PLOS ONE

RESEARCH ARTICLE

Statistical patterns of human mobility in emerging Bicycle Sharing Systems

Xiangyu Chang[1], Jingzhou Shen[1], Xiaoling Lu[2], Shuai Huang[3]*

1 Center of Data Science and Information Quality, Department of Information Management and E-business, Xi'an Jiaotong University, Xi'an, China, **2** Center for Applied Statistics, School of Statistics, Renmin University of China, Beijing, China, **3** Department of Industrial and Systems Engineering, University of Washington, Seattle, United States of America

* shuaih@uw.edu

Citation: Chang X, Shen J, Lu X, Huang S (2018) Statistical patterns of human mobility in emerging Bicycle Sharing Systems. PLoS ONE 13(3): e0193795. https://doi.org/10.1371/journal.pone.0193795

Editor: Kewei Chen, Banner Alzheimer's Institute, UNITED STATES

Received: September 25, 2017

Accepted: February 20, 2018

Published: March 15, 2018

Data Availability Statement: All relevant data are included within the paper and its Supporting Information files.

Funding: Chang was partially supported by the National Natural Science Foundation of China (Project No. 11771012, 91546119) and the Major Program of National Natural Science Foundation of China (Project No. 71731009, 71742005). Lu was partially supported by the National Natural Science Foundation of China (Project No. 61502342). The authors also acknowledge funding support from

Abstract

The emerging Bicycle Sharing System (BSS) provides a new social microscope that allows us to "photograph" the main aspects of the society and to create a comprehensive picture of human mobility behavior in this new medium. BSS has been deployed in many major cities around the world as a short-distance trip supplement for public transportations and private vehicles. A unique value of the bike flow data generated by these BSSs is to understand the human mobility in a short-distance trip. This understanding of the population on short-distance trip is lacking, limiting our capacity in management and operation of BSSs. Many existing operations research and management methods for BSS impose assumptions that emphasize statistical simplicity and homogeneity. Therefore, a deep understanding of the statistical patterns embedded in the bike flow data is an urgent and overriding issue to inform decision-makings for a variety of problems including traffic prediction, station placement, bike reallocation, and anomaly detection. In this paper, we aim to conduct a comprehensive analysis of the bike flow data using two large datasets collected in Chicago and Hangzhou over months. Our analysis reveals intrinsic structures of the bike flow data and regularities in both spatial and temporal scales such as a community structure and a taxonomy of the eigen-bike-flows.

Introduction

Understanding human mobility pattern is a longstanding scientific pursuit of mankind [1–5]. Many new data resource, such as GPS trajectory [6–8] and mobile phone data [3, 9, 10], are nowadays powerful social microscopes that bring new opportunities for us to study human mobility in new mediums, allowing to "photograph" the main aspects of the society and to create a comprehensive picture of human mobility behavior. Recently, the Bicycle Sharing System (BSS) has been spreading over 1,000 cities around the world [11] as a powerful approach to improve the first/last mile connection to other transportations. Comparing with the first-generation BSS such as the White Bicycle Plan deployed in Amsterdam in 1960s, the third generation BSS highlights the integration of information technology that enables users to borrow

the National Science Foundation under Grant CMMI-1536398.

Competing interests: The authors have declared that no competing interests exist.

bike from any station and return the bike to any station in a city. As nowadays the trips could be automatically recorded, this data provide a great opportunity to understand the human mobility in a short-distance trip, which could lead to better management and operation of BSSs in traffic prediction [12, 13], station placement [14–16], usage pattern analysis [17, 18], bike reallocation [19, 20], and inventory management [21, 22], all are crucial aspects to better manage BSSs to meet the population's dynamic needs.

Generally, there are two different schools of approaches to analyze the data of BSS. One considers individual's trip as a basic study object. For example, to provide an efficient service schedule for bike reallocation, Zhang et al. [18] considered a trip destination and duration prediction model on the individual level. Chen et al. [14] formulated the bike station placement issue as a bike trip demand prediction problem. Zhang and Yu [12] studied a trip route planning problem for individuals. Studying one trip information leads to analytical tractability, however, methodologies developed from this perspective would find limitations when considering decision-makings on the system level involving all stations and all users over time. Thus, another type of approaches aggregate individual's trips in a time window (commonly refereed as *bike flows*, as a bike flow is the collection of all trips from an ingress station to an egress station in a time window). For instances, Li et al. [13] provided a hierarchical prediction model to predict the bike flows that will be rent from/returned to each station in a future period so that reallocation of imbalance bikes can be executed in advance. Etienne and Latifa [17] proposed a model-based clustering algorithm to classify bike stations for efficient management.

In this paper, we focus on the bike flow data as it provides system-level information. Comparing with other existing works that analyzed the bike flow data, we notice that most of the existing works largely focus on prediction using the bike flow data rather than inquiring the data for extracting system-level statistical patterns. Probably because of this, none of them aimed to conduct a delicate analysis of the variation structure in the bike flow data. On the other hand, a series of challenges arise for analyzing the bike flow data. First, it has been found that the bike flow data is very noisy, showing an intrinsic uncertainty structure in both spatial and temporal domains. This often raises up the concern of how much regularity (which then determines predictability) is embedded in the BSS data as the bikes are shared by massive users all over the city, not to mention other uncontrollable conditions such as weather, transportation infrastructure, daily transportation conditions, demographics, and geographical disparities. Second, speaking of the bike flow data as a statistical object, significant dependence has been observed among the bike flows. The dependence makes the analysis of bike flow data difficult since many classical models assume independent assumption. Third, bike flows have a high-dimensional structure. Consider a BSS with N bike stations, there are $O(N^2)$ bike flows. The high dimensionality and dependency of the bike flows present major difficulties for statistical analysis.

To overcome the aforementioned challenges, we realize that a crucial step is to decide on what spatial scale the data should be analyzed. Solid evidences are identified in our study that we should first use clustering approach to detect the community structure among stations, and then, build the analysis on these clusters rather than on individual stations. By aggregation of the bike flows in or between detected communities (called as *aggregate bike flows* (ABFs)), not only the number of bike flows is reduced into a manageable size, but also the statistical regularity embedded in the bike flow data is sharpened which can be statistically articulated by the Principal Component Analysis (PCA). Both hierarchical clustering and PCA are not model-based methods, so they do not rely on the independent assumption of bike flows. In this paper, we show that this assembly of statistical analysis pipeline could reveal interesting city-wide statistical patterns on both datasets from Chicago and Hangzhou. Note that, in this paper, we use

Table 1. Chicago data are public and released every two quarters. Hangzhou data are private and shared by the company of Hangzhou Public Bicycle System for research purpose only.

BSS	Data set	trip	station
Chicago	2016 Q1-Q4	3,595,383	581
Hangzhou	2013 0809-1113	29,998,826	2974

https://doi.org/10.1371/journal.pone.0193795.t001

lower-case letters, e.g., x, to represent scalars, bold-face lower-case letters, e.g., $\mathbf{x}$, to represent vectors, and bold-face upper-case letters, e.g., $\mathbf{X}$, to represent matrices.

Results

Data description

The two datasets used in this analysis were provided by Chicago and Hangzhou BSSs (see S1 and S2 Datasets). In the two datasets, each trip records the user ID, the trip start and end time, and the origin and destination stations. In the Chicago data, we only focus on regular subscribers of the system that form the majority of the users. Table 1 shows detailed descriptions of the two datasets.

