Imputation of trip data for a docked bike-sharing system

Milan Mathew Thomas[1], Ashish Verma[2,*] and Sai Kiran Mayakuntla[3]

[1]Department of Civil Engineering, Rajiv Gandhi Institute of Technology, Kottayam 686 501, India
[2]Department of Civil Engineering, Indian Institute of Science, Bengaluru 560 012, India
[3]Department of Civil Engineering, Transport Division, Universidad de Chile, Chile

Mobile application-based transportation services are reshaping the urban transportation industries of both the developed and developing worlds. They generate massive amounts of data, which have the potential to provide deeper insights into urban travel activity than ever before. The bike-sharing service (BSS) market is growing at a breakneck pace with new service providers entering the arena. However, we have seen the failure of several BSS start-ups in India in recent years. All these cases have one aspect in common: user dissatisfaction because of insufficient/ineffective rebalancing approaches. The BSS operators rely on data insights to drive their policies and strategies. However, the data generated by these services are found to have several incomplete records as a result of various technical errors, like missing origin/destination. As most BSS modelling focuses on trip origin and destination, completely ignoring (or listwise deleting) trips with missing information results in the loss of valuable data that are still present in other observed variables, which include trip duration, date and time of the trip, and so on. This study proposes two methods for imputing missing data: (i) a probabilistic approach based on Bayes' theorem, and (ii) a machine learning approach based on the *k*-nearest neighbor algorithm. The methodologies for their analyses are presented in detail. Data from a BSS that operated in the Indian Institute of Science campus, Bengaluru, India, are used to illustrate the proposed approaches. This is followed by a brief discussion of the results and a comparison of the performance.

Keywords: Bike-sharing system, imputation, incomplete records, origin and destination, probabilistic and machine learning approaches, trip data.

RECENT innovations in communication technology are revolutionizing the landscape of the urban mobility industry. Mobile application-based transport services have gained a significant market share in recent years in the urban centres of the developed and developing economies alike. Similarly, bike-sharing systems (BSSs) have been introduced in several cities around the world, which have gained prominence during the pandemic, unlike the other transport modes/systems. They have proven to be essential in maintaining the social distancing requirements in urban communities around the world[1]. Besides, its potential role in transit-oriented development has been recognized even before the pandemic[2]. The current generation of BSSs are often mobile application-based and make use of GPS. Such GPS data collected from these applications can be used to understand the travel behaviour of the users[3]. The BSS service providers rely on these data to improve their operational efficiency and to better serve their customers. However, part of the data is observed to have incomplete records due to technical issues in the mode of data collection, etc. Some trips in the real-world BSS dataset may have missing data; the most straightforward approach is to delete the trips with missing observations. However, in this approach, the amount of data available for further analysis is reduced. In such cases, newcomers to the BSS industry suffer the most.

Any given dataset containing missing data points can be divided into two mutually exclusive subsets: completely observed dataset (COD) and partly missing dataset (PMD). The data points in the COD are completely observed, while those in the PMD have missing values. The handling strategy of PMD is decided based on the causes and extent of the missingness. Three general patterns of missingness have been noted in the literature: (i) missing completely at random (MCAR), (ii) missing at random (MAR) and (iii) missing not at random (MNAR). In the MCAR type, missingness is entirely random and induces no bias in any analysis performed on the data. In contrast, the missingness in MAR type can be accounted for by the observed variables, which would induce bias in the analysis. When the value of the missing variable itself is related to the missingness, it is described as an MNAR-type. Liu[4] provides formal mathematical definitions for these types of missingness.

The handling of the missing data has been extensively discussed in the literature[5–9]. The task of assigning a plausible value for a missing variable is called data imputation. Figure 1 gives a broad classification of the imputation strategies generally used. Zhang *et al.*[10] identified two approaches in the data imputation – imputation based on COD and by utilizing information within the PMD. In the second approach, missing values are replaced with the

*For correspondence. (e-mail: ashishv@iisc.ac.in)

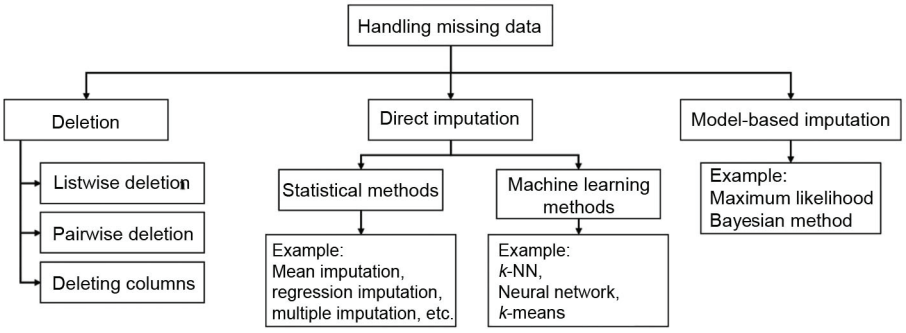

Figure 1. Handling of missing data.

