, G., Wang, Q., Ouyang, W., & Lin, L. (2019). *Contextualized spatial-temporal network for taxi origin-destination demand prediction*. Preprint, aArXiv:1905.06335.

Liu, X., Kang, C., Gong, L., & Liu, Y. (2016). Incorporating spatial interaction patterns in classifying and understanding urban land use. *International Journal of Geographical Information Science*, *30*(2), 334–350. https://doi.org/10.1080/13658816.2015.1086923

Lu, M., Hsu, S.-C., Chen, P.-C., & Lee, W.-Y. (2018). Improving the sustainability of integrated transportation system with bike-sharing: A spatial agent-based approach. *Sustainable Cities and Society*, *41*, 44–51. https://doi.org/10.1016/j.scs.2018.05.023

Marrocu, E., & Paci, R. (2013). Different tourists to different destinations. Evidence from spatial interaction models. *Tourism Management*, *39*, 71–83. https://doi.org/10.1016/j.tourman.2012.10.009

Mitas, L., & Mitasova, H. (1999). Spatial interpolation. In P. A. Longley, M. F. Goodchild, D. J. Maguire, & D. W. Rhind (Eds.), *Geographical information systems: Principles, techniques, management and applications* (pp. 481–494). Chichester, UK: John Wiley & Sons.

Noland, R. B., Smart, M. J., & Guo, Z. (2016). Bikeshare trip generation in New York City. *Transportation Research Part A: Policy and Practice*, *94*, 164–181. https://doi.org/10.1016/j.tra.2016.08.030

NYC Department of Transportation. (2019). *New York City mobility report*. Retrieved from https://www1.nyc.gov/html/dot/downloads/pdf/mobility-report-singlepage-2019.pdf

Oja, P., Titze, S., Bauman, A., de Geus, B., Krenn, P., Reger-Nash, B., & 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, T. M. (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, T. M. (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

Oshan, T. M. (2021). The spatial structure debate in spatial interaction modeling: 50 years on. *Progress in Human Geography*, *45*(5), 925–950. https://doi.org/10.1177/0309132520968134

Otero, I., Nieuwenhuijsen, M. J., & 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

Pan, L., Cai, Q., Fang, Z., Tang, P., & Huang, L. (2019). A deep reinforcement learning framework for rebalancing dockless bike sharing systems. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, Honolulu, HI (pp. 1393–1400). Palo Alto, CA: AAAI. https://doi.org/10.1609/aaai.v33i01.33011393

Pourebrahim, N., Sultana, S., Thill, J.-C., & Mohanty, S. (2018). Enhancing trip distribution prediction with Twitter data: Comparison of neural network and gravity models. In *Proceedings of the Second ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery*, Seattle, WA (pp. 5–8). New York, NY: ACM. https://doi.org/10.1145/3281548.3281555

SafeGraph. (2020). *The impact of Coronavirus (COVID-19) on foot traffic*. Retrieved from https://www.safegraph.com/data-examples/covid19-commerce-pattern

Schuessler, N., & Axhausen, K. W. (2009). Processing raw data from global positioning systems without additional information. *Transportation Research Record*, *2105*(1), 28–36. https://doi.org/10.3141/2105-04

Shaw, S.-L., & Sui, D. Z. (Eds.). (2018). *Human dynamics research in smart and connected communities*. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-73247-3

Si, H., Shi, J., Wu, G., Chen, J., & Zhao, X. (2019). Mapping the bike sharing research published from 2010 to 2018: A scientometric review. *Journal of Cleaner Production*, *213*, 415–427. https://doi.org/10.1016/j.jclepro.2018.12.157

Sibson, R. (1981). A brief description of natural neighbor interpolation. In V. Barnett (Ed.), *Interpreting multivariate data* (pp. 21–36). Chichester, UK: John Wiley & Sons.

Signorino, G., Pasetto, R., Gatto, E., Mucciardi, M., La Rocca, M., & Mudu, P. (2011). Gravity models to classify commuting vs. resident workers. An application to the analysis of residential risk in a contaminated area. *International Journal of Health Geographics*, *10*(1), 11. https://doi.org/10.1186/1476-072X-10-11

Šimbera, J., & Aasa, A. (2019). Areal interpolation of spatial interaction data. In *Proceedings of the 15th International Conference on Location Based Services*, Vienna, Austria (pp. 177–184). https://doi.org/10.34726/lbs2019.61

Stopher, P., Jiang, Q., & FitzGerald, C. (2005). Processing GPS data from travel surveys. In *Proceedings of the Second International Colloquium on the Behavioural Foundations of Integrated Land-Use and Transportation Models: Frameworks, Models and Applications*, Toronto, ONT, Canada.