As shown in Fig 1(a), stations of both BSSs are usually located next to each other, forming a certain spatial pattern. Also, it could be seen that the number of trips per day exhibits a mix of regularity and uncertainty, as shown in Fig 1(b). For instance, the amount of trips of Hangzhou BSS is mostly stable, but could be occasionally small such as the amount of trips on the day of 7th Oct, 2013. Because this was the last day of National Holiday in China, many tourists were leaving Hangzhou city and many local customers were resting at home. This phenomenon is called short-lived property of BSS that has been reported in the literature [13, 18]. In Fig 1(c) and 1(d), we count the average number of trips on hourly basis in one day, and the

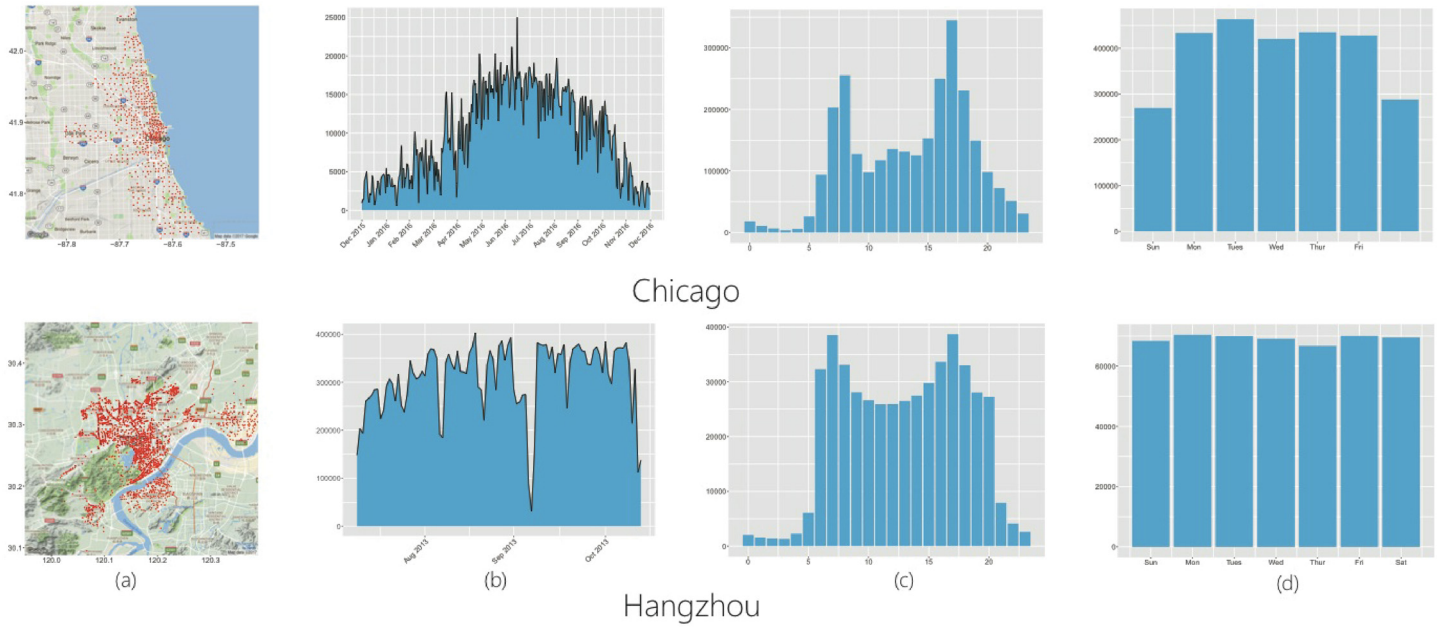

Fig 1. (a) Location information of the BSS stations; (b) number of trips per days; (c) average number of trips on hourly basis; and (d) average number of trips of each day in a week.

https://doi.org/10.1371/journal.pone.0193795.g001

average number of trips of each day in one week, respectively. It is observed that users in Chicago like using the bike in workday and rush hours, indicating that those users may have found usage of the BSS for home-workplace commutes. The average number of trips of Hangzhou BSS is relatively stable over days in one week, and have two peaks in the rush hours of one day.

The community structure

In many real-world networks, it is common that some nodes in the network would be recognized as hubs that either connect with many other nodes or contribute substantially to the network activities. It is interesting that in the bike flow data from both cities, we didn't identify significant hub stations that can account for the amount of bike flow traffic that is significantly larger than average. To show that, we present the cumulative distribution functions (CDFs) of the number of trips in the bike stations in Fig 2. It indicates that, to account for 80% of the total trip records, it took 36% of the stations for Chicago and 42% of the stations for Hangzhou.

Although we didn't discover significant hubs, we identified a community structure of the stations [23, 24], i.e., showing a pattern that there are many bike flows within the stations in the same cluster but less bike flows between stations in different clusters. Specifically, we used the classical hierarchical clustering method to detect the community structure. Assume that $\mathbf{A}^t = (a_{ij}^t) \in \mathbb{R}^{N^2}$ is the adjacent matrix of a BSS, where $a_{ij}^t (t = 1, \ldots, T)$ denotes the total number of bike flows between the i-th and j-th stations in the t-th time epoch, T is the total number of time epochs, and N is the number of stations. Also, denote the distance at the t-th time epoch between the i-th and j-th stations as $s_{ij}^t = 1/a_{ij}^t$. Based on the distance matrix $\mathbf{S}^t = (s_{ij}^t) \in \mathbb{R}^{N^2}$, we can construct the total distance matrix $\mathbf{S} = \sum_{t=1}^{T} \mathbf{S}^t$. By applying the classical hierarchical clustering on $\mathbf{S}$, the bike stations are grouped 15 and 40 communities for Chicago and Hangzhou respectively in Fig 3. Note that the time interval in the paper is one hour. Thus, $T = 24 \times 366 = 8{,}784$ for Chicago BSS and $T = 24 \times 96 = 2{,}304$ for Hangzhou BSS, respectively.

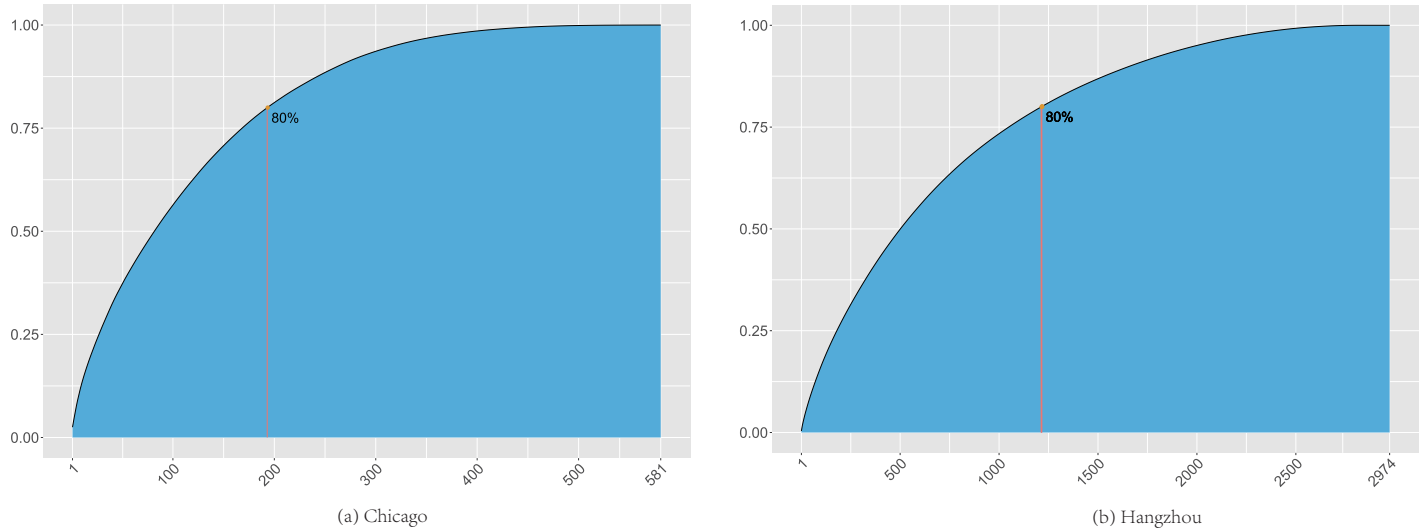

Fig 2. CDF of number of bike stations v.s. number of trips.

https://doi.org/10.1371/journal.pone.0193795.g002

Fig 3. Community structures detected in both Chicago and Hanzhou bike flow data.

https://doi.org/10.1371/journal.pone.0193795.g003

Aggregate bike flow (ABF) based on the community structure

Following the insight revealed by the community structure, we aggregate bike flows of the stations on the basis of clusters and conduct further statistical analysis on the aggregated bike flows (ABFs). Specifically, based on $\mathbf{A}^t$, we define that $\mathbf{B}^t = (b^t_{kl}) \in \mathbb{R}^{K^2}$ where $b^t_{kl} = \sum_{ij} a^t_{ij} I(i \in C_k, j \in C_l), k, l = 1, \ldots, K$, K is the number of clusters, and $I(i \in C_k, j \in C_l)$ is the indicator function that equals to one only if the i-th station is in the k-th cluster (denoted as C_k) and the j-th station is in the l-th community (denoted as C_l). Denote $\tilde{\mathbf{x}}_t = (b^t_{11}, \ldots, b^t_{1K}, b^t_{21}, \ldots, b^t_{2K}, \ldots, b^t_{KK})^\top \in \mathbb{R}^{K^2}$, and $\mathbf{X} = (\tilde{\mathbf{x}}_1, \tilde{\mathbf{x}}_2, \ldots, \tilde{\mathbf{x}}_T)^\top \in \mathbb{R}^{T \times P}$ where $P = K^2$. The columns $\mathbf{x}_p$, $p = 1, \ldots, P$ of $\mathbf{X}$ are the time series of the p-th bike flow in a BSS that is called *aggregate bike flows*(ABFs).