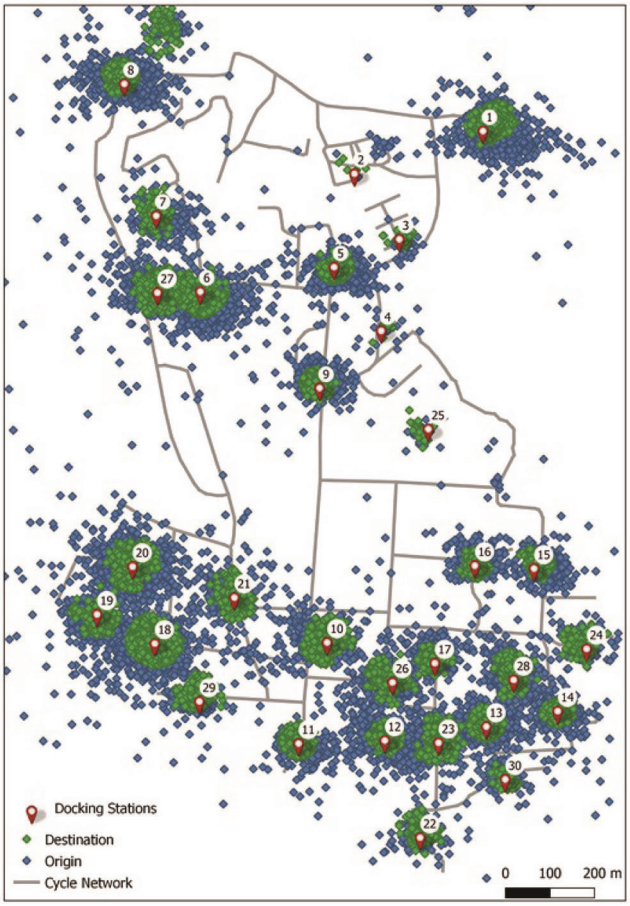

Figure 2. Trip ends and docking stations.

plausible values that are generated from the COD based on evidence from PMD.

With regard to the GPS-based trip dataset, majority of imputation studies have been done to identify the purpose of the trip[3] and the mode of transport[11]. Yao *et al.*[12] have used graph convolution networks to impute the origin–destination (OD) flow with the Beijing taxi trip data in China[12], but there are limited studies in the imputation of the trip ends. In the BSS domain, El Esawey[13] developed the Markov chain Monte Carlo multiple imputation (MCMC MI) method to estimate the missing cycling counts. Liu *et al.*[14] used GPS data obtained from a BSS operation to characterize roadways. The present study is intended to develop a simple and easy-to-implement missing data handling method to retain the data points in the BSS dataset. Two imputation methods are presented here: (i) a probabilistic approach and (ii) a machine learning approach. It was preferred to limit the complexity of the imputation methods and choose simple methods in both approaches.

The probabilistic approach discussed here employs the Bayesian method, which is strong in making intuitive inferences and the attributes are considered random[15,16]. The *k*-nearest neighbor (*k*-NN) algorithm employed in the second approach, is a simple, yet powerful algorithm that has been used for handling all three types of missing data[17].

Methodology

The dataset used in this study is from the Indian Institute of Science (IISc), Bengaluru, India, which had 140 bikes distributed across 30 docking stations. The BSS operated on the campus from April 2019 to January 2020, and service was stopped due to multiple issues. This system is station-based, which means there are no specific racks for the bikes; instead they are parked around the name post of the stations (Figure 2). Even though the dataset has only 2% of data points with incomplete observations, it was concentrated during the operational period of the service. The overall outcome of this study is to extract the maximum inference from the dataset generated from the BSS operation.

Data description

During the operational period, 32,958 trips were made on the campus. Table 1 describes the attributes recorded for each trip. Whenever the user locks/unlocks a bike, the corresponding time and coordinates are logged. The bikes are geofenced and the trips can only be started or terminated within the proximity of name post of a docking

Table 1. Attributes of the dataset

Variable name	Description	Type
booking_id	Unique ID generated by the system for each trip	Trip information
trip_start_time	Bicycle pick-up time from the origin station	Trip information
trip_end_time	Bicycle drop-off time to the destination station	Trip information
actstartloc	Coordinates of the origin station	Trip information
actendloc	Coordinates of the destination station	Trip information
timemin	Trip duration (min)	Trip information
actamt	Cost of the trip (Rs)	Trip information
user_id	Unique ID assigned to each user	User information
mobile_no	Contact number of the user	User information
license_plate	Unique ID assigned to each bicycle	Bicycle information

Source: Bounce Share, 2019.

