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

A R T I C L E I N F O

Keywords:
Bike-sharing
Cycling
Cities
Commuters

A B S T R A C T

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-

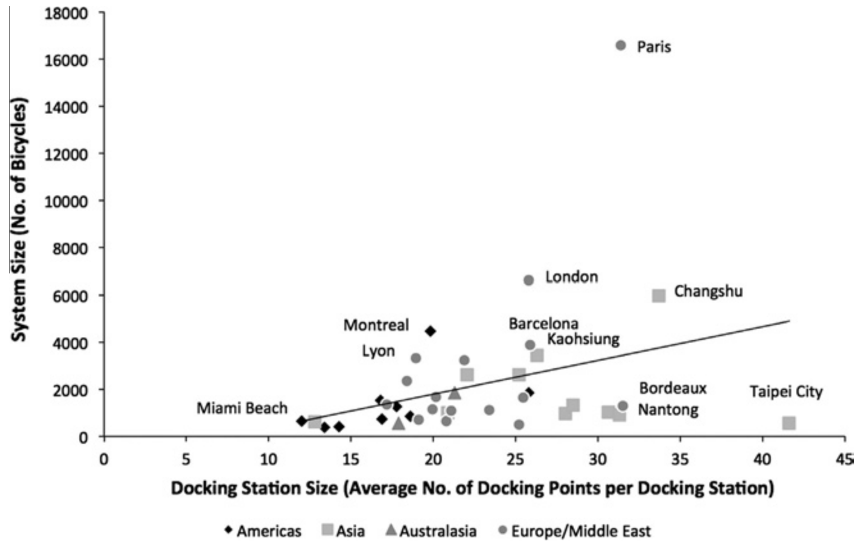

Fig. 1. System size and docking station size. The line shows the weak direct correlation.

tions, and the spatio-temporal changes in bicycle distribution within them.

Of the 38 bike-shares studied, 16 fall in Europe and the Middle East, 11 in Asia, 9 in the Americas and 2 in Australasia. The bike-shares studied have at least 40 docking stations, a clean feed of data and are each contained within a single city. A small number of the selected systems have been more recently set up for data collection, or have substantially changed in size during this collection period but their data, upon examination, have been found to be sufficient for inclusion.

Looking at the relative locations of docking stations in each system, and the diurnal and weekly variations in the aggregated load factor (the proportion of docking points across the docking stations with bicycles docked to them) a number of characteristics can be compared and contrasted. These features have been grouped into three types – aggregate characteristics that provide simple (non-spatial) measures for each system, spatial characteristics that look at the placing of docking stations within each system, and temporal characteristics.

Aggregate characteristics measured include the maximum number of docking stations, the maximum number of bicycles in the docking stations, and the biggest change between the daily maximum and minimum number of bicycles – a measure of maximum simultaneous use). From these, the maximum load factor and maximum intraday load factor change can be simply calculated. These characteristics are generally measured during September 2012.

Spatial characteristics of the docking station footprints analysed include the latitude of the centroid of the system, the system's area of influence, the observed mean distance between docking stations, the *Z*-score (which describes whether the system's footprint is statistically clustered, random or dispersed) and the compactness ratio (a measure of the system's shape compared with a theoretical circular footprint around its centre).

Temporal characteristics include the load factor and a normalised measure of the redistribution needed to even out the load across the system, and how both these measures vary on an intraday and weekday/weekend basis. The measures are obtained by regularly counting up the number of full and empty docking points, within each docking station. For the load factor measure, these are simply aggregated across the system. The redistribution measure compares the deviation of each docking station's load factor with the average across the system at that time.

The following sections detail the results of analysis of these three sets of characteristics in turn that are then used as a basis for classifying the different bike-share systems.

3.1. Aggregate characteristics

The size of a system can be expressed in terms of the number of docking stations, the number of docking points (which are grouped into docking stations) or the number of bicycles available to use in the system. As a bicycle is essential for use of the system, the latter metric is used here. This is measured by examining the number of bicycles available and taking the maximum number typically observed on a normal day during the period of study (shown in Fig. 1), that is a day where a typical cycle of usage is seen, so excluding days with special events. The number of bicycles available dips during periods of high usage such as at the beginning and end of a working day, and reaches a maximum typically around 3 a.m. A caveat is that some systems may be gradually expanding (although not significantly during the relatively short period of study here) or may have different numbers available on weekdays and at weekends, to manage differing usage levels and spatial patterns.

Operator redistributions can also affect availability although a temporal analysis of the load factor, which is detailed later, reveals that most systems see little change in the value during the early morning, indicating that operator redistribution does not contribute significantly to the low usage of bicycles during these quiet periods. Indeed, operators may not want to redistribute overnight as the morning commuters will likely want the bicycles to be in the same place that they left them the previous evening.

The system size can be contrasted with the docking station size – average number of docking points in each system's set of docking stations. Smaller-scale systems will often use smaller docking stations for economic reasons. There is a weak direct correlation ($R^2 = 0.1$, although rising to $R^2 = 0.3$ if excluding Asia), shown in Fig. 1, between docking station size and system size. For clarity, only a selection of cities are named in each figure in this section, but all cities, and their values plotted, are listed in Table 1.

Paris stands out as the largest system in this study – it is believed to be the third largest in the world at present behind two Chinese systems, Wuhan and Hangzhou (Jinran and Xiaodong, 2012), for which data are not currently available. It also has relatively large docking stations.

Table 1
Values measured and displayed in (Figs. 1, 2, 4 and 5). ME = Middle East.