By applying PCA to $\mathbf{X}$, a low intrinsic dimensionality of the ABFs could be found in both BSS datasets, as shown in the scree plots in Fig 4. This indicates that a vast majority of the temporal variability of the ABFs is contributed by the first few eigen-bike-flows (around 5), which is much lower than the number of ABFs. As shown in Fig 5, we randomly select two ABFs and show that the two ABFs can be sufficiently approximated by the top 5 eigen-bike-flows. This observation could be generalized on all ABFs, as shown in Fig 6 that the relative reconstruction errors (RRE) via the first k eigen-bike-flows decrease dramatically as k increases, where $\text{RRE} = \|\hat{\mathbf{X}}_k - \mathbf{X}\|_F / \|\mathbf{X}\|_F$, $\|\mathbf{X}\|_F = (\sum_{i,j} X^2_{ij})^{1/2}$ and $\hat{\mathbf{X}}_k$ is denoted by Eq 8.

Taxonomy of the eigen-bike-flows

The aforementioned analysis of the ABF data emphasizes the central role of eigen-bike-flows in understanding the ABFs. It seems that the eigen-bike-flows can be divided into two categories: deterministic eigen-bike-flows (d-flows) and spike eigen-bike-flows (s-flows). To show this, randomly selected d-flows and s-flows from the two BSS data sets are shown in Figs 7 and 8. The d-flows in Fig 7 show periodic trends. These periodicities are reflected by the hourly (rush hour and off-peek time) and diurnal (weekday and weekend) activities. On the other hand, the s-flows in Fig 8 illustrate certain short-lived spikes, which may correspond to occasional burst of usage due to holidays or particular weather conditions.

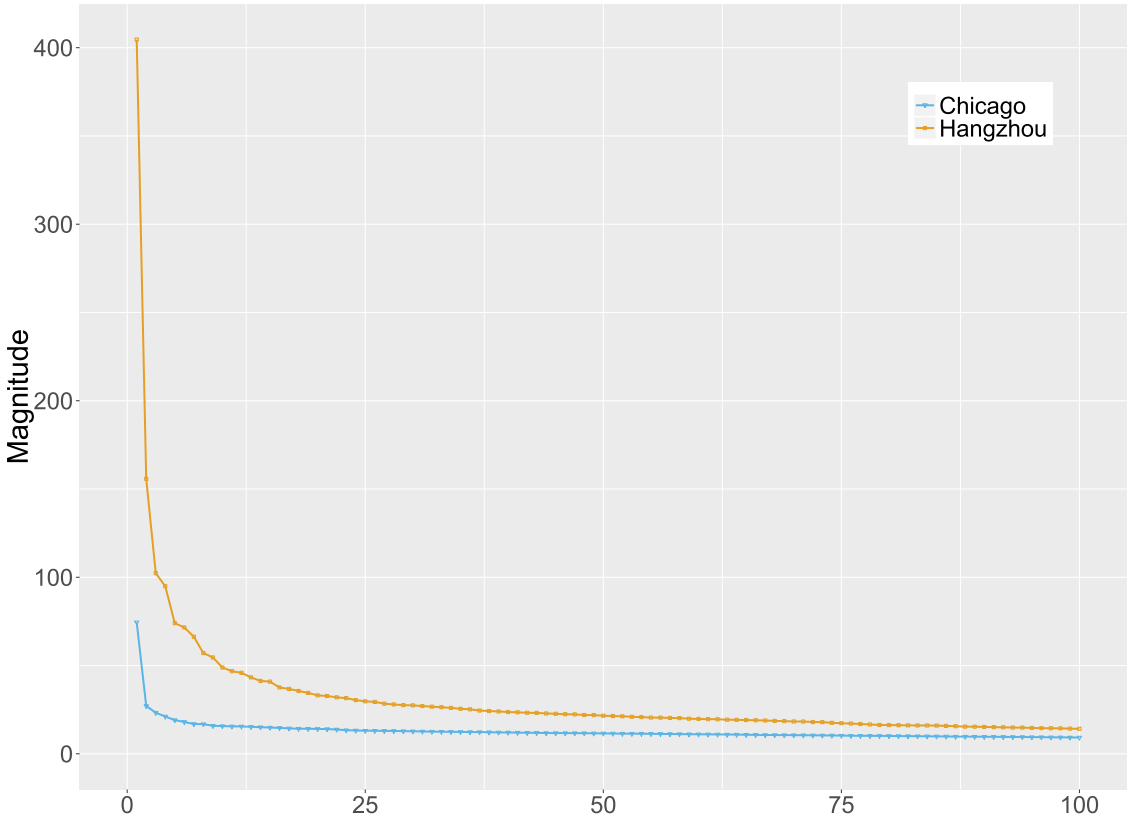

Fig 4. Scree plots for ABFs of Chicago and Hangzhou BSSs.

https://doi.org/10.1371/journal.pone.0193795.g004

To detect the d-flows, we conduct Fourier analysis of the eigen-bike-flows. The fourth column of Fig 7 shows that the spectrum of the selected d-flows all exhibit a significant standalone peak at twenty-four hours. To find the s-flows, the 5-sigma rule can be employed: whether there is a point whose distance from the mean exceeds 5 times standard deviations. The category of each eigen-bike-flow can be determined according to the criteria aforementioned. However, there are eigen-bike-flows who belong to more than one category. To overcome this contradiction, we define the d-flows as the eigen-bike-flows that have a significant standalone peak in the spectrum regardless the existence of the spikes.

Statistical representational power of the eigen-bike-flows

As shown in the Method Section, each ABF can be reconstructed as a weighted sum of eigen-bike-flows (e.g., see Eq 6). Particularly, each row of the principal matrix **V** specifies the extent to which each eigen-bike-flow contributes to the corresponding ABF. Thus, we are interested to examine the rows of **V** to discern the structure of the ABFs. Particularly, for an ABF, the entries of the corresponding row of **V** whose magnitudes are remarkably larger than a threshold indicate the significant eigen-bike-flows that constitute the ABF. Here, we set the threshold as $1/\sqrt{P}$, i.e., this is because that, in an extreme situation that all the eigen-bike-flows contribute to one ABF equally, all the entries of the corresponding row of **V** will be $1/\sqrt{P}$ due to the unit norm constrain of the columns of **V**.

Furthermore, CDFs of the number of entries which exceed $1/\sqrt{P}$ in their magnitudes are shown in Fig 9. The figure indicates that overall each ABF only has a small set of constitutional