Table 2. Snapshot of the dataset

booking_id	user_id	startunix	endunix	Duration (sec)	Origin	Destination
39995853	3427145	1579613615	1579613924	309	7	5
39965503	831013	1579602417	1579602516	99	8	8
39450375	1888619	1579271951	1579272793	842	14	
39314595	2109019	1579184672	1579184755	83	5	5
39266867	976352	1579162686	1579163771	1085	14	27
39228872	877175	1579143300	1579143621	321	1	5
39160486	3454851	1579087965	1579088018	53	13	13
39142031	1888619	1579077872	1579078432	560	5	13
39137941	1888619	1579075633	1579076124	491	8	5
39113490	2002191	1579061593	1579062825	1232	27	
39011982	3406736	1578990601	1578990858	257	27	27
39008422	3070606	1578988558	1578988634	76	27	27

Source: Bounce Share, 2019.

Figure 3. Station-based bike-sharing systems.

station, the trip end coordinates are mapped to the corresponding coordinates and ID of the docking station. The destination data value of 643 trips were noted to be missing. This may be attributed to the poor cellular connectivity within the campus. Table 2 shows a sample of the dataset.

Data exploration

The characteristics of the data logged during the operation can be comprehended from the exploratory data analysis. Following are key observations from the analysis.

Docking stations: All the bikes are typically picked-up/dropped-off within a few metres of any docking station. Figure 2 shows that the drop-off locations are more clustered around the docking stations. This is presumably because of a stronger connectivity requirement by the bike with the name post of the stations for dropping-off, leading to more accurate locations. Since there are no fixed racks for holding the bikes there is a piling up of the bikes around that location (Figure 3). As a result, the user may also find it difficult in terminating the trips.

Demand variation: Nearly 30% of all trips were observed to be round trips. Also, the demands showed recurring patterns of activity on weekdays and weekends. Figure 4 shows the variation of demand across the OD pairs. During the morning peak hour a good proportion of the trips

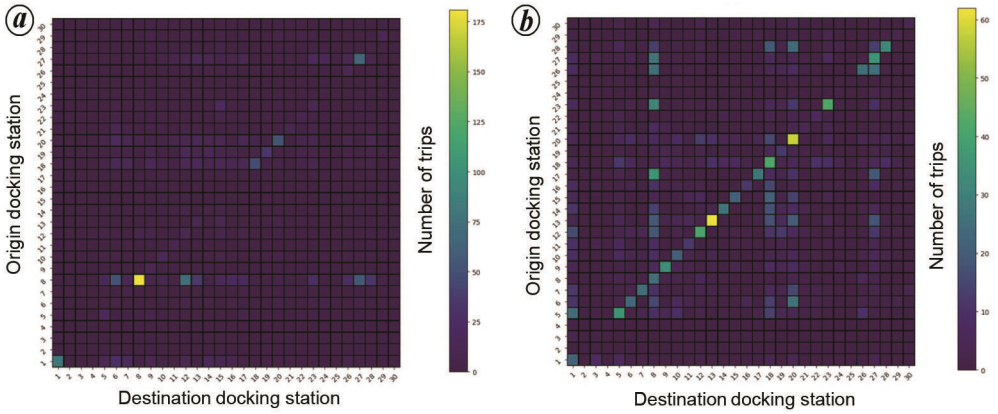

Figure 4. Spatio-temporal demand variation during (***a***) 09:00–10:00 (morning peak), (***b***) 17:00–18:00 (evening peak).

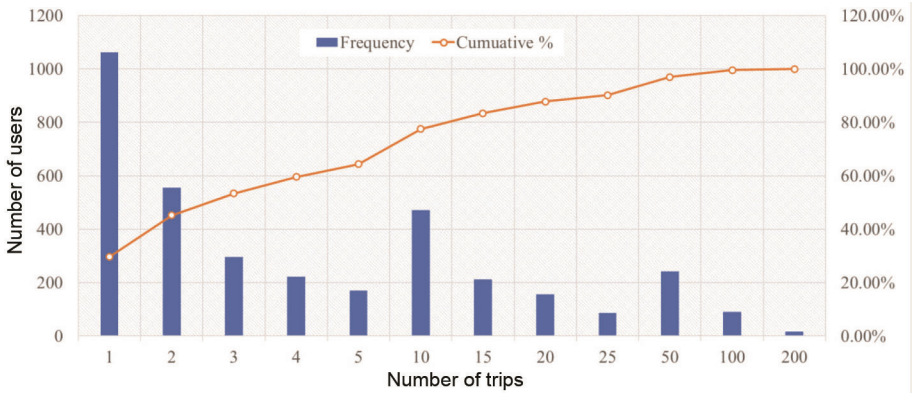

Figure 5. Distribution of the number of trips made by the user.