System	Region	System size	Average docking station size	Maximum daily concurrent use (%)	Maximum load factor (%)	Area of system (km)	Average distance between docking stations (m)	Compactness ratio	Z-score
Boston	Americas	734	17	29.4	45.8	64	605	0.73	0.0
Denver	Americas	411	14	28.2	54.4	40	519	0.60	−2.6
Mexico City	Americas	1845	26	14.2	44.7	35	197	0.77	−3.6
Miami Beach	Americas	641	12	29.0	47.2	41	249	0.56	−4.5
Minneapolis	Americas	1266	18	11.5	49.2	117	707	0.47	−2.1
Montreal	Americas	4476	20	20.1	55.8	114	382	0.55	−4.2
Rio de Janeiro	Americas	382	13	87.7	50.0	55	580	0.51	−7.4
Toronto	Americas	850	19	20.7	57.1	26	374	0.92	2.6
Washington DC	Americas	1518	17	35.5	47.8	134	529	0.41	−7.7
Changshu	Asia	5966	34	17.3	50.3	127	306	0.48	−11.0
Changwon	Asia	2613	22	28.0	50.4	169	518	0.30	−16.0
Daejeon	Asia	619	13	45.0	46.5	60	389	0.49	−7.0
Heihe	Asia	971	28	34.9	55.9	23	390	0.80	−1.1
Kaohsiung	Asia	1013	31	47.8	44.7	90	714	0.37	−4.1
Nantong	Asia	898	31	27.1	47.8	35	444	0.79	0.5
Shaoxing	Asia	1328	28	38.0	57.6	43	397	0.74	−1.1
Suzhou	Asia	3434	26	18.9	41.1	192	436	0.34	−18.8
Taipei City	Asia	567	42	34.9	32.4	43	553	0.59	−0.8
Wujiang	Asia	998	21	21.7	46.9	51	353	0.66	−3.4
Zhongshan	Asia	2619	25	25.0	51.7	84	338	0.53	−6.7
Brisbane	Australasia	1854	21	4.8	57.7	37	222	0.84	−4.5
Melbourne	Australasia	548	18	11.5	61.2	33	501	0.80	0.9
Barcelona	Europe/ME	3871	26	36.3	35.8	77	247	0.78	−8.5
Bordeaux	Europe/ME	1305	32	31.5	48.7	115	941	0.30	−1.8
Brussels	Europe/ME	3241	22	15.2	52.4	118	601	0.70	5.5
Dublin	Europe/ME	489	25	36.7	44.1	16	451	0.92	4.4
Lille	Europe/ME	1664	20	23.9	45.4	78	510	0.54	−8.4
London	Europe/ME	6613	26	38.7	45.4	103	327	0.73	−1.8
Luxembourg City	Europe/ME	690	19	7.6	50.2	43	545	0.71	−2.8
Lyon	Europe/ME	3308	19	24.7	50.6	85	366	0.81	0.7
Milan	Europe/ME	1637	25	36.3	44.3	33	308	0.90	2.1
Nice	Europe/ME	1360	17	14.4	46.1	56	316	0.58	−10.5
Paris	Europe/ME	16,564	31	27.3	43.7	201	329	0.68	−0.4
Rennes	Europe/ME	649	21	17.0	37.6	47	584	0.79	3.2
Saragossa	Europe/ME	1087	21	28.1	40.3	36	377	0.89	3.8
Tel Aviv	Europe/ME	1153	20	30.9	38.5	76	455	0.64	−2.4
Valencia	Europe/ME	2337	18	31.2	45.4	64	369	0.78	6.7
Vienna	Europe/ME	1116	23	27.3	49.7	57	625	0.75	0.6

American systems generally build smaller docking stations (that is, with fewer docking points for each station, on average) when compared to other regions in this study, but this may reflect different manufacturer preferences and system specification as much as usage differences. By contrast, Asian systems generally have larger than average docking stations. Larger docking stations allow for the bursts of asymmetric, concentrated flows in particular areas that are characteristic of commuters.

An additional aggregate characteristic is the proportion of docking points that are filled by bicycles at the point of maximum availability. This is known as the maximum load factor or alternatively the maximum Normalised Available Bicycles, or NAB (Froehlich et al., 2009):

$$L_{\max} = B_{\max}/D = B_{\max}/(B_{\max} + S(B_{\max}))$$

$$\begin{aligned} L_{\max} &= \text{Maximum Load Factor}, \quad B_{\max} \\ &= \text{Maximum bicycles available}, D = \text{Docking points}, \quad S(B_{\max}) \\ &= \text{Spaces available at point where maximum bicycles available} \end{aligned}$$

As well as the maximum of the load factor being a useful statistic for a system, the temporal variations in the load factor are also a key metric that is discussed further below.

A final aggregate characteristic is one of popularity. This can be inferred by observing the peak utilisation of the bicycles and is calculated by looking at the maximum change, in a single day during September 2012, the period of study, of the number of bicycles available:

$$U_{\mathrm{dmax}} = (B_{\mathrm{dmax}} - B_{\mathrm{dmin}})/B_{\mathrm{dmax}}$$

$$\begin{aligned} U_{\mathrm{dmax}} &= \text{Maximum concurrent usage for day}, \quad B_{\mathrm{dmax}} \\ &= \text{Maximum bicycles available in day}, \quad B_{\mathrm{dmin}} \\ &= \text{Minimum bicycles available in day} \end{aligned}$$

The result represents a popular day's usage within each system, probably during good weather and with a minimum of docking stations disabled for maintenance. The day could have been a weekend day or weekday, depending on when a system's most popular usage occurs. It is important to note that it is the maximum concurrent usage is being measured, rather than the popularity directly. This is considered to be proportional to the total daily use of the system, in the absence of journey data being available. In systems where the usage is low or the journeys are short in duration, then the concurrent maximum measurement may underestimate the total daily use, while still providing a useful indication of activity.

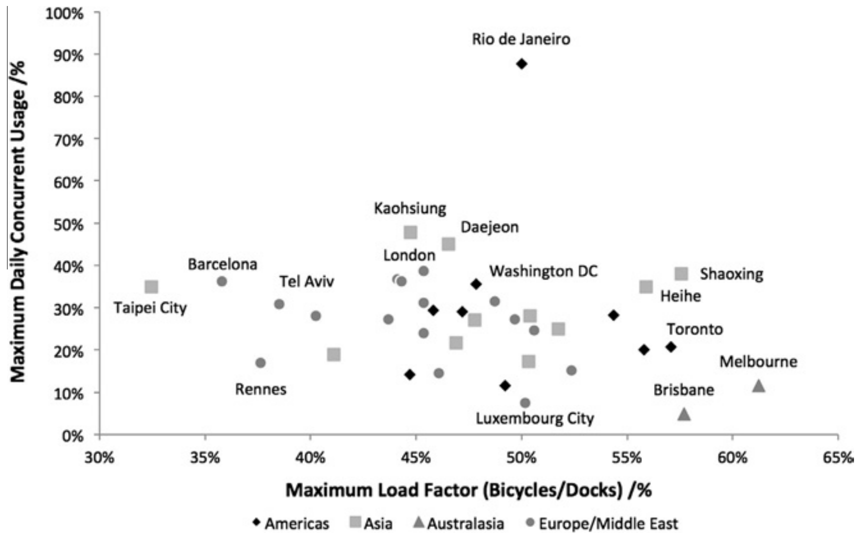

Fig. 2. Maximum daily concurrent usage and maximum load factor.

Less popular systems may use a higher maximum load factor so that bicycles are more readily available for use in most docking station locations without a danger of overloading the system, and to visually attract potential users. More popular systems will lower the load factor to allow the system to work efficiently during periods of high simultaneous usage – particularly systems which have a strong uni-directional flow at certain times, e.g. commutes into a city centre. Popular systems where usage is predominately tourist or utility based, so less likely to see uni-directional flow, can however work efficiently with a higher load factor due to the more random nature of the movement of the bicycles.