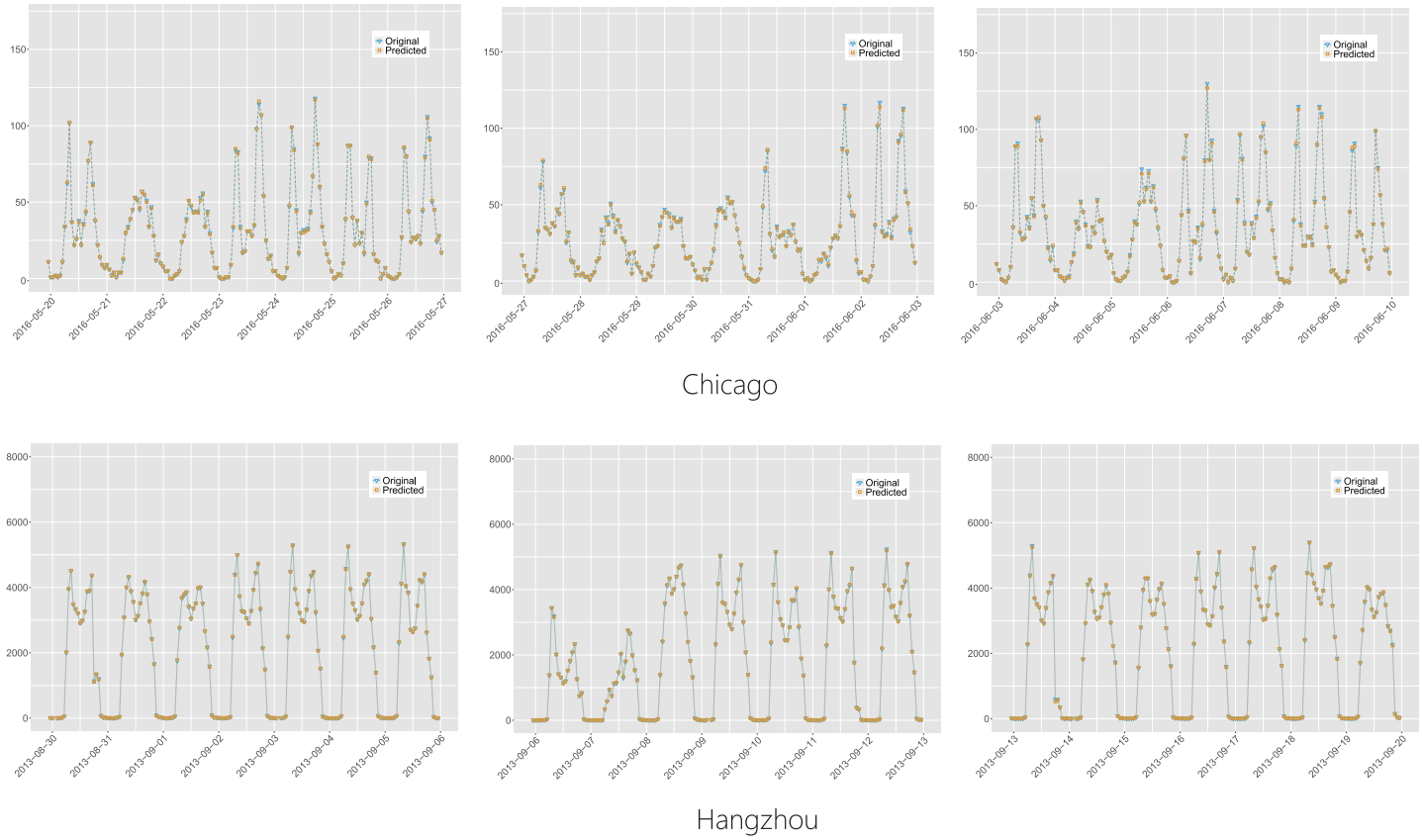

Fig 5. Reconstructing two ABFs with 5 principal components. For presentational simplicity, we only exhibit 3 weeks of the ABFs data.

https://doi.org/10.1371/journal.pone.0193795.g005

eigen-bike-flows. We further show the entries of $\mathbf{V}$ whose magnitudes exceed the threshold for the two data sets in Fig 10, after the rows of $\mathbf{V}$ are sorted by the variance of their corresponding ABFs. The top rows in each plot indicate the eigen-bike-flows that are significant in forming the strongest ABFs, and the bottom rows show the significant eigen-bike-flows for the weakest ABFs. Two interesting observations can be drew. First, from the vertical direction, the elements of one ABF are clustered in a small region. Second, from the horizontal direction, the elements of the ABF with larger variance are mainly top eigen-bike-flows, while the ones with smaller variance are mainly less significant eigen-bike-flows.

Temporal stability of the bike flow structure

It is of interest to see if the bike flow structure revealed in aforementioned sections could remain stable over time. To examine its temporal stability, here, we divide the measurement matrix $\mathbf{X} \in \mathbb{R}^{T \times P}$ into $\mathbf{X}_1 \in \mathbb{R}^{T_1 \times P}$ and $\mathbf{X}_2 \in \mathbb{R}^{T_2 \times P}$ where $T = T_1 + T_2$. Then, we apply PCA on $\mathbf{X}_1$ and use the obtained eigen-structure to predict $\mathbf{X}_2$. Our rationale is that, if the eigen-structure learned from $\mathbf{X}_1$ is stable, then it could show significant prediction power for $\mathbf{X}_2$. Details of how we could leverage the eigen-structure learned from $\mathbf{X}_1$ to predict $\mathbf{X}_2$ in shown in the Section Methods. The performance of the one-step prediction of $\mathbf{X}_2$ is shown in Fig 11, which shows the root mean square error (RMSE) per ABF in $\mathbf{X}_2$, while the ABFs are ordered with decreasing variances from left to right and $T_1 = T_2 = T/2$. From Fig 11, it can be seen that the eigen-structure learned from $\mathbf{X}_1$ could lead to accurate prediction of $\mathbf{X}_2$. Accurately forecasting the ABFs will no doubt benefit many decision-makings for managing the BSSs such as station

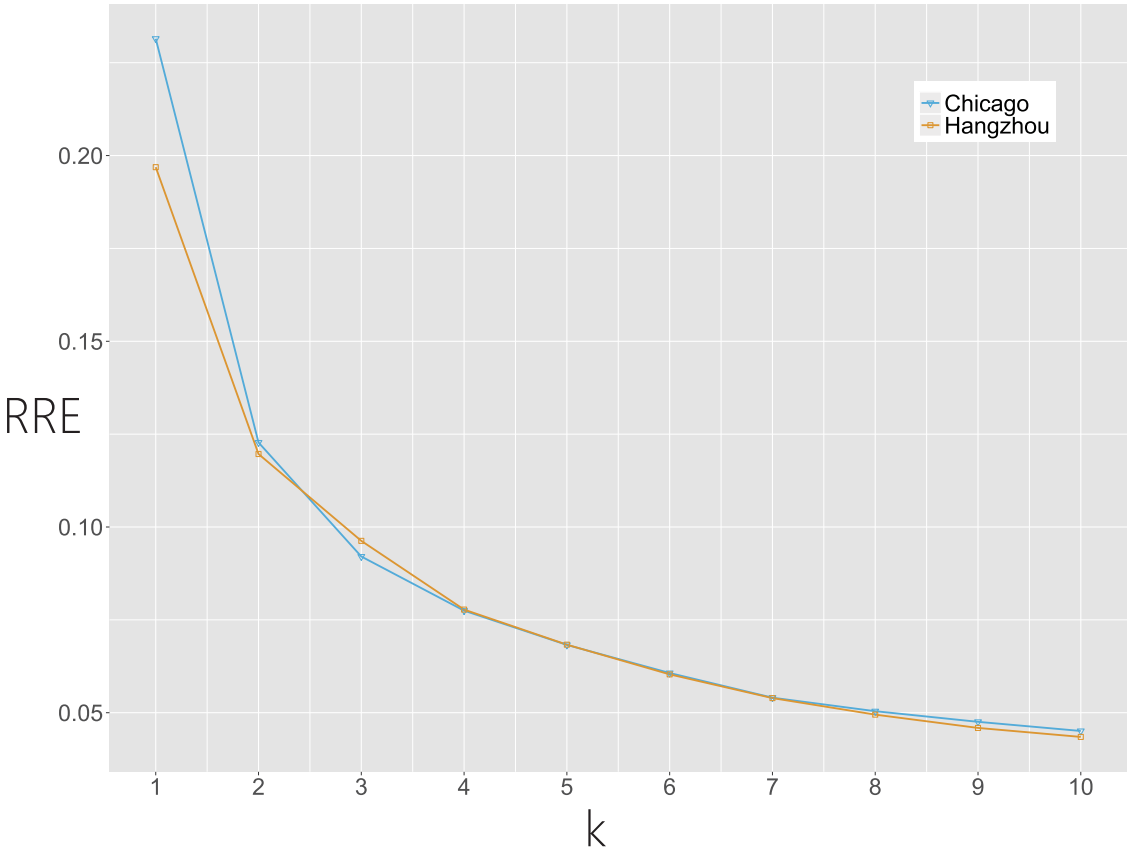

Fig 6. Relative reconstruction errors via the first $k = 1, 2, \ldots, 10$ eigen-bike-flows.

https://doi.org/10.1371/journal.pone.0193795.g006

placement [14, 15] and bike reallocation [19, 20]. The commonly accepted approaches in bike flow forecasting consider each ABF as a time series, and then, use some time series models such as the Autoregressive Integrated Moving Average (ARIMA for short) method [25] to predict the bike flows. Here, Fig 11 also shows that the performance of the PCA-based prediction model is better than ARIMA model for most ABFs of $\mathbf{X}_2$.