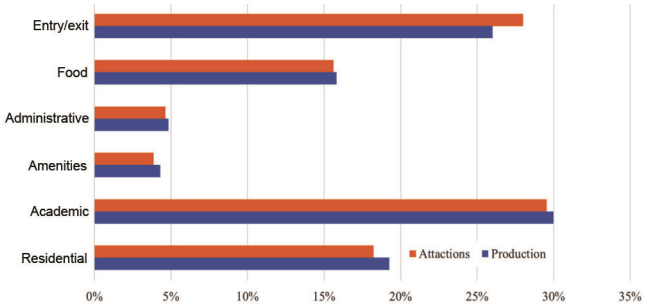

Figure 6. Trip productions and attractions from different activity centres.

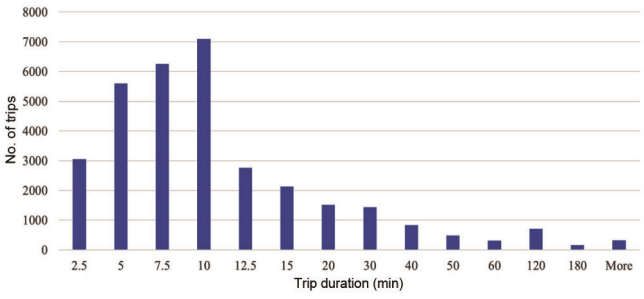

Figure 7. Distribution of trip duration.

are round trips from station 8, while the evening peak demand is distributed among different stations. This points to the fact that the spatial patterns of trip origin and destination vary with the hour of the day. The destination choice is also influenced by the station from which the trip has been initiated.

Active users: During the observational period, the BSS served about 3583 users, of which 45% were single-time users. Figure 5 depicts that only a small fraction used the system frequently and showed a rich trip history. Trips made by the users are defined by their purpose. Figure 6 shows the share of trips originating and terminating near different activity areas. Nearly 60% of the trip ends are near academic buildings and the entry/exit gates followed by trips to eating places and residential areas.

Trip duration: Figure 7 shows the distribution of trip duration; most of the trip durations fall in the range 5–10 min. The cycling distance between each station pair ranges from 100 m to 2.5 km. If the bikes are not locked properly, then the timer will continue resulting in unrealistic values for the trip duration. Furthermore, the operator charged Rs 3 (0.014 USD) for 10 min, which is inexpensive and

encourages users to hold the cycle in interim locations and chain trips. So, the value for the trip ends can be imputed if information about the user and trip duration is available.

Imputation methods

As mentioned before, the raw dataset is divided into COD and PMD (Tables 3 and 4 respectively). Each row in the PMD is referred to as a query instance, for which the destination value is to be imputed. From data exploration, it can be inferred that the plausible value for the missing destination is influenced by: (i) the origin – it was found that the flow between some OD pairs is zero, so one could boldly remove those docking stations from the plausible set of stations; (ii) the start time of the trip – for instance, in the morning demand was more in the docking stations near the campus gates, which means their probability for being the reasonable impute will be comparatively high, and (iii) the behaviour of the user – say, if the imputation is based on trip duration. The above information and their combination retained in the PMD are utilized for predicting the missing destination values.

Following are the notations used in this study to explain the imputation methods.

O, Origin docking station from where the user starts his/her trip.
D, Destination docking station where the user terminates his/her trip.
d, Duration of the trip from O to D.
S, Number of docking stations in the BSS network.
n, Number of trips in COD.
m, Number of trips in PMD.
O', Origin docking station in the query instance.
D', Missing destination docking station.
d', Trip duration in the query instance.
user_id, Unique ID assigned to the user by the BSS operator.

Bayesian method: The Bayesian method is a probabilistic approach for the imputation of destination values. It makes use of the Bayes theorem for calculating the conditional probability of a destination i being the missing value based on prior knowledge of a trip. The probability of any station in the BSS network being the plausible values for D' is $1/S$ in the equally likely scenario. However, the probabilities of individual docking stations vary based on many factors like time of the day the trip was made, who made the trip, from which docking station the trip was made – called prior information. For instance, during 9–10 am, users will be taking BSS to reach their lecture rooms/laboratories from their place of residence. So the docking stations near the academic buildings will have a higher probability. When evidence from the query instance is considered, the probabilities of the stations can be estimated more effectively. The Bayesian equation (eq. (1)) is used to calculate such conditional probabilities of the stations.

$$P\left(\frac{D_i}{E}\right) = \frac{P\left(\frac{E}{D_i}\right) \times P(D_i)}{\sum_{j=1}^{s} P\left(\frac{E}{D_j}\right) \times P(D_j)}, \quad (1)$$

where $P(D_i/E)$ is the posterior probability of the ith destination given the evidence E, $P(D_i)$ the prior probability of the ith destination, $P(E/D_i)$ the likelihood of evidence E if the destination is D_i, $P(D_j)$ the prior probability of the jth destination and $P(E/D_j)$ is the likelihood of evidence E if the destination is D_j.

