Sensing and Predicting the Pulse of the City through Shared Bicycling

International Workshop on Spatio-temporal Data Mining for a Better Understanding of Human Mobility: The Bicycle Sharing System Case Study
December 5th, Paris, France
A quick story re: traveling to this workshop.
I never take Vélib’. The redistribution is completely broken.

— Girl On Train
Parisian Exchange Student from England
bike share research stakeholders & benefits

1. **Operators**: benefit from more accurate models of demand for load balancing

2. **End-users**: benefit from understanding and forecasting how system will be used for trip planning

3. **Urban planners**: can use bike models to improve the bikeability of the city

4. **Social scientists & human geographers**: can study and better understand human mobility and routine
The social sciences can finally have access to masses of data that are of the same order of magnitude of their older sisters, the natural sciences.

Bruno Latour, 2007
French philosopher and sociologist
data
my research
this talk
my research
bike share data

Station Capacity Data

Data collection method
(1) Data dump from operator
(2) Scrape bikeshare website
bike share data

granularity

Station Capacity Data

Data collection method
(1) Data dump from operator
(2) Scrape bikeshare website
bikeshare data

granularity

Station Capacity Data

Data collection method
(1) Data dump from operator
(2) Scrape bikeshare website
(3) 3rd-party DIY APIs
bike share data

granularity

Station Capacity Data

Data collection method
(1) Data dump from operator
(2) Scrape bikeshare website
(3) 3rd-party DIY APIs

O/D Bike Data

Data collection method
(1) Data dump from operator
(2) ...

Data collection method

Capital Bikeshare releases anonymous trip data
by David Alpert • January 11, 2012 3:32 pm

Programmers or analysts interested in studying Capital Bikeshare patterns or creating useful apps can now do a lot more. Capital Bikeshare has followed through on its promise and posted data files with individual (but anonymous) trip data.

The files, one for each quarter going back to late 2010, list individual trips, including the time each started and ended, duration, which station it started and ended at, and an identifying number for the individual bike. It doesn't say anything about the member who used the bike, except whether they are a "registered" (annual or monthly) member or a "casual" member (daily or 3- or 5-day).

Now, people can generate tables or graphics showing the most popular station pairs, or where people most often go from an individual station, or what weather patterns make usage heavier or lighter, or where the nighttime activity is, and much more.

This data has been available for some time for London, allowing people to create animations of a day's CaBi usage and diagrams of a single bike's path over several days. The folks who built those and other tools can now even adapt their code to work for Capital Bikeshare, if they're so inclined.
Trip History Data

Overview of the Trip History Data

When a rental occurs within the system our software collects basic data about the trip. That data can be exported from our system and used for various types of analysis or research. By making this data available we hope to stimulate that analysis or research amongst a much wider community. Please note, all private data including member names has been removed from these files.

What is included?

Each .csv file contains data for one quarter of the year. Within each file there are 7 columns.
Some Quick **Fun** Facts

Analysis Enabled by Anonymized O/D Bike Data

(1) “Average” trip is downhill (by -1.94 meters)

(2) Last mile usage: four most common trips are short and seem to cover areas that subway/bus do not

(3) Sixth most common trip is a return trip from Smithsonian station back to the Smithsonian station

(4) Casual vs. members usage fees: 40.7% casual riders incur fees vs. 3.3% of members incur fees

MV Jantzen, CaBi Trips, http://youtu.be/-h0rV7tw1Eo
Using **heuristics** to infer bike flow.
At 2010-10-04 06:01:00 there were 3 bikes in use.
The routing is done using OpenStreetMap data & Routing routing scripts optimized for bike usage (i.e., constant speed on all road types, obeying one-way roads and taking advantage of marked cycleway). I’ve tweaked the desirability of road types, so that the trunk and primary roads are only slightly less desirable than quieter routes.

— Oliver O’Brien
UCL CASL

http://oliverobrien.co.uk/2011/02/boris-bikes-flow-video-now-with-better-curves
bike share data

granularity

Station Capacity Data

- Data collection method
  1. Data dump from operator
  2. Scrape bikeshare website
  3. 3rd-party DIY APIs

O/D Bike Data

- Data collection method
  1. Data dump from operator
  2. ...

O/D User Data

- Data collection method
  1. Data dump from operator
  2. Scrape user account logs
"Working collaboratively with Transport for London (TfL), customer records reporting a unique customer identifier, gender and postcode, have been made available. So too has a complete set of user journeys [for those customer ids]"
Rentals (72)

Capital Bikeshare is currently undergoing a system update which affects your rental history statistics from being calculated accurately. During this update, your distance, calories burned and CO2 lbs saved statistics will not be displayed. We are working to finalize the update and will provide correct statistics for your rental history again upon its completion. Your other rental history data has not been affected and is accurately displayed on this page. Thanks for your patience!