Discussion

The emerging bike sharing systems provide a new data source for us to understand human mobility. A unique value of this new data, comparing with existing mobility datasets such as GPS trajectory [6–8] and mobile phone data [3, 9, 10], lies on its characterization of human mobiliy in short-distance trips. Thus, as a deep understanding of the population on short-distance trip is currently lacking, we conduct a systematic analysis of the bike flow data collected in two major cities in the United States and China. Understanding the statistical characteristics of bike flow data holds great potential to develop informed decision-makings for better traffic prediction, infrastructure design such as station placement, real-time bike reallocation, and inventory management.

By analyzing the bike flow datasets from the two cities, we found statistical regularities underlying the irregular surface of the bike flow data. Our basic approach is inspired by the recognition of the spatial organizing principles of the bike flow, such that stations could be first clustered into distinct clusters. Then, by using PCA to analyze the aggregated bike flow data on the cluster level, we could identify a taxonomy of constituting eigen-flows that

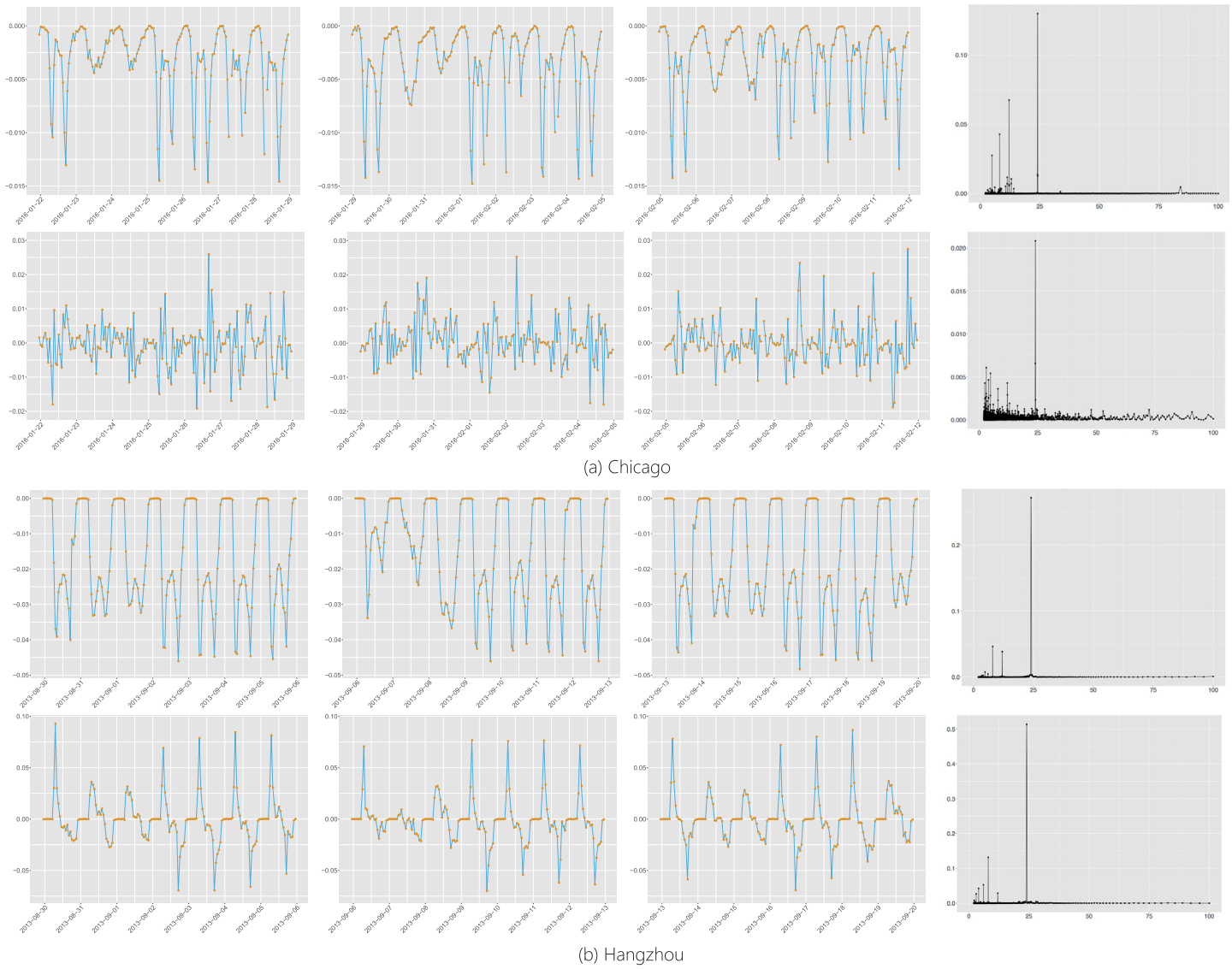

Fig 7. The first three columns are the examples of d-flows. The fourth column is the periodograms of d-flows. The x-axis of periodograms indicates their periods.

https://doi.org/10.1371/journal.pone.0193795.g007

correspond to routine and outburst in the bike flow, i.e., (i) deterministic eigen-bike-flows, which capture the periodic trends that have been reported in previous works of other bike flow data of similar nature [13, 18]; (ii) spike eigen-bike-flows, which capture the occasional short-lived bursts [13, 18] in BSSs. Besides the interpretability of the eigen-flows, we also find out that with a small set of eigen-flows, the ABFs could be accurately reconstructed, demonstrating their statistical significance and efficiency. We further study the temporal stability of the eigen-structure by using it to predict on unseen bike flow. Thus, although irregularity could be observed from the surface of the data, regularity emerges when looking into the spatial structure embedded in data. Further, on top of the spatial structure, temporal regularity could also be detected. On what level we should interrogate the data and ask what questions seems to be a crucial precondition for us to properly understand the data and translate that understanding into better decision makings.

