Applications of bus location data

Traffic Data Research Team
School of Information Sciences
University of Tampere, Finland
Contact: jyrki.nummenmaa@uta.fi

In collaboration with Tampere University of Technology, Finland, and Sichuan University, Chengdu, China
University of Tampere started computer science in 1965 (first CS professor in Scandinavia).
Tampere – Background info

- Tampere is a third largest city in Finland and Tampere region is the largest outside of capital area with a population a little over 500,000 (small in Chinese scale)
- There are about 150 buses in traffic during daytime
- Buses have GPS sensors. Locations are sent to background system
- Background system shares bus locations once a second through internet.
- We have collected and stored this data for analysis for nearly 2 years.
Tampere Bus Location Data

- **Real-time** data stream from Tampere public transportation bus fleet
  - > 100 vehicles
- in **SIRI format**
- Updates **every second**
- Includes e.g.
  - GPS location
  - Line number, direction and departure
  - Delay
Data Quality Problems

- Service breaks with no data (not very often)
- Connection to bus lost or some other technical problem
  - Shows the same position
  - Last time the position was recorded is shown
- Buses that are not in any busline are included
- GPS accuracy
Open data

• The bus location data is open data, provided by the City of Tampere.
• However, we need to store the data to have the history data.
• The data, cleaned up by us, has been used e.g. in a traffic data visualization competition.
Travel time analysis

- Travel times of the route 16 during the daytime hours
- 42-60 minutes, 46 stops
- Travel time in minutes per stops
  Direction 1
  Direction 2
Bus delays

- Delay is a difference between timetable and arrival time
- Delay is calculated and included in the data (every second)
- For example, see below delay on Route 16 at bus stops during daytime hours
- If a bus starts late it will be late at the end
- Delay increases in the city center and some intersections

Direction 1

Direction 2
Bus delay analysis

- From **stored bus location data**, analyze the traffic fluency
  - From all the observations with delay > 5min, use **frequent itemset mining** to find the lines, locations and times of most regular delays
- **Compare** a large set of delayed journeys to not delayed journeys to find out the bottlenecks along the bus routes
- Take the best and worst **quartiles and compare** to find out the bottlenecks
Step 1: Where, when and on which lines do the delays typically occur?

- **8 AM – 9 AM**: 
- **3 PM – 4 PM**: 
- **4 PM – 5 PM**: 
- **5 PM – 6 PM**:

The diagrams show the areas with delays at different times of the day.
Step 2: Time spent on fast vs slow journeys

Fastest 25% on mid-day quiet hours

Slowest 25% on afternoon peak hours

- Driving
- Traffic signals
- Bus stops
- Bus stops or traffic signals
Step 3: Identify the bus stops that most contribute to the journey time

Compare the variation of the time spent at the bus stops during the day.
Step 3: Identify the route segments between stops that have highest variation in journey time
Step 4: Action

The effect of improving public transportation traffic signal priorities on the main street.
On-going further work

- Mining areas and times with exceptional delays.
- Compare the delay values in grid cells (3-dimensional grid, z-axis is time) with city average delays.
- If we find a grid cell where delay values are exceptional, we try to extend the grid in different coordinate directions.
Friday
Scanning time 16:00~17:00
Timetables

- The City of Tampere bus timetables show when they hope that the bus comes to the bus stop.

- We have calculated *data-driven* information, which show the reliability of the arrival (www.pysakilla.fi pysakilla = at stop)

  green: ok, 0-4 mins wait  
yellow: quite ok, 0-8 mins wait,  
red: unreliable
Route planner

• The best thing since public transportation was invented?
• You want to get from place A to place B – the route planner calculates the necessary connection, tells you when to leave, where to change the bus, etc.
• ”What could go wrong”?
The data for risk estimates

- To estimate the risks involved in connections, we will use the data on how buses really arrive and depart.
- The location of each bus is available each second. Let’s first have a look at Bus line 18 (second bus) actual arrival times at our final destination.

