ARTICLE IN PRESS

Journal of Transport Geography xxx (2013) xxx–xxx

Contents lists available at SciVerse ScienceDirect

Journal of Transport Geography

journal homepage: www.elsevier.com/locate/jtrangeo

Mining bicycle sharing data for generating insights into sustainable transport systems

Oliver O'Brien *, James Cheshire, Michael Batty

Centre for Advanced Spatial Analysis, University College London, Gower Street, London WC1E 6BT, United Kingdom

ARTICLE INFO

Keywords:
Bike-sharing
Cycling
Cities
Commuters

ABSTRACT

Bicycle sharing systems (bike-shares) are becoming increasingly popular in towns and cities around the world. They are viewed as a cheap, efficient, and healthy means of navigating dense urban environments. This paper is the first to take a global view of bike-sharing characteristics by analysing data from 38 systems located in Europe, the Middle East, Asia, Australasia and the Americas. To achieve this, an extensive database depicting the geographical location and bicycle occupancy of each docking station within a particular system has been created over a number of years to chart the usage in the chosen systems (and others) and provide a consistent basis on which to compare and classify them. Analysis of the variation of occupancy rates over time, and comparison across the system's extent, infers the likely demographics and intentions of user groups. A classification of bike-shares, based on the geographical footprint and diurnal, day-of-week and spatial variations in occupancy rates, is proposed. The knowledge of such patterns and characteristics identifiable from the dataset has a range of applications, including informing operators and policymakers about the maintenance of a suitable balance of bicycles throughout the system area (a nontrivial problem for many bike-shares), the location of new docking stations and cycle lanes, and better targeting of promotional materials to encourage new users. Within the context of transport research, the systems utilised here are part of relatively small, closed environments that can be more easily modelled and validated. Such work lays foundations for the analysis of larger scale transport systems by creating a classification of the different systems and seeks to demonstrate that bike-shares have a lot to offer both as an effective method of transport and a rich source of data.

1. Introduction

Bicycle sharing systems (bike-shares) are a relatively new form of transport in many urban areas. There are around 450 such systems currently operating worldwide (DeMaio and Meddin, 2012) and analysing the relatively freely available data for many of them generates insights into the habits of their users and, by proxy, movements within urban areas (Padgham, 2012).

Bike-shares are typically single systems located in and around the commercial or business centres of their host towns and cities. Exceptions include suburb-based systems such as Mexico City (Kazis, 2012) or Taipei City (Tso, 2009), or those extending well beyond the city core, as is the case for Barcelona. Current bike-shares [characterised in the literature as "third-generation" (see DeMaio, 2009; Haverman, 2010)], make use of technology to operate on a largely automated basis. There are two typical configurations – the docking point/docking station model where bicycles are hired from a docking point in one of a fixed number of docking stations in the host area, then later returned to a docking point within another (or the same) docking station; and the free-placement model where bicycles are obtained typically from crossroads in the system area. This paper will focus on systems using the former configuration as, having a fixed set of locations for the start and end of each journey with a measureable number of bicycles at these locations, such systems tend to produce more readily available and usable data. An example implementation of the latter approach, not studied further here, is Berlin's Callabike system.

Bike-shares attract a range of users from professional commuters to students, local residents running errands, leisure users and tourists (JZTI, 2010, p. 36–7). System operators can influence usage behaviour by, for example, prohibiting certain user types, such as in Barcelona, where users must live in the city (introduced to avoid a perceived impact on an existing manual tourist cycle hire business) (OBIS, 2011, p. 14). The result is that the temporal characteristics of the dominant flows of cyclists will vary between systems presenting each operator with a unique set of challenges to ensure the bicycles are appropriately distributed to meet demand. For example, bike-shares that have a dominant commuter pattern, such as that in London, often suffer from particularly asymmetric flows, making effective redistribution an important part of the system's success.

* Corresponding author. Tel.: +44 203 108 3883.
E-mail address: o.obrien@ucl.ac.uk (O. O'Brien).

0966-6923/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.jtrangeo.2013.06.007

The latest bike-share systems enable users to monitor cycle availability and docking station spaces via near real-time online maps. These websites often specify and supply an applications programming interface (API) for external software developers to access the underlying data. In addition, a number of system operators release datasets pertaining to individual journeys made over a particular time period. Both types of data offer insights in the usage of particular bike-shares and provide a ready basis for utilisation in transport research. A small number of previous studies have been undertaken and generally concern the characteristics of a single city's system, often with a focus on user demographics. Jensen et al. (2010), for example, analysed 11.6 million journeys of the Vélo'v bicycle sharing system in Lyon, constructing a map showing the likely flows of the bicycles across the city. Several characteristics emerged; namely greatly enhanced usage during public transport strikes, and variations in average speeds through the day such as for example, a small but significant increase in speed just before 9 a.m. as cycle commuters hurry to complete their journeys before the start of normal working hours. One intriguing result was that the average speed during the morning commute was greatest on Wednesdays, the authors conjecturing that this was due to a greater proportion of users on Wednesdays being men, due to the tradition of at-home childcare by women on this day.