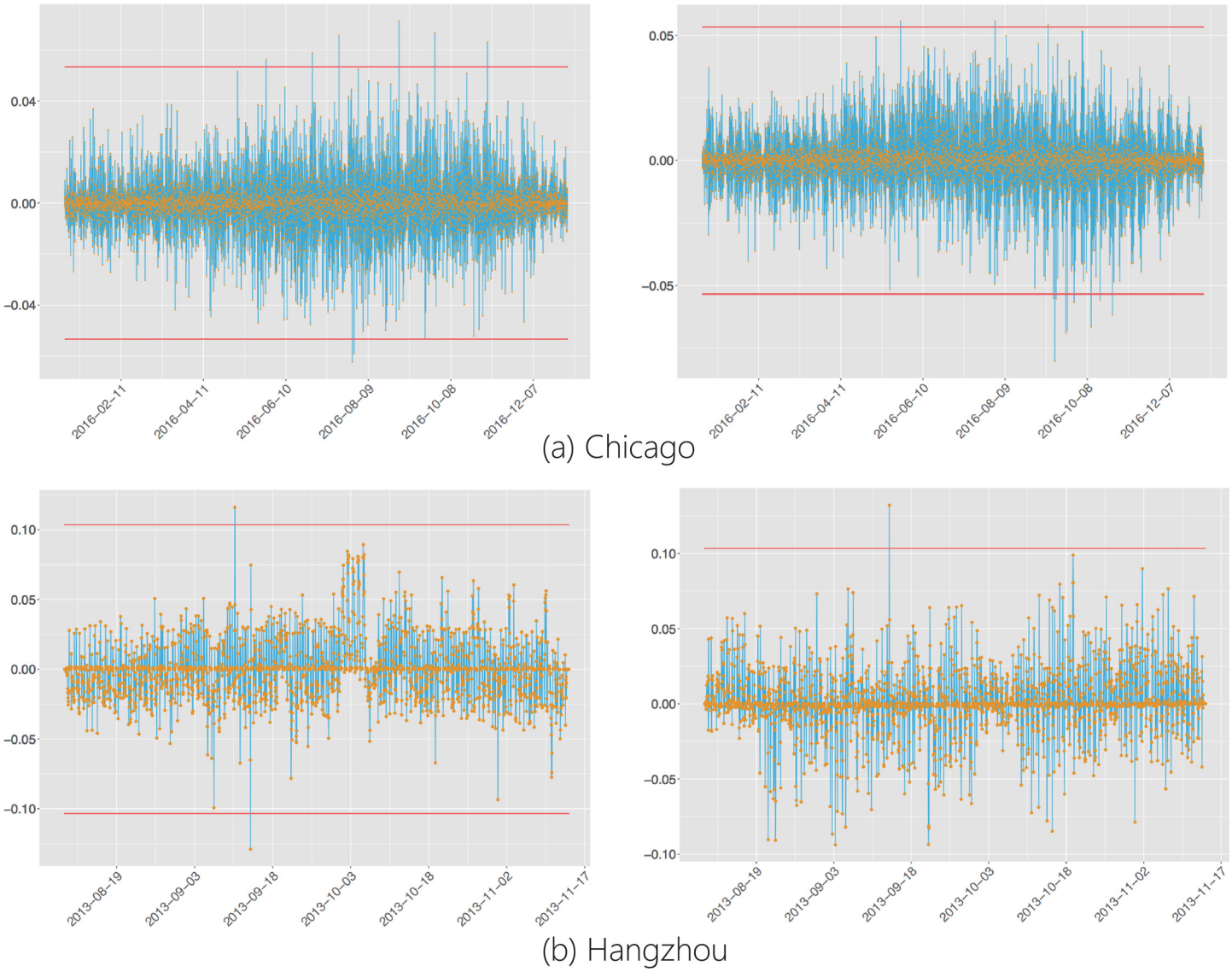

Fig 8. Examples of s-flows.

https://doi.org/10.1371/journal.pone.0193795.g008

In contrast to some popular assumptions made in some operations research and management methods for BSS that emphasize statistical simplicity and homogeneity, our analysis reveals far more intrinsic structures, heterogeneous patterns, and statistical complexities in both spatial and temporal scales in the bike flow data. Thus, our study anticipates further development of more realistic operations research and management methods which could optimize their performances to account for the unique statistical characteristics embedded in the BSS data. On the other hand, comparing with other existing works that used the bike flow data, we notice that most of the existing works largely focus on prediction using the bike flow data rather than inquiring the data for extracting system-level statistical patterns. Probably because of this, none of them aimed to conduct a delicate analysis of the variation structure in the bike flow data. One exception [13] in these prediction works exploited the idea of first

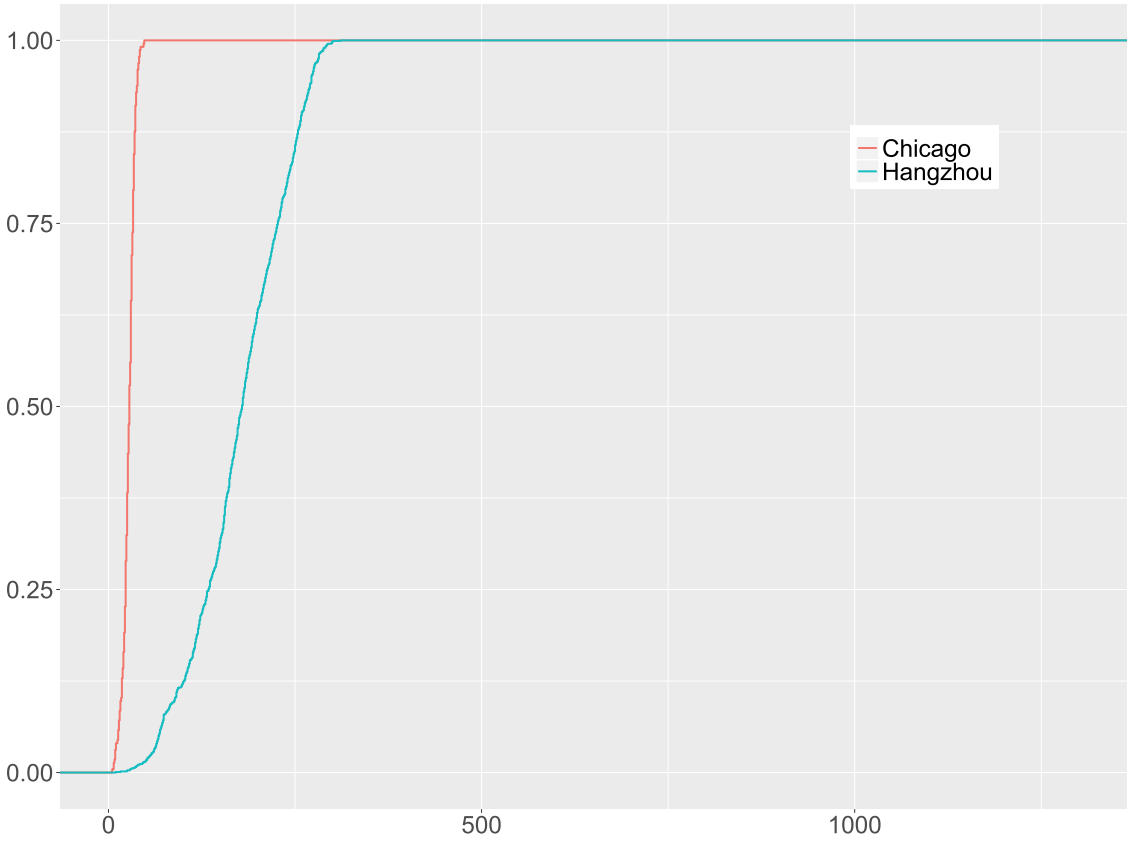

Fig 9. Number of eigen-bike-flows that constitute each ABF.

https://doi.org/10.1371/journal.pone.0193795.g009

spatially clustering the stations, and then, predicting on the combined bike flows very much like the ABF in our paper. While this study showed positive evidences to backup our finding, it was motivated by gaining prediction accuracy rather than a systematic revelation of the spatial structure and temporal eigen-structure in our paper.

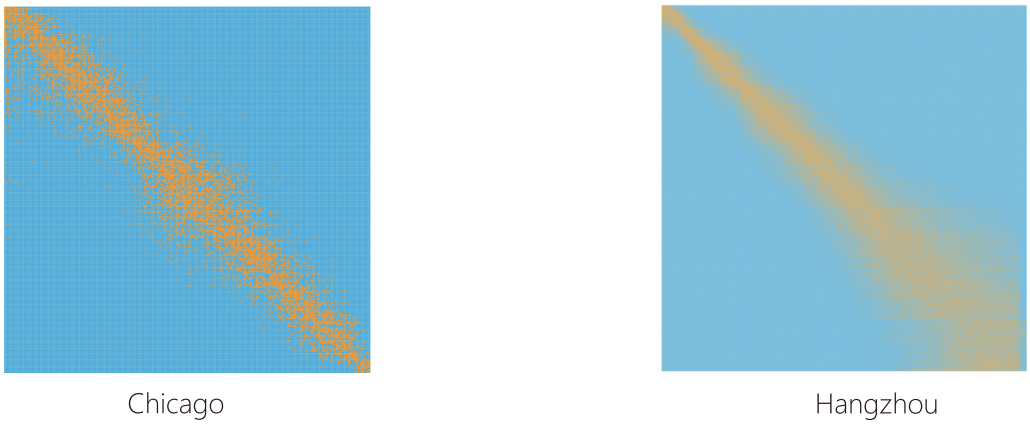

Fig 10. Indices of the eigen-bike-flows constituting each ABF. Note that the x-axis is the eigen-bike-flow index that are organized by convention in decreasing order of the singular values, and y-axis is ordered according to the decreasing ABF rate as well.