COD is filtered for prior information such as the trip origin, user id, day and time. The corresponding prior probabilities are then computed, as well as the probability of the evidence for a given destination. To measure the posterior probabilities of the destinations, these values are fed into the Bayes equation. Figure 8 shows the estimation of the posterior probabilities. For each docking station, posterior probabilities are measured and the station with a higher value is imputed for the missing data.

k-NN method: The k-NN algorithm can predict a point value by considering the values of the points that are closest to it[17]. It essentially imputes the missing information based on the corresponding values from its 'nearest neighbor' sample[18], which is defined by similarity with the available information. k-NN imputation is robust to bias with a higher percentage of missing values[5]. Both the discrete and continuous attributes can be imputed using

Table 3. Snapshot of the completely observed dataset

user_id	Duration (sec)	Origin	Destination
3427145	309	7	5
831013	99	8	8
2109019	83	5	5
976352	1085	14	27
877175	321	1	5
3454851	53	13	13
1888619	560	5	13
1888619	491	8	5
3406736	257	27	27
3070606	76	27	27

Table 4. Snapshot of the partly missing dataset

user_id	Duration (sec)	Origin	Destination
1888619	842	14	
2002191	1232	27	
1076352	602	4	
76352	109	12	

k-NN and it can be easily adapted to any attribute by just selecting those for distance calculation[19].

The k-NN model depends on the number of instances that are taken into consideration, which is depicted by the parameter k. For the imputation of data, trip duration is considered as the distance metric and has a discrete value as output (destination ID in this case). If the k value is too large, destination IDs with a greater number of samples in the training dataset can have more control than the smaller ones and will introduce bias in the results. On the other hand, if the k value is too small, the advantage of having many samples in the training dataset is not exploited[20]. The calculation of proximity plays a key role in this imputation. The evidences like O', d' and user_id are utilized for the imputation.

The training dataset for k-NN is obtained by filtering COD for O', which makes the imputation process user-specific. We assume that the users ride bicycles at an average speed and there is no halt during a trip. Thus, the trip duration can be approximated to the cycling time between origin and destination.

The closeness of d' to all the observed trip durations made by an user from O is obtained. Mathematically, we have pairs $(d_1, D_1), (d_2, D_2), \ldots, (d_n, D_n)$ taking values in $R \times \{1, 2, \ldots, k\}$, where D is the destination label of d. Given Euclidean norm $\|\cdot\|$ on R and a duration $d \in R$, let $(d_{(1)}, D_{(1)}), (d_{(2)}, D_{(2)}), \ldots, (d_{(n)}, D_{(n)})$ be the reordering of the data, such that $\|d_{(1)} - d'\| \leq \cdots \leq \|d_{(n)} - d'\|$. Gathering the destination labels of the nearest neighbours, use the simple majority of the destination labels to predict the missing values. Figure 9 shows the concept of this.

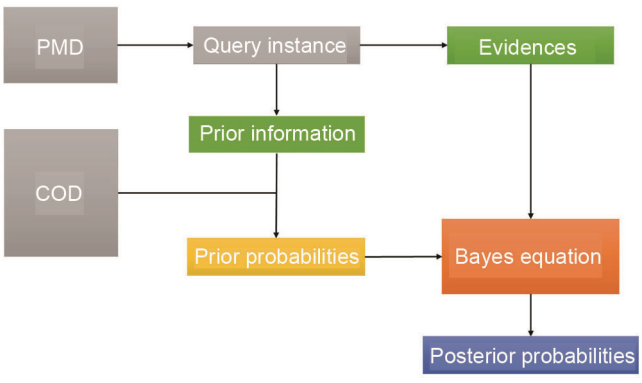

Figure 8. Flow chart of the Bayesian method.

Table 5. Training dataset for the query instance

user_id	Duration (sec)	Origin	Destination
1888619	863	14	8
1888619	855	14	8
1888619	1085	14	27
1888619	837	14	8
1888619	491	14	5
1888619	660	14	5

(1) The training dataset is generated from COD for each query instance in the PMD. The training dataset contains all the trips made by the user_id from O'. As shown in Table 5, for the first query instance d' is 842 sec. The training dataset for this query instance is extracted from the observed data. COD is filtered with user_id = 1888619 and origin = 14 to develop the training data for the query instances.

(2) From Table 5, the number of trips made by the user 1888619 from station 14 is 6. Then ordered pair of duration and destination label will be

$$(d_1, D_1), (d_2, D_2), \ldots, (d_6, D_n)$$

$$= (863, 8), (855, 8), \ldots, (660, 5).$$

(3) Reorder the training dataset according to the proximity of the duration to the query instance. Then, the ordered pair of duration and destination label will be rearranged as

$$(d_{(1)}, D_{(1)}), (d_{(2)}, D_{(2)}), \ldots, (d_{(6)}, D_{(n)})$$

$$= (837, 8), (855, 8), (863, 8), \ldots, (491, 5).$$

(4) Gather the destination labels of the k nearest neighbours. Let $k = 4$; then the possible destination labels are $\{8, 8, 8, 5\}$.