Time-based charges mean that non-availability of a docking station at the end of a journey is considered a much worse situation than non-availability of bicycles at the start, hence the need for operators to specify their maximum load factors carefully, taking into account predominant usage types and the degree of uni-directional flow that may happen.

A maximum load factor of just under 50% (one bicycle for two docking points) is typical for most bike-shares, with Europe's average of 45% being slightly lower, and America's average of 50% being slightly higher, than the worldwide average of 48%. This may reflect the different predominant cultural usage types of bicycles in the different regions. It is acknowledged that the differences in maximum load factors vary much more between systems within each region, than between the regions. This reflects that practical considerations within each city, such as the geographical structure of a city, population density and the system's size (both absolute and relative to the city population), equipment funding and redistribution resources, ultimately are more likely key determinants of each system's implemented maximum load factor.

The London system was specified with a docking point to bicycle ratio of 1.7 to 1, equating to an unusually high maximum load factor of 59% (Transport for London, 2009, p. 10) and launched with a load factor of 54%. It then gradually reduced to around 50% as the overall system size increased over the first year of operation – the total number of bicycles in the system being in fact reduced slightly as the same time as new docking stations were opened. During September 2012 it was at 45%, but then increasing gradually in bicycles available during the wintertime, viable because of lower usage, and a corresponding decrease in summertime.

Australasia's systems, relatively unpopular in this study due to local bylaws requiring helmet use, also have noticeably higher maximum load factors than the other regions. Aside from this, the implied inverse correlation between popularity and load factor is only weakly visible (see Fig. 2).

Taipei City's system stands out in both sets of comparative charts above, as having the largest average docking station size and the lowest maximum load factor of all the systems studied –

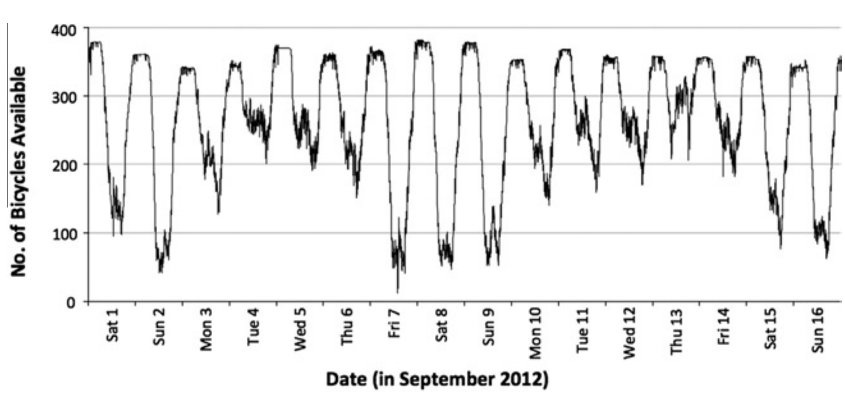

Fig. 3. A snapshot of temporal variation in bicycle availability for Rio de Janeiro's bicycle sharing system, across the first 16 days in September 2012. The 7th was a public holiday. A trough indicates a high number of bicycles in use, resulting in fewer being available for others to use. For example, the figure shows that during the middle of the second day, only around 50 of the approximately 380 bicycles are in the system are not being used at that time. The pattern of usage between weekdays and weekends is noticeably different.

this may be due to the compact and off-centre location of its central business district compared with the rest of the city, resulting in fewer, but larger docking stations, and a low load factor needed to manage large bursts of commuter usage to and from this area.

Rio de Janeiro's system is small but sees many days of extremely high usage – especially on weekends – where few bicycles remain available for hire by additional users for a significant portion of the day. The strong weekend bias is likely due to use of the system by tourists, as this bike-share is located beside several large beaches – a high tourist use is also suggested by the period of high usage extending throughout much of the day – tourists are likely to be less price-sensitive and so less likely to return their bicycle during the first hire time band (typically an hour) where the usage charge is minimal. Fig. 3 shows the typical change in availability of bicycles across the first part of September 2012, the larger troughs corresponding to weekend days or public holidays. Weekdays show a slight "double dip" of likely commuter usage but contained within an overarching single large dip, again likely due to high tourist use regardless of it being a weekend or weekday. These weekly, daily and intraday patterns continue throughout the year, with the city's warm year-round weather resulting in only small seasonal variations.

These variations in load factor, which provide some insight into the high maximum load factor aggregate characteristic for Rio, are studied in more detail below.

3.2. Spatial characteristics

For each system a 1 km buffer around each docking station is applied with the resulting shape approximating the area of influence of the system. One kilometre is chosen as a compromise between the maximum straight-line distance that someone would likely walk in order to reach a docking station (see Daniels and Mulley, 2011) and that which a user would likely cycle beyond the physical system boundary. We concede this is a generous criterion for many systems; for example in designing the London bicycle sharing system a guideline separation of 300 m was targeted (Mayor of London, 2010, p. 1). This is equivalent to specifying a buffer of 150 m, that if applied, creates a large number of "holes" that clearly are not present in the context of known usage characteristics. The resulting geographic extents of the 38 systems studied here vary considerably from 20 km^2 to 200 km^2, with a significant cluster around 50 km^2. In addition, nearest-neighbour analysis was also performed to measure the average distance between each docking station and its closest partner. Results from this indicate that approximately 50% of the systems have an average around 300–400 m.

It should be noted that Euclidean distances are used here rather than network distances. It is assumed that the two are directly proportional, that the latter distances for cycling journeys are not significantly longer, and that this proportion is approximately the same across the systems such that using Euclidean distances is a justifiable simplification. All the systems studied are based in urban areas, which typically have a high density of intersections, allowing journeys to be made in any direction by using road or bicycle lane infrastructure, with the distance travelled unlikely to be far beyond the straight-line distance. The physical effort exerted in cycling means that cyclists are much less predisposed to taking long detours than car drivers or other transport users (e.g. metro riders) who are more constrained within their network. Cities also generally act to maximise the permeability of movement for pedestrians and cyclists. A Euclidean distance simplification is less likely to be valid for rural areas, with a lower road and intersection density.

Based on the relationship between these two attributes, Paris and Bordeaux are distinct; Paris has the largest area but also maintains a high density of docking stations whilst Bordeaux's system has a compact core, but also many docking stations separated by several kilometres from their nearest neighbour in the suburbs. There is no significant variation of the areas or densities seen across world regions.