https://doi.org/10.1371/journal.pone.0193795.g010

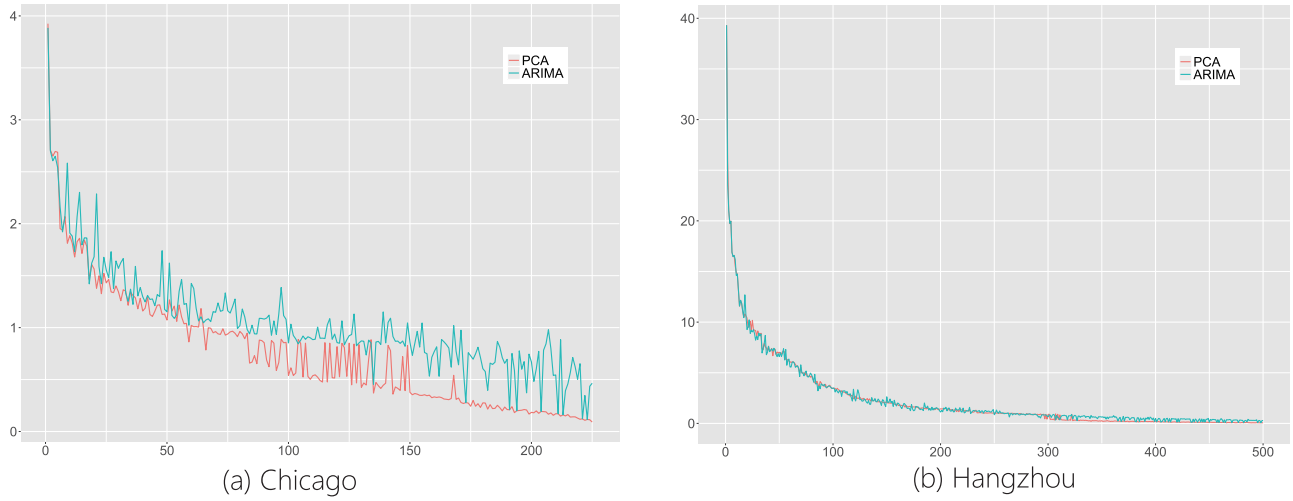

Fig 11. RMSE of each ABF to show the performance of ARIMA and PCA models. Note that the ABFs are ordered with decreasing variances from left to right. Furthermore, we did the Kolmogorov–Smirnov test between RMSE of ARIMA and PCA models. The p-values are all significantly less than 0.05 for Hangzhou and Chicago BSS respectively.

https://doi.org/10.1371/journal.pone.0193795.g011

In summery, this paper is, to the best of our knowledge, the first attempt to comprehensively investigate the human mobility patterns in short-distance trips, characterized by their manifestation on bike sharing systems. Consistent patterns have been discovered from datasets collected in two major cities in the US and China, implying that these patterns may represent universal conditions that shape the bike flow activities in real-world. This study has limitations. First, the methods used in this study, the classic hierarchical clustering and PCA methods, reveal interesting structures but also impose limitations on the structures they could identify. Particularly, as shown in the taxonomy of the eigen-bike-flows, some d-flows contain both periodic trends and spikes. This indicates the limitations of PCA and suggests that methods that can clearly separate the underlying signals could reveal further structures in the data. Second, while PCA is useful in analyzing the ABFs, more delicate time series analysis tools or signal processing methods could be used to study the dynamics embedded in the time series data. Last but not least, the two datasets used in this study may not fully present the complexity of the BSS data in other cities. While the observations made in this study are interesting and inspiring, this study lays the foundation for further inquiries such as traffic prediction, infrastructure design such as station placement, real-time bike reallocation, and inventory management.

Methods

Principle component analysis

PCA, as an unsupervised statistical learning method for studying the underlying structure in complex data, has been used for coordinate transformation and dimension reduction tasks [26, 27]. It maps the original data onto a new set of axes via coordinate transformation. The new axes are referred to principal components that point to the directions with the largest variance or energy in the data. Under the assumption that the most important structure exists along the new coordinate with the largest variance, the first few principal components may well capture the concerned structure in complex data. Due to the superiority of PCA, it has been widely used in many scientific fields, such as eigenfaces for recognition [28], network traffic analysis [29] and human mobility modeling [30].

Let $\mathbf{X} \in \mathbb{R}^{T \times P}$ be the measurement matrix. The p-th column denotes the p-th ABF and the t-th row represents an instance of all the ABFs at time t. Deriving the principal components is to solve the eigen-decomposition problem for matrix $\mathbf{X}^\top\mathbf{X}$, because $\mathbf{X}^\top\mathbf{X}$ measures the covariance between ABFs. The mathematical formulation is

$$\mathbf{X}^\top\mathbf{X}\mathbf{v}_p = \lambda_p\mathbf{v}_p, p = 1, \ldots, P \tag{1}$$

where λ_p is the p-th eigenvalue corresponding to eigenvector $\mathbf{v}_p$. Since $\mathbf{X}^\top\mathbf{X}$ is symmetric and semidefinite, its eigenvectors $\{\mathbf{v}_p\}_{p=1}^P$ are orthogonal and the corresponding eigenvalues $\{\lambda_p\}_{p=1}^P$ are nonnegative. It is required that the eigenvectors $\{\mathbf{v}_p\}_{p=1}^P$ have unit norm. Also, the eigenvalues are arranged from largest to smallest, i.e., $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_P \geq 0$. If eigenvalues $\{\lambda_p\}_{p=1}^P$ have r nonzero values, then the rank of $\mathbf{X}$ is r. It is well known that $\{\mathbf{v}_p\}_{p=1}^r$ are the principal components of $\mathbf{X}$. According to (1),

$$\mathbf{X}^\top\mathbf{X} = \mathbf{V}\Lambda\mathbf{V}^\top, \tag{2}$$

where $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_r]$ and $\Lambda = diag(\lambda_1, \ldots, \lambda_r)$. Calculating principal components actually is intimately related to Singular Value Decomposition (SVD) [31]. SVD is matrix decomposition tool that can be expressed in the form of matrix multiplication as

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^\top \tag{3}$$

where $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_r]$, $\mathbf{u}_i^\top\mathbf{u}_j = 0$ for $i \neq j$ and $\Sigma = diag(\sigma_1, \ldots, \sigma_r)$ is an $r \times r$ diagonal matrix with singular values $\{\sigma_p\}_{p=1}^r$ on the diagonal. Therefore,

$$\mathbf{X}^\top\mathbf{X} = \mathbf{V}\Sigma^2\mathbf{V}^\top. \tag{4}$$

Comparing with (2), we find that $\lambda_p = \sigma_p^2$. Furthermore, based on (3), it has

$$\mathbf{u}_p = \mathbf{X}\mathbf{v}_p/\sigma_p, p = 1, \ldots, r, \tag{5}$$

and

$$\mathbf{x}_p = \sigma_p\mathbf{U}(\mathbf{V}^\top)_p, p = 1, \ldots, P, \tag{6}$$

where $\mathbf{u}_p$, $p = 1, \ldots, r$ are vectors of size T and orthogonal by construction, and $(\mathbf{V}^\top)_p$ is the p-th row of $\mathbf{V}$. The Eq (5) indicates that all the ABFs can be transformed into a new coordinate with weights $\mathbf{v}_p$. $\mathbf{u}_p$ captures the temporal variation common to all flows along principal axis p. Specifically, $\mathbf{u}_1$ captures the strongest temporal trend common to all ABFs, $\mathbf{u}_2$ captures the second strongest, and so on so forth. The Eq (6) shows that each ABF is in turn a linear combination of the eigen-bike-flows, weighted by $(\mathbf{V}^\top)_p$.

Using SVD, a low-rank approximation matrix of $\mathbf{X}$ can be constructed as follows. The approximation form is

$$\hat{\mathbf{X}}_k = \mathbf{U}_k\Sigma_k\mathbf{V}_k^\top \tag{7}$$

where $\hat{\mathbf{X}}_k$ is an approximation of $\mathbf{X}$ with rank $k < r$, $\mathbf{U}_k$ and $\mathbf{V}_k$ are the first k columns of $\mathbf{U}$ and $\mathbf{V}$, respectively, and Σ_k is the top-left part of Σ of size k. The low-rank approximation of $\mathbf{X}$ is actually a dimension reduction approach via PCA with the form

$$\hat{\mathbf{X}}_k = \sum_{p=1}^{k} \sigma_p\mathbf{u}_p\mathbf{v}_p^\top. \tag{8}$$

Prediction of bike flow

By assuming temporal stability of the bike flow data, we could leverage the eigen-structure revealed in a previous data, denoted as $\mathbf{X}_1$, to predict the data in next, denoted as $\mathbf{X}_2$. According to SVD, $\mathbf{X}_1 = \mathbf{U}_1\Sigma_1\mathbf{V}_1^\top$ where each column of $\mathbf{U}_1$ is an eigen-bike-flow and each column of $\mathbf{V}_1$ is a principal component of $\mathbf{X}_1$. We denote that $\mathbf{U}_1^d$ as the d-flows of $\mathbf{X}_1$ which have periodic trends. Therefore, we can apply the ARIMA model on each column of $\mathbf{U}_1^d$ to predict the estimated d-flows of $\hat{\mathbf{X}}_2^d$. Then, based on the temporal stability assumption of the principal components, the estimation of $\hat{\mathbf{X}}_2$ can be constructed by

$$\hat{\mathbf{X}}_2 = \hat{\mathbf{U}}_2^d\Sigma^d(\mathbf{V}_1^d)^\top \tag{9}$$

where Σ^d and $\mathbf{V}_1^d$ are sub-matrices of Σ and $\mathbf{V}$, respectively, that correspond to the d-flows in $\mathbf{X}_1$. The prediction performance could be evaluated by RMSE. Here, the RMSE is defined as

$$RMSE_i = 1/\sqrt{T_2}\|\mathbf{X}_2^i - \hat{\mathbf{X}}_2^i\|, i = 1, \ldots, d \tag{10}$$

where T_2 is the number of rows in $\mathbf{X}_2$, d is the number of d-flows, $\mathbf{X}_2^i$ and $\hat{\mathbf{X}}_2^i$ are the i-th columns of $\mathbf{X}_2$ and $\hat{\mathbf{X}}_2$ respectively.