<table>
<thead>
<tr>
<th>Start Station</th>
<th>Start Date</th>
<th>End Station</th>
<th>End Date</th>
<th>Duration</th>
<th>Cost</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lamont &amp; Mt Pleasant NW</td>
<td>10-27-2012 10:40 am</td>
<td>New York Ave &amp; 15th St NW</td>
<td>10-27-2012 11:22 am</td>
<td>42 minutes, 35 seconds</td>
<td>$1.50</td>
</tr>
<tr>
<td>Connecticut Ave &amp; Newark St NW/Cleveland Park</td>
<td>10-23-2012 9:09 pm</td>
<td>Lamont &amp; Mt Pleasant NW</td>
<td>10-23-2012 9:20 pm</td>
<td>10 minutes, 49 seconds</td>
<td>$0.00</td>
</tr>
<tr>
<td>Adams Mill &amp; Columbia Rd NW</td>
<td>10-21-2012 10:51 am</td>
<td>Adams Mill &amp; Columbia Rd NW</td>
<td>10-21-2012 10:51 am</td>
<td>9 seconds</td>
<td>$0.00</td>
</tr>
<tr>
<td>20th &amp; Crystal Dr</td>
<td>10-20-2012 11:25 am</td>
<td>Braddock Rd Metro</td>
<td>10-20-2012 11:55 am</td>
<td>29 minutes, 30 seconds</td>
<td>$0.00</td>
</tr>
<tr>
<td>Lamont &amp; Mt Pleasant NW</td>
<td>06-03-2014 10:51 am</td>
<td>Lamont &amp; Mt Pleasant NW</td>
<td>June 3, 2014</td>
<td>32 minutes</td>
<td>$0.00</td>
</tr>
</tbody>
</table>
Human route repetition in car driving.
Route Prediction from Trip Observations

14,468 trips / 240 subjects

<table>
<thead>
<tr>
<th>Description</th>
<th>Average</th>
<th>Median</th>
</tr>
</thead>
<tbody>
<tr>
<td>trip distance (miles)</td>
<td>7.7</td>
<td>4.2</td>
</tr>
<tr>
<td>trip time (min)</td>
<td>16.3</td>
<td>11.5</td>
</tr>
<tr>
<td>num trips / day</td>
<td>4</td>
<td>3.9</td>
</tr>
<tr>
<td>num trips / subject</td>
<td>60.3</td>
<td>50</td>
</tr>
<tr>
<td>num days of data / subject</td>
<td>15.1</td>
<td>13</td>
</tr>
</tbody>
</table>

High Level Trip Stats

[Greater Seattle Area

[From Froehlich & Krumm, Route Prediction From Trip Observations, SAE 2008]
After approximately 1 month of observation, the number of repeat trips reaches 50%.
Avg Cumulative Distribution of Trips in Routes

Percentage of Trips

Route (Ordered By Most Frequently Traveled)
The top ten most frequently traveled routes account for 50% of a driver’s trips.

The most frequently traveled route account for 12% of a driver’s trips.
**bike share data**

**Event Based**
Station I/O with cloud when bike docked/hired

- Station Capacity Data
- O/D Bike Data
- O/D User Data

**Continuous**
Use sensors to track movement

- Instrumented Cities (e.g., CCTVs)
- Instrumented Bicycles (e.g., GPS sensor)
- Instrumented Users (e.g., mobile phones)
bike share data

**Event Based**
Station I/O with cloud when bike docked/hired

**Continuous**
Use sensors to track movement

- Station Capacity Data
- O/D Bike Data
- O/D User Data
- Instrumented Cities (e.g., CCTVs)
- Instrumented Bicycles (e.g., GPS sensor)
- Instrumented Users (e.g., mobile phones)
sensing and predicting the movement of a city via shared bicycling

[Froehlich et al., UrbanSense2008; IJCAI2009]
Summer 2008:
- 373 stations
- 6,000 bicycles
- 150,000 subscribers

bicing
barcelona, spain
Mapa de estaciones

A continuación se pueden visualizar en el siguiente plano las estaciones de bicicletas actualmente en funcionamiento. Así como, ver en tiempo real las disponibilidades de bicicletas en cada una de ellas.