Finally, two dimensionless statistical measures of system shape and layout are obtained. The compactness ratio, also known as the circularity ratio, describes how circle-like each system's shape is. It compares the area of the polygon formed by the buffer creation around each docking station, with the area of a circle that has a circumference equal to the buffer polygon. A perfectly circular, or compact, system has a compactness ratio of 1. Systems with very low compactness ratios are more irregularly shaped.

$$C = 4\pi A/L^2$$

$$C = \text{Compactness Ratio}, \quad A = \text{buffer area}, \quad L = \text{circumference of the buffer}$$

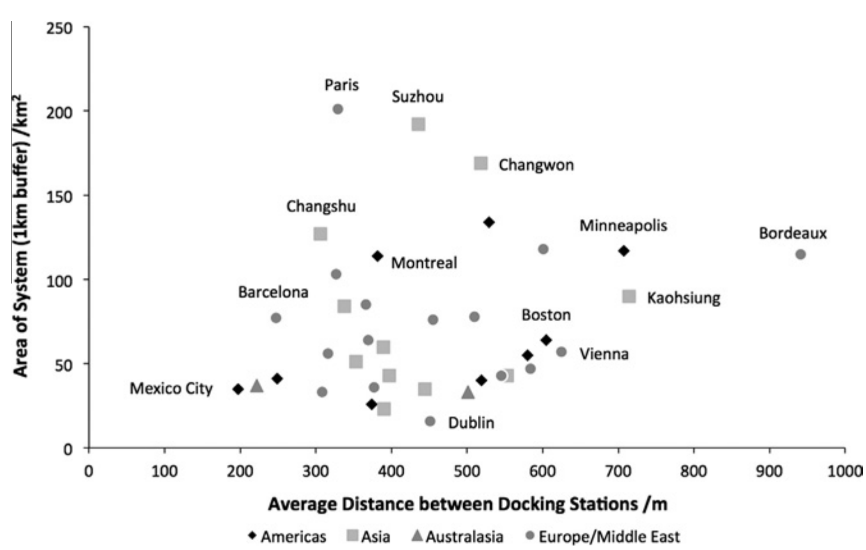

Fig. 4. System area and nearest-neighbour distances.

Please cite this article in press as: O'Brien, O., et al. Mining bicycle sharing data for generating insights into sustainable transport systems. J. Transp. Geogr. (2013), http://dx.doi.org/10.1016/j.jtrangeo.2013.06.007

Asian systems have generally lower compactness ratios (averaging 0.55) than European/Middle Eastern ones (averaging 0.72). This notable difference suggests larger, less compact settlements that need to meander in shape around local geographies, are more present in Asia, European systems being more likely to concentrate in the traditionally compact cores of their cities. It is also likely to reflect the tendency for Asian systems to cover separate, distinct communities within one interoperating system in a large city, serving journeys within sub-systems each covering a local community while using the same technology across the city for simplicity. As such, the perimeter of the system is more complex. A side effect of the fixed 1 km buffer also means that systems which have the nearest neighbour measurement approaching the same distance – notably Bordeaux – will result in a low compactness ratio as a significant proportion of docking stations have a buffer that does not overlap with those adjacent to it.

The second dimensionless measure is the *Z*-score, also known as the standard score. It is beyond the scope of this paper to detail this, but it can be simply interpreted as the "orderliness" of how the docking stations appear across the system's extent, with respect to each other. It does not look at the calculated area, but instead looks at the distribution of distances between all the points (docking stations in this case) normalised by the distance between each point and the central mean point of the system. It calculates whether this distribution is random or exhibits signs of clustering or dispersal.

Clustered systems predominately have groups of docking stations in one or more smaller areas within the larger system – perhaps in middle business districts of a city, or distinct neighbourhood populated by a demographic that the system designers expect would use the bicycles extensively. They have a large negative *Z*-score. Dispersed systems have been deliberately designed to be spaced out within their urban areas – allowing a near-even coverage and confidence to a user that, as long as they remain in the system area, they are not normally more than a short distance from the nearest docking station. Such systems have a large positive *Z*-score. Systems exhibiting a near random distribution of docking stations have a *Z*-score of around zero. These systems will have concentrations of docking stations in key areas – major transport hubs and city centres – while also having a sparse distribution in outlying neighbourhoods and other areas of low population density and therefore likely ridership.

The *Z*-score statistic is being used here merely to quantitatively describe the type of layout of a system. Since docking station locations are designated by urban planners to serve a particular residential area, high-density office complex, public transport hub or tourist attraction, and are normally limited by planning requirements, population distributions, demographic considerations or physical constraints of geography, each city is likely to have a natural *Z*-score statistic that would be unlikely to change without a significant redeployment of docking stations. A substantial change would be seen if a city decided to fill in the gaps in their system to increase the evenness of coverage, or deploy additional docking stations to remote neighbourhoods.

There are a wide variety of *Z*-scores across the 38 systems studied, as shown in Fig. 5, with European/Middle Eastern systems generally having a higher *Z*-score (average −0.6) than those in Asia (−6.3) – suggesting a more regularly spaced set of docking stations across the system. The *Z*-score correlates somewhat with the compactness ratio – again this is most likely because of larger systems serving several multiple, relatively disconnected communities in larger Asian cities, compared with more uniform cross-city European systems. Nice's unusually low Z-score (for Europe) is due to the area's steep topography and the system having a long drawn-out strand of docking stations along the coast and up several adjacent valleys, in effect serving communities separate from the system's main core of docking stations in Nice itself.

In the system footprints shown in Fig. 6, Changwon's complex shape and wide variations in the distance between adjacent docking stations – a high degree of spatial clustering – means it has a low compactness ratio (0.3) and very negative *Z*-score (−16). In contrast, Valencia's grid-like structure – highly dispersed – and approximately circular shape results in a high compactness ratio (0.8) and a positive *Z*-score (6.7). Barcelona has approximately the same compactness ratio as Valencia's, but a significantly lower *Z*-score (−8) because it clusters more strongly along some major roads, within the main system area. There is a weak correlation between the two measures, as shown in Fig. 5, as smaller systems in particular are likely to be concentrated only in a city centre, which is typically of uniformly high building density and free of geographical obstructions, allowing a circular and even structure – simple to design for and straightforward to use – to be easily constructed.