Supporting information

S1 Dataset. BSS data of Chicago.
(CSV)

S2 Dataset. BSS data of Hangzhou.
(ZIP)

Acknowledgments

Chang was partially supported by the National Natural Science Foundation of China (Project No. 11771012, 91546119) and the Major Program of National Natural Science Foundation of China (Project No. 71731009, 71742005). Lu was partially supported by the National Natural Science Foundation of China (Project No. 61502342). The authors also acknowledge funding support from the National Science Foundation under Grant CMMI-1536398.

Author Contributions

Conceptualization: Xiangyu Chang, Xiaoling Lu, Shuai Huang.

Data curation: Jingzhou Shen, Xiaoling Lu.

Formal analysis: Xiangyu Chang, Jingzhou Shen, Shuai Huang.

Investigation: Xiangyu Chang, Shuai Huang.

Methodology: Xiangyu Chang, Shuai Huang.

Project administration: Xiangyu Chang.

Validation: Jingzhou Shen.

Visualization: Jingzhou Shen.

Writing – original draft: Xiangyu Chang.

Writing – review & editing: Xiangyu Chang, Shuai Huang.

References

1. Brockmann D, Hufnagel L, Geisel T. The scaling laws of human travel. Nature. 2006; 439:462–465. https://doi.org/10.1038/nature04292 PMID: 16437114
2. González MC, Hidalgo CA, Barabási AL. Understanding individual human mobility patterns. Nature. 2008; 453(7196):779–782. https://doi.org/10.1038/nature06958 PMID: 18528393
3. Song C, Qu Z, Blumm N, Barabási AL. Limits of predictability in human mobility. Science. 2010; 327 (5968):1018–1021. https://doi.org/10.1126/science.1177170 PMID: 20167789
4. Jiang S, Ferreira J, González MC. Clustering daily patterns of human activities in the city. Data Mining and Knowledge Discovery. 2012; p. 1–33.
5. Pappalardo L, Simini F, Rinzivillo S, Pedreschi D, Giannotti F, Barabási AL. Returners and explorers dichotomy in human mobility. Nature communications. 2015; 6. https://doi.org/10.1038/ncomms9166 PMID: 26349016
6. Giannotti F, Nanni M, Pinelli F, Pedreschi D. Trajectory pattern mining. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2007. p. 330–339.
7. Monreale A, Pinelli F, Trasarti R, Giannotti F. Wherenext: a location predictor on trajectory pattern mining. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM; 2009. p. 637–646.
8. Pappalardo L, Rinzivillo S, Qu Z, Pedreschi D, Giannotti F. Understanding the patterns of car travel. The European Physical Journal Special Topics. 2013; 215(1):61–73. https://doi.org/10.1140/epjst/e2013-01715-5
9. Phithakkitnukoon S, Smoreda Z, Olivier P. Socio-geography of human mobility: A study using longitudinal mobile phone data. PloS one. 2012; 7(6):e39253. https://doi.org/10.1371/journal.pone.0039253 PMID: 22761748
10. Wang P, González MC, Hidalgo CA, Barabási AL. Understanding the spreading patterns of mobile phone viruses. Science. 2009; 324(5930):1071–1076. https://doi.org/10.1126/science.1167053 PMID: 19342553
11. DeMaio P. Bike-sharing: History, impacts, models of provision, and future. Journal of Public Transportation. 2009; 12(4):3. https://doi.org/10.5038/2375-0901.12.4.3
12. Zhang J, Philip SY. Trip Route Planning for Bicycle-Sharing Systems. In: Collaboration and Internet Computing (CIC), 2016 IEEE 2nd International Conference on. IEEE; 2016. p. 381–390.
13. Li Y, Zheng Y, Zhang H, Chen L. Traffic prediction in a bike-sharing system. In: Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems. ACM; 2015. p. 33.
14. Chen L, Zhang D, Pan G, Ma X, Yang D, Kushlev K, et al. Bike sharing station placement leveraging heterogeneous urban open data. In: Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing. ACM; 2015. p. 571–575.
15. Liu J, Li Q, Qu M, Chen W, Yang J, Xiong H, et al. Station site optimization in bike sharing systems. In: Data Mining (ICDM), 2015 IEEE International Conference on. IEEE; 2015. p. 883–888.
16. O'Mahony E, Shmoys DB. Data Analysis and Optimization for (Citi) Bike Sharing. In: AAAI; 2015. p. 687–694.
17. Etienne C, Latifa O. Model-based count series clustering for bike sharing system usage mining: a case study with the Vélib'system of Paris. ACM Transactions on Intelligent Systems and Technology (TIST). 2014; 5(3):39.
18. Zhang J, Pan X, Li M, Philip SY. Bicycle-sharing system analysis and trip prediction. In: Mobile Data Management (MDM), 2016 17th IEEE International Conference on. vol. 1. IEEE; 2016. p. 174–179.
19. Pfrommer J, Warrington J, Schildbach G, Morari M. Dynamic vehicle redistribution and online price incentives in shared mobility systems. IEEE Transactions on Intelligent Transportation Systems. 2014; 15(4):1567–1578. https://doi.org/10.1109/TITS.2014.2303986
20. Shu J, Chou MC, Liu Q, Teo CP, Wang IL. Models for effective deployment and redistribution of bicycles within public bicycle-sharing systems. Operations Research. 2013; 61(6):1346–1359.
21. Raviv T, Kolka O. Optimal inventory management of a bike-sharing station. IIE Transactions. 2013; 45 (10):1077–1093. https://doi.org/10.1080/0740817X.2013.770186
22. Lin JR, Yang TH, Chang YC. A hub location inventory model for bicycle sharing system design: Formulation and solution. Computers & Industrial Engineering. 2013; 65(1):77–86. https://doi.org/10.1016/j.cie.2011.12.006

23. Girvan M, Newman ME. Community structure in social and biological networks. Proceedings of the national academy of sciences. 2002; 99(12):7821–7826. https://doi.org/10.1073/pnas.122653799
24. Fortunato S. Community detection in graphs. Physics reports. 2010; 486(3):75–174. https://doi.org/10.1016/j.physrep.2009.11.002
25. Box GE, Jenkins GM, Reinsel GC, Ljung GM. Time series analysis: forecasting and control. John Wiley & Sons; 2015.
26. Hotelling H. Analysis of a complex of statistical variables into principal components. Journal of educational psychology. 1933; 24(6):417. https://doi.org/10.1037/h0071325
27. Friedman J, Hastie T, Tibshirani R. The elements of statistical learning. vol. 1. Springer series in statistics Springer, Berlin; 2001.
28. Turk M, Pentland A. Eigenfaces for recognition. Journal of cognitive neuroscience. 1991; 3(1):71–86. https://doi.org/10.1162/jocn.1991.3.1.71 PMID: 23964806
29. Lakhina A, Papagiannaki K, Crovella M, Diot C, Kolaczyk ED, Taft N. Structural analysis of network traffic flows. In: ACM SIGMETRICS Performance evaluation review. vol. 32. ACM; 2004. p. 61–72.
30. Sun J, Wang Y, Si H, Yuan J, Shan X. Aggregate Human Mobility Modeling Using Principal Component Analysis. JoWUA. 2010; 1(2/3):83–95.
31. Jennings A, McKeown JJ. Matrix computation. John Wiley & Sons Inc; 1992.