<table>
<thead>
<tr>
<th>date</th>
<th>1505 actual time (scheduled: 15:58)</th>
<th>1515 actual time (scheduled: 16:08)</th>
<th>1515 delay seconds</th>
</tr>
</thead>
<tbody>
<tr>
<td>2013-11-21</td>
<td>16:00:20</td>
<td>16:21:35</td>
<td>13:35</td>
</tr>
<tr>
<td>2013-11-22</td>
<td>15:58:02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2013-11-26</td>
<td>16:05:17</td>
<td>16:14:09</td>
<td>6:09</td>
</tr>
<tr>
<td>2013-11-29</td>
<td>16:07:05</td>
<td>16:07:08</td>
<td>-52</td>
</tr>
<tr>
<td>2013-12-02</td>
<td>15:58:30</td>
<td>16:18:17</td>
<td>10:17</td>
</tr>
<tr>
<td>2013-12-04</td>
<td>16:06:06</td>
<td>16:14:34</td>
<td>6:34</td>
</tr>
<tr>
<td>2013-12-05</td>
<td>16:02:35</td>
<td>16:14:50</td>
<td>6:50</td>
</tr>
</tbody>
</table>
Chances to change the bus?

- Now, let’s combine arrivals of Line 30 and departures of Line 18.

<table>
<thead>
<tr>
<th>Date</th>
<th>Latest viable Line 30 bus</th>
<th>Latest viable connecting Line 18 bus</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DVJR</td>
<td>Scheduled time</td>
</tr>
<tr>
<td>20/11/2013</td>
<td>1500</td>
<td>15:28</td>
</tr>
<tr>
<td>21/11/2013</td>
<td>1500</td>
<td>15:28</td>
</tr>
<tr>
<td>26/11/2013</td>
<td>1500</td>
<td>15:28</td>
</tr>
<tr>
<td>28/11/2013</td>
<td>1510</td>
<td>15:38</td>
</tr>
<tr>
<td>30/11/2013</td>
<td>1500</td>
<td>15:28</td>
</tr>
<tr>
<td>03/12/2013</td>
<td>1510</td>
<td>15:38</td>
</tr>
<tr>
<td>04/12/2013</td>
<td>1510</td>
<td>15:38</td>
</tr>
<tr>
<td>05/12/2013</td>
<td>1500</td>
<td>15:28</td>
</tr>
</tbody>
</table>
Choices available for the traveler

• The low-risk option is to take the Line 30/DVJR 1450 bus, connecting to the Line 18/DVJR 1505 bus. On every day for which we have data, using this sequence of journeys would have resulted in the traveller arriving at Runkokatu by 16:15 time deadline.

• The higher-risk option is to take the Line 30/DVJR 1500 bus.
  • On 6 out of 12 days, it arrived at Koskipuisto sufficiently early to get the Line 18/DVJR 1505 bus which arrived at Runkokatu before the time deadline.
  • On a further three days, the Line 30 bus arrived at Koskipuisto sufficiently early for the traveller to catch the Line 18/DVJR 1505 bus and this latter bus arrived at Runkokatu by the time deadline.
  • On just two days for which we have data would this option have resulted in the traveller arriving at Runkokatu after the time deadline.
## Second example

### Route suggestions: Vilunen, Tampere - Runkokatu, Tampere

<table>
<thead>
<tr>
<th>Date</th>
<th>Line 22 Actual departure time from Vilunen</th>
<th>Line 22 scheduled arrival time at Koskipuisto</th>
<th>Line 22 actual arrival time at Koskipuisto</th>
<th>Line 18 actual departure time from Koskipuisto</th>
</tr>
</thead>
<tbody>
<tr>
<td>20/11/2013</td>
<td>15:20</td>
<td>15:36</td>
<td>15:34</td>
<td>15:41</td>
</tr>
<tr>
<td>21/11/2013</td>
<td>15:02</td>
<td>15:14</td>
<td>15:15</td>
<td>15:30</td>
</tr>
<tr>
<td>26/11/2013</td>
<td>15:20</td>
<td>15:36</td>
<td>15:34</td>
<td>15:36</td>
</tr>
<tr>
<td>27/11/2013</td>
<td>15:28</td>
<td>15:44</td>
<td>15:44</td>
<td>15:45</td>
</tr>
<tr>
<td>02/12/2013</td>
<td>14:56</td>
<td>15:14</td>
<td>15:10</td>
<td>15:29</td>
</tr>
<tr>
<td>03/12/2013</td>
<td>15:12</td>
<td>15:29</td>
<td>15:28</td>
<td>15:29</td>
</tr>
<tr>
<td>04/12/2013</td>
<td>15:21</td>
<td>15:36</td>
<td>15:35</td>
<td>15:41</td>
</tr>
</tbody>
</table>
Calculating our chances