(5) Use the simple majority of the destination labels to predict the value of the query instance. Thus, the imputed destination (D') will be 8 for the query instance (d') 842 sec.

Results

The output of the proposed imputation methods is assessed using a random sample of size 300 from COD. The performance metric considered is the percentage of query instances that the methods are able to impute correctly.

Bayesian method

The method is evaluated with different combinations of evidence and prior information. The user_id and O' from the query instances were used as evidence for the calculation of posterior probabilities. The query instance also has information about the day and time during which the trip was made, which was used for calculating the prior probabilities. Table 6 shows the variation of imputation accuracy with prior information and evidence. In this case, the best imputation results are obtained when posterior probability estimations are performed using the user id as evidence. The results show that imputation accuracy decreases when more prior information is used in the calculation.

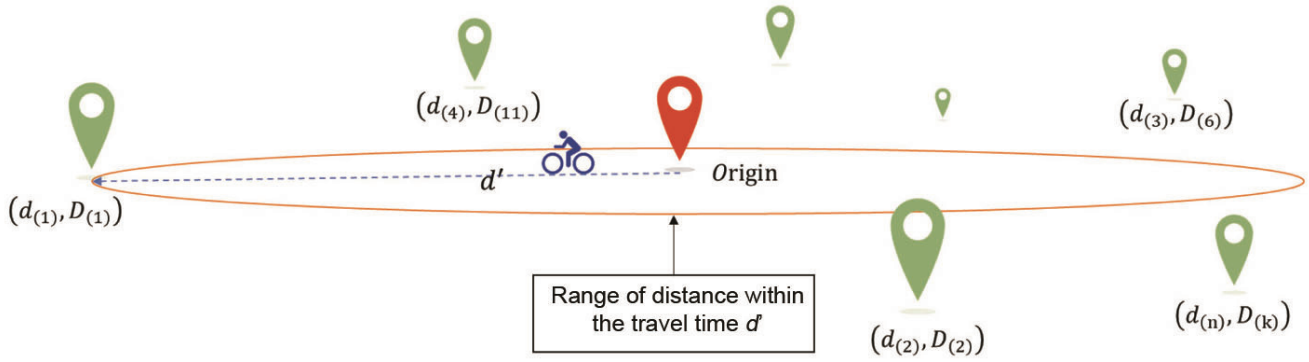

Figure 9. Concept of the *k*-nearest neighbour (*k*-NN) method.

Table 6. Summary of the Bayesian method

Prior information Choice of the destination depends on						
Origin	Hour of day	Day of week	User	Evidence	Posterior probability	Accuracy of imputation (%)
				Origin	*P* (destination | origin)	25
✓				Origin	*P* (destination | origin)	27
	✓			Origin	*P* (destination | origin)	22
		✓		Origin	*P* (destination | origin)	25
			✓	Origin	*P* (destination | origin)	14
✓	✓			Origin	*P* (destination | origin)	16
✓		✓		Origin	*P* (destination | origin)	27
✓			✓	Origin	*P* (destination | origin)	6
	✓	✓		Origin	*P* (destination | origin)	19
	✓		✓	Origin	*P* (destination | origin)	7
		✓	✓	Origin	*P* (destination | origin)	7
✓	✓	✓		Origin	*P* (destination | origin)	5
✓	✓	✓	✓	Origin	*P* (destination | origin)	6
				User_ID	*P* (destination | user ID)	37
✓				User_ID	*P* (destination | user ID)	32
	✓			User_ID	*P* (destination | user ID)	34
		✓		User_ID	*P* (destination | user ID)	34
			✓	User_ID	*P* (destination | user ID)	11
✓	✓			User_ID	*P* (destination | user ID)	18
✓		✓		User_ID	*P* (destination | user ID)	26
✓			✓	User_ID	*P* (destination | user ID)	6
	✓	✓		User_ID	*P* (destination | user ID)	25
	✓		✓	User_ID	*P* (destination | user ID)	8
		✓	✓	User_ID	*P* (destination | user ID)	5
✓	✓	✓		User_ID	*P* (destination | user ID)	15
✓	✓	✓	✓	User_ID	*P* (destination | user ID)	6

k-NN method

If the training dataset has at least *k* number of journeys, the query instance will be imputed; otherwise, the query instance will be denied. The number of trips in the training dataset and the *k* value determine whether the query instance is rejected or imputed in the *k*-NN process. Figure 10 shows that as the *k* value increases, the share of trips getting imputed decreases and the rejection increases. Moreover, as the *k* value increases, the accuracy of the imputation increases. As a result, the *k* value is set as a compromise between imputation accuracy and the number of trips to be made.