3.3. Temporal characteristics

To extract the temporal signatures of each bike-share, the total number of bicycles available at each moment are summed and

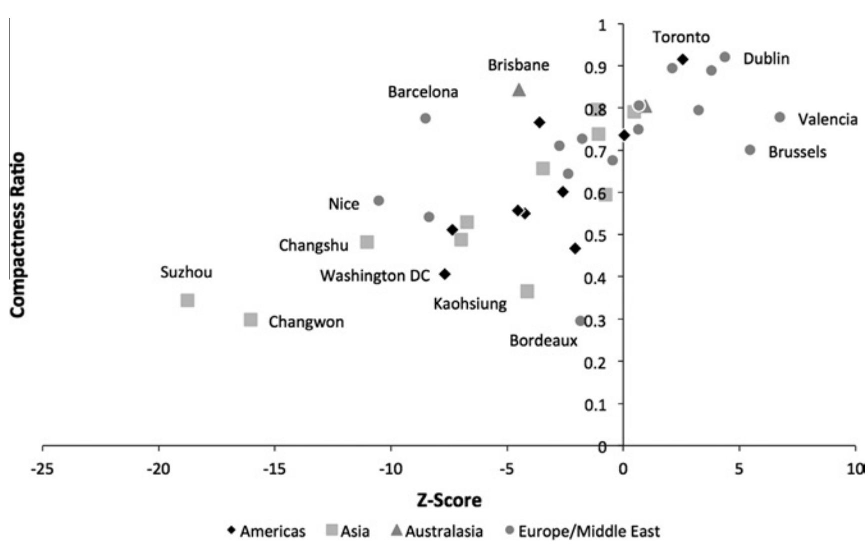

Fig. 5. Shape and density characteristics: compactness ratio and *Z*-score.

Please cite this article in press as: O'Brien, O., et al. Mining bicycle sharing data for generating insights into sustainable transport systems. J. Transp. Geogr. (2013), http://dx.doi.org/10.1016/j.jtrangeo.2013.06.007

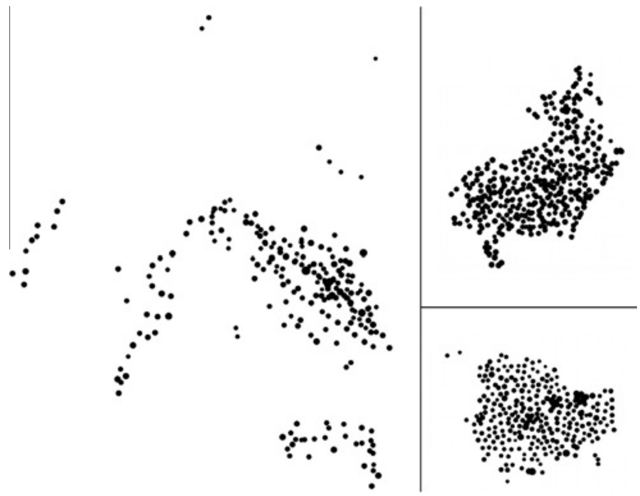

Fig. 6. Spatial footprints of systems in Changwon (left), Barcelona (top right) and Valencia (bottom right) shown at the same scale. Each circle represents a docking station. The area of each circle is directly proportional to the number of docking points it contains.

divided by the total number of docking points available, to obtain the load factor. A value of 0 at a particular point in time means there are no bicycles available to hire, while a value of 100% means that every docking point in each docking station is full of bicycles. In practice, as discussed earlier, most systems operate under an approximate load factor of just under 50% (i.e. just over two docking points for each bicycle). The measured value drops below this during periods of heavy use, returning to the "steady state" value overnight. Load factor will rarely fall below 30%, as this would probably result in large parts of a system's area being without bicycles available for hire, restricting additional use.

Redistribution and maintenance activities carried out by system operators can result in spurious activity being seen, particularly overnight. It is difficult to remove such activity from the analysis. We do not, however, consider this as significant when viewing the macroscopic characteristics of each system, particularly when the data are aggregated over longer periods of time. In commuter-driven systems, such as London and Barcelona, the very unequal distribution of bicycles across the system following each commuting time-period is observed to by-and-large remain until the next commute time. On transitioning into weekend use, the less ordered movement of weekend users often acts as a remarkably effective redistribution from a Friday-evening post-commuter distribution in itself.

Many bike-shares are based in city centres and show characteristic usage peaks (corresponding to temporary troughs in the load factor) during morning and evening commutes, and a single peak during weekend afternoons, such as is seen in London in Fig. 7.

A peak in hiring activity leads to a drop in the measured load factor across the bicycle sharing system, so such bursts of activity appear as troughs in the graphs shown in Fig. 8.

Fig. 8 highlights obvious differences between systems, but it is also notable that many systems show common traits, despite substantial differences in geographical footprint, system size and density. In particular, a double-trough weekday usage and a wide single-trough weekend usage is a characteristic shared in a substantial number of the systems we have studied here. Ultimately, the ubiquitous "9–5" working day, five days a week, plays out across most of the systems studied. The high capital costs associated with setting up a bicycle sharing system also ensures that most system designers will likely wish to emulate existing successful systems in terms of the user demographics and habits.

3.4. System classification based on temporal characteristics

Based on observations of the 30 systems shown in Fig. 8, and of the other 8 systems studied, we have developed a classification, shown in Table 2, that summarises the temporal characteristics. It incorporates the number of the peaks per day for weekdays and weekends, the relative difference between weekend and weekday usage, and average load factor. We have also stated the likely demographic characteristics of the cyclists themselves. The proposed demographic types are:

- Commuters: Use bicycles to travel between home/transport hub and office.
- Utility Users: Use bicycles throughout the weekday for shopping, errands.
- Leisure Users: Generally cycle at weekends for fun and exercise.
- Tourist Users: Use bicycles to get to beach or explore city.

3.5. Description of systems based on their comparative characteristics

On observing the relative positions of the individual systems on the graphs showing the non-spatial and spatial attributes, and plotting time-series data on systems as we have done in Fig. 8, the commonalities and differences between the respective systems become apparent.

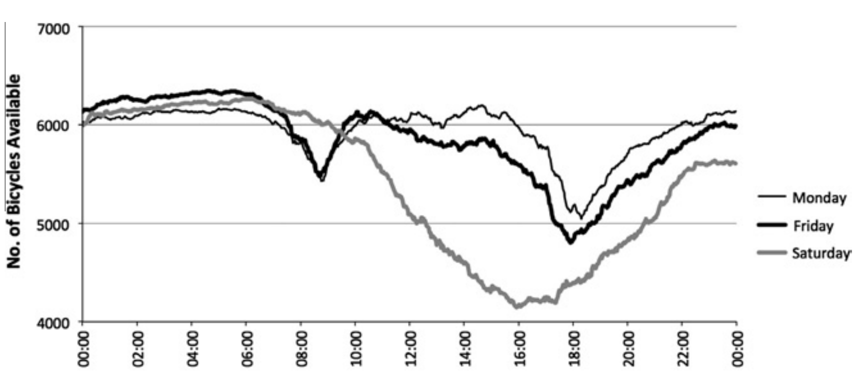

Fig. 7. Load factor variation across three specimen days (midnight to midnight) in May 2012 for London. The troughs correspond to peaks in activity, resulting in a drop of available bicycles. Monday's data shows a typical weekday. Friday was a weekday during very good weather, the trend sees the normal "double dip" weekday use but also a strong influence of the weekend pattern – high additional use in the afternoon.