• Based on our data (the table in the previous slide), we can now present the success rate of getting to Koskipuisto by particular times.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Arrival on time on this proportion of days</td>
<td>11/11</td>
<td>10/11</td>
<td>8/10</td>
<td>7/10</td>
<td>6/10</td>
<td>5/10</td>
<td>4/11</td>
<td>2/11</td>
<td>1/11</td>
<td>0/11</td>
</tr>
</tbody>
</table>

• Please note that this computation is data-intensive: for another time and connection, we need a different data set. In case of a large amount of users, we might like something a bit faster…
Predicting bus change success

• We can separate three different situations for estimating the bus change probability:
  1. The buses we are interested in are already in the move, and we can use their position and delay data in the estimate.
  2. The buses we are interested in are not yet in the move, but we want to use some other buses (e.g. same-day previous buses) in the estimate.
  3. We do not want to consider near-history data separately, but we want to make the estimate based on our collected history data.

We have some results on 2 and 3 and are working on 1.
Arrival distributions

- We want to find the arrivals time distribution for a bus on a given line arriving to a given stop.
- First guess: Normal distribution
- Generally not, if all departures of the day are considered together (rush hour and quiet time are too different).
- Best chance of normality is when we calculate the distribution for each (linenumber, direction, departure, busstop) combination -> worked well enough with us.

<table>
<thead>
<tr>
<th>Date</th>
<th>Line 22 scheduled arrival time at Koskipuisto</th>
<th>Line 22 actual arrival time at Koskipuisto</th>
</tr>
</thead>
<tbody>
<tr>
<td>20/11/2013</td>
<td>15:36</td>
<td>15:34</td>
</tr>
<tr>
<td>21/11/2013</td>
<td>15:14</td>
<td>15:15</td>
</tr>
<tr>
<td>22/11/2013</td>
<td>15:21</td>
<td>15:21</td>
</tr>
<tr>
<td>25/11/2013</td>
<td>15:36</td>
<td>15:37</td>
</tr>
<tr>
<td>26/11/2013</td>
<td>15:36</td>
<td>15:34</td>
</tr>
<tr>
<td>27/11/2013</td>
<td>15:44</td>
<td>15:44</td>
</tr>
<tr>
<td>28/11/2013</td>
<td>15:21</td>
<td>15:19</td>
</tr>
<tr>
<td>29/11/2013</td>
<td>15:36</td>
<td>15:35</td>
</tr>
<tr>
<td>02/12/2013</td>
<td>15:14</td>
<td>15:10</td>
</tr>
<tr>
<td>03/12/2013</td>
<td>15:29</td>
<td>15:28</td>
</tr>
<tr>
<td>04/12/2013</td>
<td>15:36</td>
<td>15:35</td>
</tr>
</tbody>
</table>
Compute parameters: map/reduce

Each night, for yesterday:

1. **Map**: by key (line, direction, origin, destination, departureTime, data) - additionally only arrivalTime and position are needed from the raw data (~800 Mb per day, we use 60 days, but every night just the last day’s data is calculated – previous data does not change.)

2. **Reduce**: Find the information of the bus stop, and for each stop, find for each bus the arrival and departure time (a bit of calculations with the coordinates are involved.

3. **Store** the resulting stopCode, arrivalTime, departureTime data (~5 Mb)

Now for the latest 60 weekdays (~300 Mb or arrival and departure data)

1. **Map**: by 
   (line,direction,origin,destination,originDepartureTime,stopCode)

2. **Reduce**: compute the distributions

3. **Results** are saved for on-line prediction (~10 Mb)
On-line service

• An open-source route planner is used to experiment
• The user chooses the origin and destination
• The route planner suggests routes
• Statistics is used to evaluate the routes
  • If the required arrivalTime and departureTime are both normally distributed, then so is their difference
• Using the arrival and departure distributions, we can calculate the probability that the change will happen (by default, 1 min is used as the time needed for the change).
• A route-planner company is already interested...
On-line prediction service
On-line prediction service
On-line prediction service
On-line prediction service
Using latest bus data