Discussion

The imputation accuracy of the Bayesian method is based on prior information and evidence, while the *k*-NN method relies on the user's trip duration. The Bayesian imputation model becomes more biased as the prior information to the model increases and the imputation accuracy decreases.

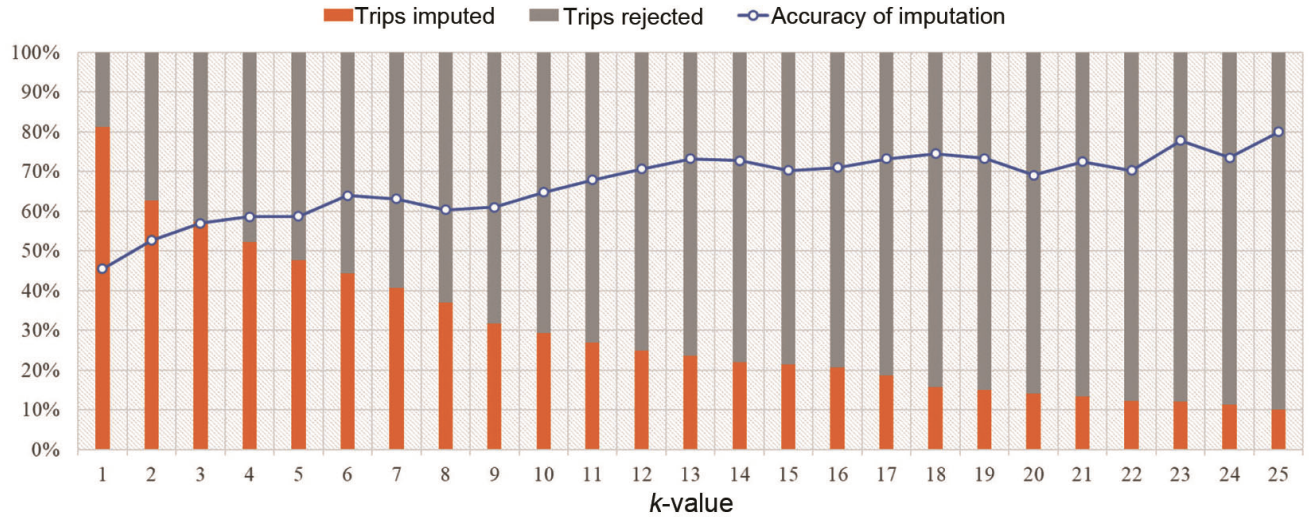

Figure 10. Performance of *k*-NN imputation for different *k* values.

The *k*-NN approach can only be used by those who have taken at least *k* trips in the past. The Bayesian method yielded 37% accuracy in this situation, while the *k*-NN method ($k = 1$) yielded 45% accuracy. The *k*-NN approach provides flexibility to control the quality and quantity of imputation. If the dataset is used for a holistic analysis after imputation, the quantity of data should take precedence over removing all of the missing trips in the *k*-NN process. When the dataset is used for microscopic analysis, however, the quality of the dataset is more important; so a higher *k* value is preferred.

Conclusion

The rebalancing of bikes across docking stations in a BSS operation is critical to ensuring their availability to the users. Rebalancing operations should be efficient and strategic because they account for a significant portion of the operational costs borne by the BSS service providers. As a result, these decisions must be based on data. To calculate bike surplus and deficiency at the docking stations, the origin and destination of each trip are translated as demand on each docking station. It is critical for newcomers to the market to make the best possible use of the data gathered from their operations.

This study discusses two methods for imputing missing trip ends in the BSS trip dataset based on other available information. To reduce the complexity of the model, we developed imputation methods based on the Bayes theorem and the *k*-NN algorithm, which also simplifies implementation. The Bayesian method aided in comprehending how destination preferences vary according to the user, origin and time of day of the trip. When compared to the Bayesian method, the *k*-NN method achieves a higher level of imputation accuracy. The data used to illustrate the imputation methods were sourced from a campus-based BSS operation, which resembles a controlled experiment. The imputation techniques discussed here could be extrapolated to data from a city-wide operation as well.

