Sensing and Predicting the Pulse of the City through Shared Bicycling

Jon Froehlich (U. of Washington)
Joachim Neumann (Telefonica)
Nuria Oliver (Telefonica)
contributions

1. introduce potential of shared bicycling as a data source
2. highlight spatiotemporal patterns and their relation to the city
3. prediction of usage patterns and an analysis of influential factors
understanding human mobility patterns

bicing in barcelona, spain
bicing in barcelona, spain

launched march 2007

by summer 2008:
- 373 stations
- 6,000 bicycles
- 150,000 subscribers
data collection

• scrape bicing webpage every 2 minutes

• extract
  — station’s geo-location
  — # of available bicycles
  — # of vacant parking slots
example data

station 15: num available bicycles

# of avail bicycles

0 10 20

random zero value

erroneous oscillations

time

after cleansing

station 15: num available bicycles

# of avail bicycles

0 10 20

fri sat sun mon tues wed thurs fri
data cleansing example

available bikes for all station

# of avail bicycles

0 5 10 15 20 25 30 35

1150 1160 1170 1180 1190 1200 1210

random zero value

and here as well

6 hours
# Dataset

<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>

13 weeks of observations  
(Aug 27 – Dec 1, 2008)
introducing DayViews

Mon  Tues  Wed  Thur  Fri

sept 1-5

sept 8-12

sept 15-19

Station 37 DayView

percentage full

time of day (hrs)
morning commute starts @ 7am
lunch rush begins @ 1pm
return from lunch/work at 3:30pm
10am beach goers begin arriving
morning commute starts @ 7am

Activity Score: 
$AS(t) = |B_t - B_{t-1}|$

lunch rush begins @ 1pm

return from lunch/work at 3:30pm
num checked-out bicycles across all stations

evening commute
late spanish lunch
morning commute

sleeping in on weekends
how are bicing patterns shared across stations and distributed in the city?
applied dendrogram clustering with dynamic time warping as distance metric
activity clusters

Cluster A1 (N=207)
Cluster A2 (N=76)
Cluster A3 (N=74)
Cluster A4 (N=12)
activity clusters

Cluster A1 (N=207)
Cluster A2 (N=76)
Cluster A3 (N=74)
Cluster A4 (N=12)
Cluster A5 (N=1)
avail. bicycle clusters

**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 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.
downtown station
(morning)

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

$B_{t_0} = 27$, $t_0$

$PW = 10 \text{min}$

# of available bicycles

time
Last value $P_{DLV}=(t_0, B_{t0}, PW)=B_{t0}$

$B_{t0} = 27$

$P_{DLV} = 27$, $P_{DLV} = 27$

Actual Value = 27

Actual Value = 10

$PW = 10\text{min}$

$PW = 2\text{hr}$
Historic mean \( \text{Pred}_{HM} = (t_0, B_{t0}, PW) = \bar{B}_{TB_{t0}} + PW \)
Historic mean: \( \text{Pred}_{HM} = (t_0, B_{t0}, PW) = B_{TB_{t0} + PW} \)

- \( B_{t0} = 27 \)
- \( t_0 \)  
- \( \text{Actual Value} = 27 \)
- \( \text{Pred}_{HM} = 23 \)

- \( PW = 10 \text{min} \)
- \( PW = 2 \text{hr} \)

- \( \text{Actual Value} = 10 \)

**Station's DayView**
**historic trend**

$$\text{Pred}_{HT}(t_0, B_{t0}, PW) = B_{t0} + B_{TB_{t0}} + PW - B_{TB_{t0}}$$

- **$B_{t0} = 27$**
- **$t_0$**
- **$PW = 10\text{min}$**
- **$PW = 2\text{hr}$**

**Station’s DayView**

- **Actual Value = 27**
- **$\text{Pred}_{HT} = 27 + (25 - 24) = 28$**
- **$\text{Pred}_{HT} = 27 + (15 - 24) = 18$**
- **Actual Value = 10**
Bayesian network $\text{Pred}_{BN}(t_0, B_{t_0}, PW) = B_{t_0} + \text{delta}$

- **time**: discrete observed node corresponding to hours in the day
- **bikes**: the # of avail bikes at time $t$
- **PW**: the prediction window
- **delta**: continuous Gaussian var that represents change in number of bikes at time $t + PW$

Prediction made by adding the value of the *delta* node to the most recent observation
prediction evaluation

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

3 weeks of data to build models

1 week of test data
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

Error is in normalized available bicycles (nab)

<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>

0.37 corresponds to roughly 9 bicycles.
prediction error vs. pw

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

PW > 60, HT and BN significantly better
prediction vs. activity cluster

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

- A5 (N=1)
- A4 (N=12)
- A3 (N=74)
- A2 (N=76)
- A1 (N=207)
prediction vs. bicycle cluster

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

prediction error (nab)


# of avail bicycles

time of day (hrs)
biases of human behavior
other transit sources
self-sustainable system
promote usage

at current pace:
expected arrival: 11 mins
current: 4 empty slots
predicted: 6 (11 mins)
thank you!

Jon Froehlich
jfroehli@cs.washington.edu

Joachim Neumann
joachim@tid.es

Nuria Oliver
nuriaoa@tid.es
understanding human mobility
patterns

barcelona cellular network

Message Count for Monday, 2008-03-03

12AM  7AM  4PM  12AM

Time
00:00:00 02:00:00 04:00:00 06:00:00 08:00:00 10:00:00 12:00:00 14:00:00 16:00:00 18:00:00 20:00:00 22:00:00 00:00:00

Message Count
0 100 200 300 400 500 600 700 800

00:00:00 02:00:00 04:00:00 06:00:00 08:00:00 10:00:00 12:00:00 14:00:00 16:00:00 18:00:00 20:00:00 22:00:00 00:00:00
streetline: sfpark
from Fabien Giradin's ICING 2008 talk entitled Sentient to Responsive Cities
cellular network data

Madonna Concert
Cellphone activity in Stadio Olimpico Rome
2006-08-06

At Rome’s Olympic Stadium
Located about three kilometres from the Vatican
During the song Live to Tell...
Madonna appeared against a mirrored cross
digital footprinting: uncovering tourists with user-generated content