Distrito:
Dirección:
Código postal:

Estaciones Vacías:
Estaciones Llenas:

Gran de Gràcia
Bicicletas 5
Aparcamientos 22

Estaciones con mas de una bici
Estaciones sin bici
Estaciones cercanas

RESULTADOS
Total estaciones activas: 375
Ciutat vella
Exemple
Gràcia
Pobla de Farners
Pobla Sec
Sants-Montjuïc
Sants-Sant Gervasi
Vila Olímpica
Data Collection

• Scrape bicing webpage every 2 mins

• Extract:
  – station’s geo-location
  – # of available bicycles
  – # of vacant parking slots
before cleansing

station 15: num available bicycles

0  10  20
# of avail bicycles

fri  sat  sun  mon  tues  wed  thurs  fri

time

random zero value

erroneous oscillations

after cleansing

station 15: num available bicycles

0  10  20
# of avail bicycles

fri  sat  sun  mon  tues  wed  thurs  fri
How do we know what to clean?

- available bikes for all station

# of avail bicycles

- random zero value
- and here as well

6 hours
13 weeks of observations
Aug 27 – Dec 1, 2008

<table>
<thead>
<tr>
<th></th>
<th>raw dataset</th>
<th>cleaned dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>stations</td>
<td>390</td>
<td>370</td>
</tr>
<tr>
<td>days</td>
<td>25K</td>
<td>22.7K</td>
</tr>
<tr>
<td>observations</td>
<td>26.1M</td>
<td>20.2M</td>
</tr>
<tr>
<td>parking slots</td>
<td>9831</td>
<td>9315</td>
</tr>
</tbody>
</table>
Num checked-out bicycles across all stations

- morning commute
- late Spanish lunch
- evening commute
- sleeping in on weekends
Two-spike pattern found in study of London Underground

[Lathia, Froehlich & Capra, ICDM2010]
Introducing DayViews
morning commute starts @ 7am
lunch rush begins @ 1pm
return from lunch/work at 3:30pm
10am beach goers begin arriving
morning commute starts @ 7am

lunch rush begins @ 1pm

return from lunch/work at 3:30pm

Activity Score:

\[ AS(t) = |B_t - B_{t-1}| \]
How are Bicing patterns shared across stations and distributed in the city?
Temporal Clustering

- Activity clusters
- Available bicycle clusters

Applied dendrogram clustering with dynamic time warping as distance metric.
Cluster A1 (N=207)
Cluster A2 (N=76)
Cluster A3 (N=74)
Cluster A4 (N=12)
Temporal Clustering

activity clusters

available bicycle clusters

Activity

Weekday
Available Bicycle Cluster

**flat**
- Cluster B1 (N=106)
- Cluster B2 (N=72)

**outgoing**
- Cluster B3 (N=42)
- Cluster B4 (N=62)

**incoming**
- Cluster B5 (N=50)
- Cluster B6 (N=38)
Can Bicing station usage be predicted?
Why Care?

• load balancing
• assist urban planners / city officials about expected activity
• provide new web/mobile services to bicing users
76% of respondents had difficulty finding a bicycle.
66% of respondents had difficulty finding a parking slot
50% of respondents avoid Bicing when they are traveling to a place where they must be on time.
Station Models

$B_{t_0} = 27 \quad t_0$

$PW = 10 \text{min}$
$B_{t_0} = 27$  
$t_0$  
$\text{Pred}_{LV} = 27$  
$\text{Pred}_{LV} = 27$  
$\text{Actual Value} = 10$  
$\text{Actual Value} = 27$  
$\text{PW} = 10 \text{min}$  
$\text{PW} = 2 \text{hr}$  
$\text{Last Value}$
$B_{t_0} = 27$

$\text{Last Value} = 10$

$\text{Pred}_{LV} = (t_0, B_{t_0}, PW) = B_{t_0}$

$PW = 2\text{hr}$
Historic Mean

$\text{Pred}_{\text{HM}}(t_0, B_{t_0}, PW) = B_{TB_{t_0}} + PW$

Station's DayView
Historic Mean

$$Pred_{HM} = (t_0, B_{t_0}, PW) = B_{TB_{t_0} + PW}$$

$B_{t_0} = 27$

$t_0$

Actual Value = 27, $Pred_{HM} = 23$

Actual Value = 10

Station's DayView

$PW = 10 \text{min}$

$PW = 2 \text{hr}$
**Historic Mean**

\[ \text{Pred}_{HM} = (t_0, B_{t_0}, PW) = B_{TB_{t_0} + PW} \]

- \( B_{t_0} = 27 \)
- \( t_0 \)
- \( PW = 2\text{hr} \)
- \( \text{Actual Value} = 10 \)
- \( \text{Pred}_{HM} = 14 \)