1. Park, S., Kim, B. and Lee, J., Social distancing and outdoor physical activity during the COVID-19 outbreak in South Korea: implications for physical distancing strategies. *Asia Pac. J. Public Health*, 2020, **32**, 360–362.
2. Glass, C., Appiah-Opoku, S., Weber, J., Jr., Steven L. Jones, Chan, A. and Oppong, J., Role of bikeshare programs in transit-oriented development: case of Birmingham, Alabama. *J. Urban Plann. Dev.*, 2020, **146**, 1–9.
3. Nguyen, M. H., Armoogum, J., Madre, J. L. and Garcia, C., Reviewing trip purpose imputation in GPS-based travel surveys. *J. Traffic. Transp. Eng.* (*Eng. Ed.*), 2020, **7**, 395–412.
4. Liu, X., Methods for handling missing data. In *Methods and Applications of Longitudinal Data Analysis*, Academic Press, Imprint, 2016, pp. 441–473; ISBN: 978-0-12-801342-7; http://dx.doi.org/10.1016/B978-0-12-801342-7.00014-9.
5. Acuña, E. and Rodriguez, C., The treatment of missing values and its effect on classifier accuracy. Classification, Clustering and Data Mining Applications. In Proceedings of the Meeting of the International Federation of Classification Societies (IFCS), Illinois Institute of Technology, Chicago, Springer, Berlin, Heidelberg, 15–18 July 2004; doi:10.1007/978-3-642-17103-1_60.
6. Badr, W., 6 Different ways to compensate for missing values in a dataset, 2019; https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-60-22d9ca0779 (accessed on 5 December 2019).
7. García-Laencina, P. J., Morales-Sánchez, J., Verdú-Monedero, R., Larrey-Ruiz, J., Sancho-Gómez, J. L. and Figueiras-Vidal, A. R., Classification with incomplete data. *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques* (eds Magdalena-Benedito and Serrano López, A.), IGI Global, 2010, pp. 147–175; http://doi:10.4018/978-1-60566-766-9.ch007
8. Skarga-Bandurova, I., Biloborodova, T. and Dyachenko, Y., Strategy to managing mixed datasets with missing items. In 17 International Conference on Information Processing and Management of Uncertainty in Knowledge-Based System. *Theory and Foundations* (eds

Medina, J. *et al.*). IPMU 2018, Cádiz, Spain, 11–15 June 2018. Part of *Communications in Computer and Information Science* (Book Series), Springer, Cham, vol. 854; https://doi.org/10.1007/978-3-319-91476-3_50.
9. Soley-Bori, M., Horn, M., Morgan, J. and Min Lee, K., *Dealing with Missing Data: Key Assumptions and Methods for Applied Analysis*, 2013.
10. Zhang, S., Jin, Z. and Zhu, X., Missing data imputation by utilizing information within incomplete instances. *J. Syst. Softw.*, 2011, **84**, 452–459.
11. Feng, T. and Timmermans, H. J. P., Comparison of advanced imputation algorithms for detection of transportation mode and activity episode using GPS data. *Transp. Plann. Technol.*, 2016, **39**, 180–194; doi:http://dx.doi.org/10.1080/03081060.2015.1127540.
12. Yao, X., Gao, Y., Zhu, D., Manley, E., Wang, J. and Liu, Y., Spatial origin-destination flow imputation using graph convolutional networks. *IEEE Trans. Intell. Transp. Syst.*, 2020, 1–11.
13. El Esawey, M., Using spatio-temporal data for estimating missing cycling counts: a multiple imputation approach. *Transp. A: Transp. Sci.*, 2020, **16**, 5–22; doi:10.1080/23249935.2018.1440262.
14. Liu, X. C., Taylor, J., Porter, R. J. and Wei, R., Using trajectory data to explore roadway characterization for bikeshare network. *J. Intell. Transp. Syst. Technol. Plann., Oper.*, 2018, **22**, 530–546; doi:https://doi.org/10.1080/15472450.2018.1444484.
15. Buhi, E. R., Goodson, P. and Neilands, T. B., Out of sight, not out of mind: strategies for handling missing data. *Am. J. Health Behav.*, 2008, **32**, 83–92; doi:https://doi.org/10.5993/AJHB.32.1.8.
16. Little, R., Calibrated Bayes, for statistics in general and missing data in particular 1. *Stat. Sci.*, 2011, **26**, 162–174; doi:10.1214/10-STS318.
17. Obadia, Y., The use of KNN for missing values, 2017; https://towardsdatascience.com/the-use-of-knn-for-missing-values-cf33d935c637 (accessed on 5 December 2019).
18. Mucherino, A., Papajorgji, P. and Pardalos, P. M., *k*-Nearest neighbor classification. In *Data Mining in Agriculture*, Springer, New York, USA, 2009, pp. 83–106.
19. Batista, G. and Monard, M. C., A study of *k*-Nearest neighbour as an imputation method. *Hybrid Intell. Syst. Ser. Front. Artif. Intell. Appl.*, 2002, **30**, 251–260.
20. Mucherino, A., Papajorgji, P. J. and Pardalos, P. M., Validation. In *Optimization Data Mining in Agriculture*, Series: Springer Optimization and its Applications, Springer, New York, USA, 2009, pp. 161–172; https://doi.org/10.1007/978-0-387-88615-2.

Received 22 March 2021; revised accepted 25 November 2021

doi: 10.18520/cs/v122/i3/310-318