Please cite this article in press as: O'Brien, O., et al. Mining bicycle sharing data for generating insights into sustainable transport systems. J. Transp. Geogr. (2013), http://dx.doi.org/10.1016/j.jtrangeo.2013.06.007

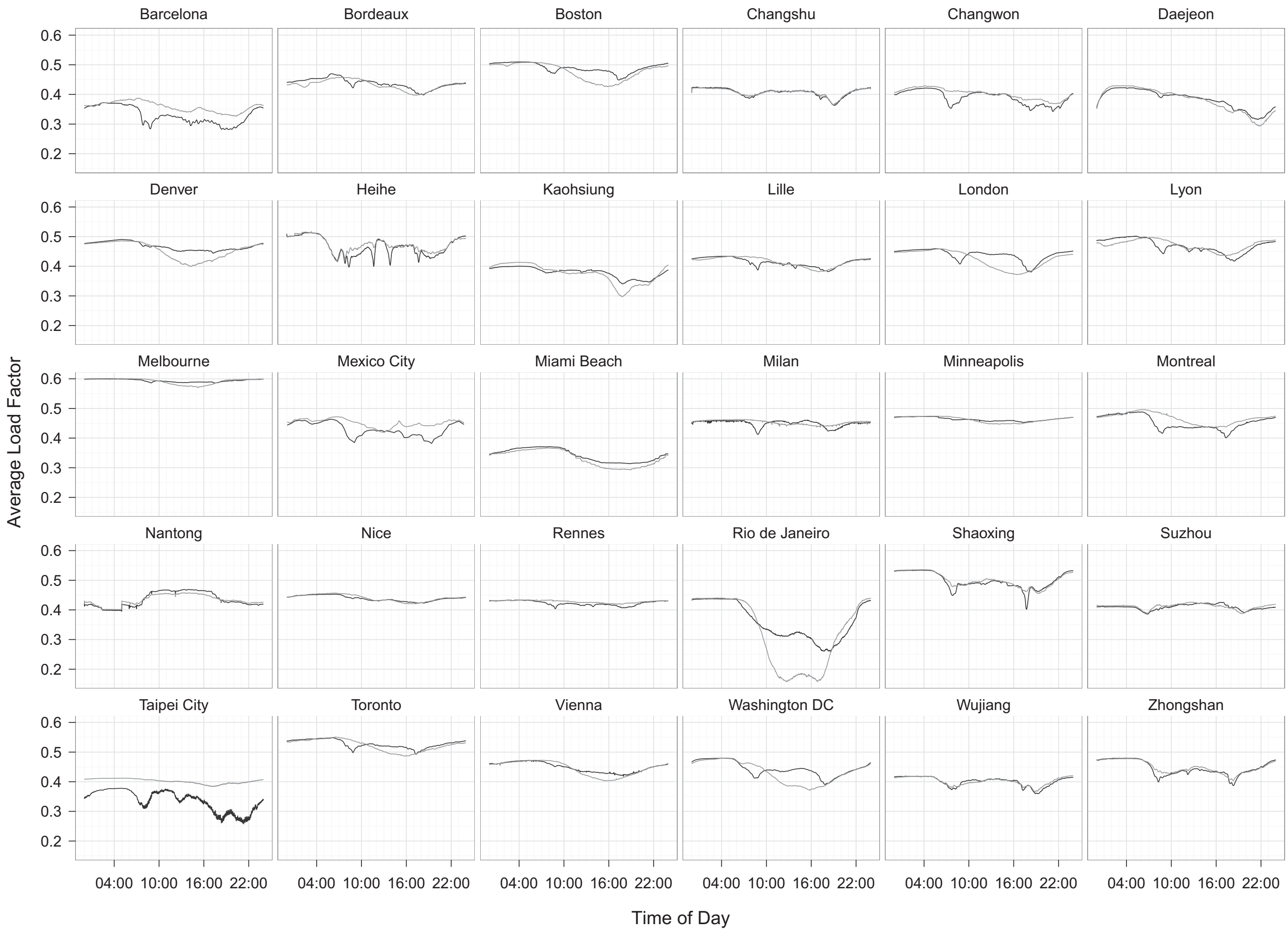

Fig. 8. Load factor variation across 24 h, for weekdays (black lines) and weekend days (grey lines), the data being averaged across April–September 2012, for 30 of the 38 cities in this study. The other eight systems' data are not shown to assuage concerns the operator may have regarding display of such data at a fine temporal resolution. For two systems here (Melbourne and Rio de Janeiro) this was across winter. Some data cleaning was employed to remove fluctuations caused by unplanned outages and data collection errors. Denver's system is closed from midnight to 6 a.m. Taipei City's data is from September 2012 only and shows some noise due to the smaller number of days it is averaged across. The time of day is local to each system.

London, Washington DC, Toronto, Boston and Montreal all see a similar pattern, with two weekday commuter peaks and a broad afternoon peak at weekends. Washington DC's weekend peak is larger relative to its weekday use, than for the other cities, likely reflecting the USA's traditional leisure and touristic orientated focus for cycling. The spatial distribution of bicycles after an evening commute period in London is shown in Fig. 9.

Changshu's system sees peaks earlier in the day than the other systems in China, reflecting that China's single time zone across a wide longitude means that local timings within the country vary significantly. Heihe's system is a small Chinese system. The lunchtime patterns suggest the bike-share users use the system to go home for lunch, and then use it again to return to work. Many of China's systems also show a weekend usage that is very similar in shape to weekday usage, often including the same commuter peaks.

Bordeaux, Toronto and Vienna see only a small but consistent weekday use, with minor but noticeable commuter peaks. However their weekend afternoon use is significantly higher. Barcelona and Saragossa show quite different patterns to the other cities. Reflecting on the different Spanish/Catalan working life style, there are three commuter peaks, the evening one being much later than for other countries. Weekend usage also sees a lull, as users avoid the afternoon sun.

Barcelona's first commuter usage peak has a "double dip" – occurring just before 8 a.m. and just before 9 a.m. A similar pattern exists in Valencia and Saragossa, the other Spanish systems studied. In many Spanish urban areas, public sector workers start work at 8 a.m. and private sector workers start at 9 a.m.

Milan shows a curious increase in load factor after the morning commuter rush, compared with the overnight values. This suggests that bicycles are added to the system by the operator, during the working day, and then removed again before the rush hour.