**Station's DayView**

- # of available bicycles
\[ \text{Pred}_{HT}(t_0, B_{t0}, PW) = B_{t0} + B_{TB_{t0}} + PW - B_{TB_{t0}} \]

Station's DayView

Actual Value=10

Actual Value=27

Pred_{HT}=27+(25-24)=28

Pred_{HT}=27+(15-24)=18

\( B_{t0} = 27 \)

\( t_0 \)

PW=10min

PW=2hr
Historic Trend

Station's DayView

$t_0 = B_{t0} = 27$

$Pred_{HT} = (t_0, B_{t0}, PW) = B_{t0} + B_{TB_{t0}} + PW - B_{TB_{t0}}$

$Actual Value = 10$

# of available bicycles

PW = 2hr
\[ \text{Pred}_{BN}(t, B_{t0}, PW) = B_{t0} + \delta \]

Bayesian Network

- **time**: discrete observed node corresponding to hours in the day
- **bikes**: the number of available bikes at time \( t \)
- **PW**: the prediction window
- **\( \delta \)**: continuous Gaussian variance that represents change in the number of bikes at time \( t + PW \)

Prediction made by adding the value of the \( \delta \) node to the most recent observation.
3 weeks of data to build models

1 week of test data

models were fed:
- the current time
- current # of avail bicycles
- each of the six pw values (10, 20, 30, 60, 90, 120 mins)
Prediction Error Metric

- Absolute difference between the predicted number of bicycles & the ground truth observation at time $t_0 + PW$
- Error is in number of bicycles – normalized by the station’s size
# High Level Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Avg Error</th>
<th>Stdev of Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>0.37</td>
<td>0.27</td>
</tr>
</tbody>
</table>

*error is in normalized available bicycles (nab)

0.37 corresponds to roughly 9 bicycles
## High Level Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Avg Error</th>
<th>Stdev of Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>0.37</td>
<td>0.27</td>
</tr>
<tr>
<td>HistoricMean</td>
<td>0.17</td>
<td>0.16</td>
</tr>
<tr>
<td>LastValue</td>
<td>0.09</td>
<td>0.14</td>
</tr>
<tr>
<td>HistoricTrend</td>
<td>0.09</td>
<td>0.13</td>
</tr>
<tr>
<td>Bayesian Network</td>
<td><strong>0.08</strong></td>
<td><strong>0.12</strong></td>
</tr>
</tbody>
</table>

*error is in normalized available bicycles (nab)*

0.08 corresponds to roughly 2 bicycles.
Prediction vs. Activity Cluster

0.1 corresponds to ~2.5 bicycles at a station with 25 slots
Topographical Influences?

- Station elevation vs. average percentage of available bicycles
- Station elevation vs. activity score
weather
other transit sources
cross city examination

BIKESHARE project

Wednesday, December 5, 2012
02:26:00 EDT

Denver
Number of stations: 0
Updated 2 minutes ago.

Brisbane
Number of stations: 153
Updated 1 minute ago.

Melbourne
Number of stations: 50
Updated 25 seconds ago.

Minneapolis
Number of stations: 145
Updated 2 minutes ago.
self-sustainable system
promote usage

at current pace:
expected arrival: 11 mins
current: 4 empty slots
predicted: 6 (11 mins)
Longitudinal Study
Related Collaborators

Neal Lathia
Nuria Oliver
Joachim Neumann
John Krumm
Licia Capra
Chris Smith

Related Publications

**Individuals Among Commuters: Building Personalised Transport Information Services From Fare Collection Systems**

**Mining Public Transport Usage for Personalised Intelligent Transport Systems**
Neal Lathia, Jon Froehlich, Licia Capra, *Proceedings of ICDM2010*

**Sensing and Predicting the Pulse of the City Through Shared Bicycling**
Jon Froehlich, Joachim Neumann, Nuria Oliver, *Proceedings of IJCAI2009*

**Measuring the Pulse of the City Through Shared Bicycle Programs**
Jon Froehlich, Joachim Neumann, Nuria Oliver, *Proceedings of UrbanSense2008*

**Route Prediction From Trip Observations**
Jon Froehlich, John Krumm, *Proceedings of SAE2008*

Sensing and Predicting the Pulse of the City through Shared Bicycling

International Workshop on Spatio-temporal Data Mining for a Better Understanding of Human Mobility: The Bicycle Sharing System Case Study
Co-organized by GERI Animatic, le Labex Futurs Urbains et l'Ecole des Ponts ParisTech
December 5th, Paris, France