Elsewhere in Europe, Barcelona's Bicing bike-share exhibits five spatial clusters of docking stations based on activity (i.e. usage) variation throughout the day and six separate spatial clusters based on the intra-day change of availability of bicycles in each docking station (Froehlich et al., 2009). In addition, Kaltenbrunner et al. (2010) looked at the system's usage patterns across seven weeks, and also developed a simple model to predict future trends. Like Froehlich et al. (2009), they used docking station data rather than data on individual journeys. Differences between weekday and weekend usage were apparent, and peak usages at different parts of the day depended on the proximity of each docking station to retail, academic and workplace locations.

More recently, Lathia et al. (2012) published results from the London bicycle sharing system's docking station data and as with Kaltenbrunner et al. (2010) and Jensen et al. (2010) characteristic usage peaks and significant weekday/weekend differences emerged. The research focused on the change in the usage patterns following the introduction of "casual" usage, where credit cards could be used in place of a dedicated key. He also identified six clusters of docking stations, grouped by similar intra-day usage patterns, and observed slight changes to these clusters once the casual usage of the system was introduced. The clusters were found to be grouped spatially, and showed a distinctive "ring and core" structure. Finally, Vogel et al. (2011) collected data for 0.74 million journeys undertaken in the Citybike bicycle sharing system in Vienna. From these, five spatially similar groups of docking stations emerged thus suggesting, in line with many of the studies cited above, distinct groups of users using the bicycles at similar times and for similar journeys.

This paper is the first to take a global view of bike-sharing patterns by analysing data from 38 systems located in Europe, the Middle East, Asia, Australasia and the Americas. To achieve this, an extensive database has been created over a number of years to chart the usage in the chosen systems (and others) and this offers a consistent basis on which to compare and classify them.

After outlining the method used to obtain and process the data, this paper discusses various metrics which can be used to gain insights into and to classify each bicycle sharing system, based both on non-spatial and spatial attributes of the docking station locations and temporal usage statistics. A tentative qualitative classification, based on the observed metrics, is proposed. The paper concludes by discussing potential further applications of the data studied, such as demographic analysis and the role of, and benefit to, operator redistribution activity.

2. Managing docking station data

The data are collected automatically (normally from operator-run websites) and include locations, capacity and current load factor of docking stations, for various systems around the world. A script, written in the Python programming language, and customised for each system, is run on a regular basis to access the bike-share's docking station data online.

The load factor, the key measure in this study, is the proportion of docking points in each docking station that currently have a bicycle available to hire. It is normally calculated from the number of bicycles and the number of free spaces in each docking station, which is the basic statistic for each docking station. The load in the "load factor" term therefore is a reference to a load of bicycles filling docking points – rather than a load of bicycles from the system being used on the streets. Systems that do not make this information available online – a key metric for users trying to discover bicycles or free spaces in their vicinity – are not included in this study.

It is recognised that the variation in load factor is not a perfect measure of the performance or popularity of a system. Theoretically, systems that are very well used but very quick to redistribute bicycles back to points of need will show a similar variation in load factor to those systems that are poorly used. In reality, the practicalities and costs of ensuring the rate and scale of bicycle redistribution required to alter the load factor in this way are prohibitive. On this basis we feel the load factor metric remains the most appropriate for this study.

In many cases, the data are extracted from embedded online mapping Application Programming Interface (API) instances (normally the Google Maps API) which typically contain a collection of pin-style markers, representing the docking stations, with the capacity and load factor appearing as statistics attached to each marker. In some cases, the operators provide dedicated APIs, typically in XML or JSON format. Such data streams are often used for mobile phone applications or dashboard monitoring of the system concerned. In some cases there are practical or technical difficulties obtaining the data in a timely fashion directly from the system operator. In such cases third party APIs, often run by volunteers based in the city concerned, have been used to access the data in a standardised format.

The data are typically collected every two minutes, except where the system's server is slow to respond, in which case the data is collected every 10 or 20 min. This frequency is sufficient to accurately show the activity and availability changes throughout the day, highlighting commuter "rush hours" and other features. Our database covers a period of up to 2 years and over 80 cities. It is therefore the most comprehensive of its kind. For this study we focus on data collected throughout September 2012.

A few systems also provide journey origin–destination data, on a historical basis. The data are normally provided by the operator on a bulk-load basis, rather than being queryable from an API or map. Such data are not used in this study because of the small number of systems which make the data available in this way, meaning that a comparative study is difficult.

3. Characterising global bike-shares based on their docking stations

This study seeks to compare and contrast the structure of various bike-shares, by looking at the "footprint" of their docking sta-