Melbourne, Minneapolis and Rennes see only a small usage – the first two having slightly greater weekend use and the latter slightly greater use during weekdays. Melbourne's result may be lower as, being the only southern-hemisphere system studied, it was wintertime during the period of study. Of note, Melbourne's load factor "at rest" (i.e. outside of normal usage hours) is much higher than for all of the other systems. The low usage allows it to have many bicycles in each docking station, presumably for visibility and publicity purposes, without system users suffering from finding their destination station full. The aggregate characteristic measure also shows this.

3.6. Bike-share system hierarchy

The final part of this paper summarises the aforementioned characteristics of each system through an inductive classification

Table 2
Simple qualitative classification of systems based on temporal characteristics.

Dominant pattern	Predicted demographic	System
Two commuter peaks during weekdays, one peak at weekend	Commuters and weekend leisure users	Bordeaux Boston Changwon London Mexico City Milan Montreal Paris Rennes Tel Aviv Toronto Washington DC
Seven-day commuter peaks	Commuters	Changshu Daejeon Kaohsiung Nantong Nice Shaoxing Suzhou Wujiang Zhongshan
More than two commuter peaks on weekdays	Commuters with some utility users	Dublin Heihe Lille Lyon Taipei City
Commuter peaks and high intra-peak usage	Utility users with some commuters	Barcelona Luxembourg City Saragossa Valencia
Mainly weekend use. Often high load factor.	Leisure users	Brisbane Brussels Denver Melbourne Minneapolis Vienna
Single peak on all days, high usage throughout the day	Tourist users	Miami Beach Rio de Janeiro

procedure. A range of attributes, listed in Fig. 10, were assigned to the 38 schemes and input into a hierarchical clustering procedure developed by Ward (1963). Ward's hierarchical clustering treats each system as unique before merging it with the most similar system (based on the input attributes) and then merging the resulting pairing with the most similar pairing and so on. This process enables a tree structure (dendrogram) to emerge with adjoining branches forming between similar groups. The resulting dendrogram is shown in Fig. 10. It offers a powerful summary of the aforementioned analysis and confirmation of the similarities/differences between different scheme characteristics. For example, Chinese systems group together, along with Lille (notable for its dispersed docking station geography). Spanish systems also group together, along with Lyon. There are some surprising partners – Montreal and Changwon, and Miami Beach and Daejeon. The cities may feel very different on the ground, but a significant number of the measures applied to the clustering – such as system size, compactness and weekday/weekend popularity comparison – may be similar, resulting in such a grouping together.

4. Applications

4.1. Demographic and community detection of the data

It is clear from the docking station data that a number of distinctive patterns are present, and can be used to form hypotheses both about the characteristics of the system users in each city, and the city itself. The interval between usage peaks during weekdays can reveal the typical working hours of a city, for example, while the position and size of the weekend usage peak provides an insight into the weekend habits of the city's dwellers. These insights however have a limitation in that the demographic concerned is that of bicycle-sharing system users, not of the city at large. There is some evidence that such users are more likely to be male and live in less-deprived areas than the general population (Ogilvie and Goodman, 2012) so such measures need to be controlled for when describing a city's behaviour as a whole. A potential practical observation that can be made is that measured fluctuations in average apparent speed might indicate episodes of road congestion or increased system users that take meandering paths.

The research can easily be extended to further understand the demographics of the system users – for example, studying "social" hires, where two or more bicycles are hired from a particular place at a close point in time, and then re-docked elsewhere at similar times later, and places where leisure cycling is most desirable, either by studying variations of journey speeds (derivable from calculating likely routes for journey-level data) in particular areas of the city, or proportions of hires that finish at the same docking station they started from. Other research has shown that clustering of docking stations can reveal communities of users, the classification showing significant spatial similarity despite

Please cite this article in press as: O'Brien, O., et al. Mining bicycle sharing data for generating insights into sustainable transport systems. J. Transp. Geogr. (2013), http://dx.doi.org/10.1016/j.jtrangeo.2013.06.007

Fig. 9. Central London, an example of a city with a commuter-dominated bike-share system following the evening rush hour. The darkened docking stations are full or nearly full, while white ones with black borders are completely empty. Large areas in the centre are completely devoid of bicycles, surrounded by a "doughnut" of largely full docking stations. The background mapping contains OpenStreetMap mapping data.

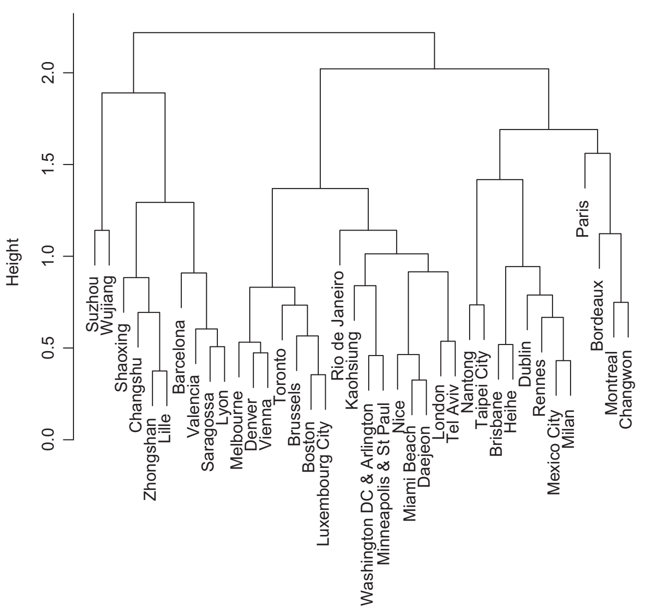

Fig. 10. A dendrogram of the systems studied, grouped using Ward's hierarchical clustering. The following input system attributes were used: number of docking stations; average docking station size; maximum simultaneous usage during September 2012; maximum load factor; system size; system area; mean nearest neighbour distance; *Z*-Score; compactness ratio; number of usage peaks each weekday; number of weekend daily usage peaks; large weekend usage indicator.

location normally not being an input into the clustering process, suggesting different user types and intentions at different times in different districts.

4.2. The redistribution problem

The operator has a particular interest in learning the daily and weekly cycles of system activity, and the effect of external events such as weather and transport strikes on movements, because effective redistribution (the act of moving bicycles from where they are located to where they are needed) is important for many of the systems – particularly if they are dominated by asymmetric flows at certain times of the day (e.g. commuters) but with other user types requiring them at other times (e.g. tourists) or they are simply too small or too popular for their city to be able to satisfy demand.

Operators can use the datasets to build up a profile of users and journeys, and monitor trend changes in the numbers of a particular user type, that may need to be addressed by changes in the redistribution strategy or frequency. While one metric used is the time a docking station remains full or empty, it may actually be desirable to retain or even affect such a state, if it is likely that a rapid change in the numbers of bicycles (caused by the users themselves rather than operator redistribution activity) can be expected shortly.