Su, X., & Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*, *2009*, e421425. https://doi.org/10.1155/2009/421425

14679671, 2022, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tgis.12933 by University Of Maryland, Wiley Online Library on [21/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography*, *46*(Suppl. 1), 234–240. https://doi.org/10.2307/143141

United Nations, Department of Economic and Social Affairs, Population Division. (2019). *World urbanization prospects: The 2018 revision (ST/ESA/SER.A/420)*. New York, NY: United Nations. Retrieved from https://esa.un.org/unpd/wup

Volkovs, M., Yu, G., & Poutanen, T. (2017). DropoutNet: Addressing cold start in recommender systems. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, Long Beach, CA (pp. 1–10).

Wang, M., & 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

Wang, X., Cheng, Z., Trépanier, M., & Sun, L. (2021). Modeling bike-sharing demand using a regression model with spatially varying coefficients. *Journal of Transport Geography*, *93*, 103059. https://doi.org/10.1016/j.jtrangeo.2021.103059

Wang, X., & Kockelman, K. M. (2009). Forecasting network data: Spatial interpolation of traffic counts from Texas data. *Transportation Research Record*, *2105*(1), 100–108. https://doi.org/10.3141/2105-13

Wang, Y., Yin, H., Chen, H., Wo, T., Xu, J., & Zheng, K. (2019). Origin-destination matrix prediction via graph convolution: A new perspective of passenger demand modeling. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Anchorage, AK (pp. 1227–1235). New York, NY: ACM. https://doi.org/10.1145/3292500.3330877

Wu, Y., Zhuang, D., Labbe, A., & Sun, L. (2020). *Inductive graph neural networks for spatiotemporal Kriging*. Preprint, ArXiv:2006.07527. Retrieved from http://arxiv.org/abs/2006.07527

Yang, H., Zhang, Y., Zhong, L., Zhang, X., & Ling, Z. (2020). Exploring spatial variation of bike sharing trip production and attraction: A study based on Chicago's Divvy system. *Applied Geography*, *115*, 102130. https://doi.org/10.1016/j.apgeog.2019.102130

Yao, X., Zhu, D., Gao, Y., Wu, L., Zhang, P., & Liu, Y. (2018). A stepwise spatio-temporal flow clustering method for discovering mobility trends. *IEEE Access*, *6*, 44666–44675. https://doi.org/10.1109/ACCESS.2018.2864662

Zhang, L., Cheng, J., & Jin, C. (2019). Spatial interaction modeling of OD flow data: Comparing Geographically Weighted Negative Binomial Regression (GWNBR) and OLS (GWOLSR). *ISPRS International Journal of Geo-Information*, *8*(5), 220. https://doi.org/10.3390/ijgi8050220

Zhang, Y., & 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

Zhang, Y., Thomas, T., Brussel, M., & van Maarseveen, M. (2017). Exploring the impact of built environment factors on the use of public bikes at bike stations: Case study in Zhongshan, China. *Journal of Transport Geography*, *58*, 59–70. https://doi.org/10.1016/j.jtrangeo.2016.11.014

Zhou, T., Huang, B., Liu, X., He, G., Gou, Q., Huang, Z., & 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

Zhou, X., Wang, M., & Li, D. (2019). Bike-sharing or taxi? Modeling the choices of travel mode in Chicago using machine learning. *Journal of Transport Geography*, *79*, 102479. https://doi.org/10.1016/j.jtrangeo.2019.102479

Zhou, Y., Chen, H., Li, J., Wu, Y., Wu, J., & Chen, L. (2019). Large-scale station-level crowd flow forecast with ST-Unet. *ISPRS International Journal of Geo-Information*, *8*(3), 140. https://doi.org/10.3390/ijgi8030140

Zhu, D., Cheng, X., Zhang, F., Yao, X., Gao, Y., & Liu, Y. (2020). Spatial interpolation using conditional generative adversarial neural networks. *International Journal of Geographical Information Science*, *34*(4), 735–758. https://doi.org/10.1080/13658816.2019.1599122

Zhu, R., Zhang, X., Kondor, D., Santi, P., & Ratti, C. (2020). Understanding spatio-temporal heterogeneity of bike-sharing and scooter-sharing mobility. *Computers, Environment and Urban Systems*, *81*, 101483. https://doi.org/10.1016/j.compenvurbsys.2020.101483

Zimmerman, D. A., de Marsily, G., Gotway, C. A., Marietta, M. G., Axness, C. L., Beauheim, R. L., … Rubin, Y. (1998). A comparison of seven geostatistically based inverse approaches to estimate transmissivities for modeling advective transport by groundwater flow. *Water Resources Research*, *34*(6), 1373–1413. https://doi.org/10.1029/98WR00003

How to cite this article: Liu, Z., & Oshan, T. (2022). Comparing spatial interaction models and flow interpolation techniques for predicting "cold start" bike-share trip demand. *Transactions in GIS*, *26*, 2081–2098. https://doi.org/10.1111/tgis.12933

14679671, 2022, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1111/tgis.12933 by University Of Maryland, Wiley Online Library on [21/09/2023]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License