In addition, the classifications offered above provide a useful basis for operators to anticipate usage patterns for planned systems and learn from systems similar to their own. It is our belief that this would enable more effective budgeting for future systems if they can better anticipate the likely redistribution efforts required and inform changes to existing systems, such as extensions and pricing strategies. Such insights stand to increase the success of both existing and future systems and therefore offer a pragmatic contribution to sustainable urban transport systems.

5. Conclusion

This paper has sought to demonstrate the insights offered by the straightforward collection of bicycle sharing system data. This work stands to benefit operators, researchers of urban behaviours and patterns, and users themselves. The usage of such systems is quite well ordered, with a relatively high degree of predictability.

It is readily apparent that a number of insightful observations can be made from a simple analysis of the docking station data from bicycle sharing systems, including many that are not discussed in this paper, and that such observations can be used to form comparisons between urban areas, and between their corresponding systems. It is clear also that there are numerous further opportunities for research with such datasets.

Please cite this article in press as: O'Brien, O., et al. Mining bicycle sharing data for generating insights into sustainable transport systems. J. Transp. Geogr. (2013), http://dx.doi.org/10.1016/j.jtrangeo.2013.06.007

Bicycle sharing systems play an important part in increasing sustainable transport options in cities. An understanding of their potential use, and impact, across many diverse types of cities and multiple user types, is becoming increasingly important.

With the creation of new systems and increased public availability of individual level origin–destination data for some systems (such as London, Washington DC, Minneapolis and Boston), the opportunities and applications of studying spatial, temporal and journey data associated with bike-sharing will continue to expand.

Acknowledgements

The authors would like to thank the anonymous reviewers for their helpful comments and suggestions, which have enhanced this paper. They would also like to thank "Eskerda" of CityBikes for capturing and standardising many of the data feeds through an API and Russell Meddin and Paul DeMaio of the Bike Sharing Blog for discovering and investigating new bicycle sharing systems around the world.

References

Daniels, R., Mulley, C., 2011. Explaining walking distance to public transport: the dominance of public transport supply. World Symposium on Transport and Land Use Research 2011. <http://sydney.edu.au/business/__data/assets/pdf_file/0013/106501/Daniels-Mulley-Explaining.pdf> (accessed 23.03.13).

DeMaio, P., 2009. Bike-sharing: history, impacts, models of provision, and future. Journal of Public Transportation 12(4), 41–56. <http://nctr.usf.edu/jpt/pdf/JPT12-4DeMaio.pdf> (accessed 23.03.13).

DeMaio, P., Meddin, R., 2012. The Bike-sharing World Map. <http://bit.ly/K9pKmO> (accessed 23.03.13).

Froehlich, J., Neumann, J., Oliver, N., 2009. Sensing and predicting the pulse of the city through shared bicycling. In: 21st International Joint Conference on Artificial Intelligence (IJCAI). <http://www.nuriaoliver.com/bicing/IJCAI09_Bicing.pdf> (accessed 23.03.13).

Haverman, R., 2010. Three Generations Bike Sharing. <http://www.bikesharing.eu/global-solutions/three-generations> (accessed 23.03.13).

Jensen, P., Rouquier, J.B., Ovtracht, N., Robardet, C., 2010. Characterizing the speed and paths of shared bicycles in Lyon. Transportation Research Part D: Transport and Environment 15(8), 522–524. <http://arxiv.org/abs/1011.6266> (accessed 23.03.13).

Jinran, Z., Xiaodong, W., 2012. Tough times in the Bicycle Rental Business. <http://www.chinadaily.com.cn/cndy/2012-09/24/content_15776527.htm> (accessed 23.03.13).

JZTI and Bonnette Consulting. 2010. Philadelphia Bikeshare Concept Study. <http://www.metrobike.net/index.php?s=file_download&id=27> (accessed 23.03.13).

Kaltenbrunner, A., Meza, R., Grivolla, J., Codina, J., Banchs, R., 2010. Urban cycles and mobility patterns: Exploring and predicting trends in a bicycle-based public transport system. Pervasive and Mobile Computing 6 (4), 455–466.

Kazis, N., 2012. With a Boost From Bike-Share, Cycling Surges on Mexico City's Mean Streets. <http://www.streetsblog.org/2012/03/22/with-a-boost-from-bike-share-cycling-surges-on-mexico-citys-mean-streets/> (accessed 23.03.13).

Lathia, N., Ahmed, S., Capra, L., 2012. Measuring the impact of opening the London shared bicycle scheme to casual users. Transportation Research Part C: Emerging Technologies 22, 88–102. <http://www.cl.cam.ac.uk/~nkl25/publications/papers/lathia_transport_research_c.pdf> (accessed 23.03.13).

Mayor of London. 2010. Twenty-third Mayor's Report to the Assembly. <http://www.london.gov.uk/sites/default/files/mr-mqt15sep10.pdf> (accessed 23.03.13).

OBIS, 2011. Optimising Bike Sharing in European Cities. OBIS Handbook. <http://www.eltis.org/docs/tools/Obis_Handbook.pdf> (accessed 23.03.13).

Ogilvie, F., Goodman, A., 2012. Inequalities in usage of a public bicycle sharing scheme: socio-demographic predictors of uptake and usage of the London (UK) cycle hire scheme. Preventative Medicine 55(1), 40–45. <http://dx.doi.org/10.1016/j.ypmed.2012.05.002> (accessed 23.03.13).

Padgham, M., 2012. Human Movement is Both Diffusive and Directed. PLoS ONE 7(5), e37754. <http://dx.doi.org/10.1371/journal.pone.0037754> (accessed 23.03.13).

Transport for London. 2009. Cycle Hire Scheme: Document 2 ST-PJ302C Cycle Hire Scheme – Business Case Submission. Version V02b. <http://northofthethames.files.wordpress.com/2010/09/cycle-hire-scheme-bc.pdf> (accessed 23.03.13).

Tso, N., 2009. Taiwan Goes Green with Bike Sharing. <http://www.time.com/time/world/article/0,8599,1898152,00.html> (accessed 23.03.13).

Vogel, P., Greiser, T., Mattfeld, D.C., 2011. Understanding Bike-Sharing Systems using Data Mining: Exploring Activity Patterns. Procedia – Social and Behavioral Sciences 20, 514–523. <http://dx.doi.org/10.1016/j.sbspro.2011.08.058> (accessed 23.03.13).

Ward, J., 1963. Hierarchical Grouping to optimize and objective function. Journal of American Statistical Association 58 (301), 236–244.

Please cite this article in press as: O'Brien, O., et al. Mining bicycle sharing data for generating insights into sustainable transport systems. J. Transp. Geogr. (2013), http://dx.doi.org/10.1016/j.jtrangeo.2013.06.007