• Here, we assume that the buses of interest are not on the move yet.
• In that case, we can use the previous buses data (on the same line).
• We can do this with Bayesian estimation:
  • We have bus arrival/departure distributions.
  • We use these distributions as *a priori* distributions and compute a *posteriori* distribution using the latest observations.
• The rest can be done as before.
• Initial tests showed that with just $N=10$ latest observations we got reasonable results.
• Obvious problem is that we need to consult the raw data for this, in the previous approach just the pre-computed distribution parameters were enough.
Analyzing traffic fluency

• We consider the travel times of the buses between bus stops
  • In places where there is no bus lane, this reflects the overall travel fluency.
• This way, we can estimate and predict general travel times.
• We also want to detect exceptional circumstances.
“Prisma” junction, left turn, bus line No 3

80% of the traveling times fit between the blue and black lines.

Half of the traveling times are below and half above the red line.

*) Computed from a sample of 5744 observations on 76 working days / winter 2014-2015, with 15 minute time resolution.
Exceptional case

- Accident in the junction on October 8th 2014 at about 8AM jammed the traffic
- The traveling time in the junction raised to about 10-fold compared to a normal morning traffic
- (In addition, we can see that the traffic signal settings are probably different at the time when the model was built compared to October 8th ⇒ the model must be continuously updated)
Daily peak “Pispalan valtatie”

*) Computed from 28002 observations on 76 days in winter 2014-2015 with 15 minute time resolution

Green: Monday 2nd February 2015
But the peak does not appear every day!
On about 10-20% of the normal working days there is no peak (additionally, not on holidays)
E.g. on Thursday 29.1.2015 there is no peak
Could we, by using some other observations, predict which days the peak will appear?
All the ~2000 between-busstops-segments in Tampere can be automatically profiled using bus history data to get the "normal profile"

- E.g. with 30 min resolution, the model is a table of ~60000 rows and ~5-10 columns of numerical information
- Fits easily in main memory
- From the normal profile, we can find the interesting links that contain some regular peaks
- All the profiles can be used in real time to detect exceptional traffic situations
Ongoing work: More accurate predictions

- We consider the use of *shapelets* to analyze delays.
- Shapelets are an increasingly popular method for time series data mining (in our case the x-axis is the distance traveled).
- The idea is to identify approximate sub-shapes.
- Using shapelets we can more accurately study the detail patterns and predict traffic fluency.
- Intuitively, we can see that different locations give different delay patterns.
Shapelet mining

• Efficiency is a major concern.
• We have started by mining simple non-increasing or non-decreasing shapes (seems ok in this domain).
Combining bus data and traffic light data

- Data source: Tampere junctions with traffic lights, the data comes from the control systems at junctions.
- Data contents:
  - Signals A
  - Induction loops
  - SyncX detector
  - Bus priority detectors
  - Pedestrian crossing
  - Other Detectors
Map of Junction 309
Real time data

- For each junction: data update in every other hundreds milliseconds in real time.
- Raw data

```
1 CCB99811BB4 10100000000001000000000000110000000000 0 2014-08-04 00:00:04
2 CCB99811BB4 10100000000001000000000000110000000000 0 2014-08-04 00:00:04
3 CCB99811BB4 10100000000001000000000000110000000000 0 2014-08-04 00:00:04
4 CCB99811BB4 10100000000001000000000000110000000000 0 2014-08-04 00:00:04
5 CCB99811BB4 10100000000001000000000000110000000000 0 2014-08-04 00:00:04
6 CCB99811BB4 10100000000001000000000000110000000000 0 2014-08-04 00:00:04
```
### Example data

<table>
<thead>
<tr>
<th>#</th>
<th>Time</th>
<th>B</th>
<th>7</th>
<th>5</th>
<th>dt(dsl)7</th>
<th>dt(dsl)5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1000</td>
<td>2014-08-04</td>
<td>red</td>
<td>0</td>
<td>1</td>
<td>0(5)</td>
<td>1(125)</td>
</tr>
<tr>
<td>1001</td>
<td>2014-08-04</td>
<td>red</td>
<td>0</td>
<td>1</td>
<td>0(3)</td>
<td>1(125)</td>
</tr>
<tr>
<td>1002</td>
<td>2014-08-04</td>
<td>red</td>
<td>0</td>
<td>1</td>
<td>0(3)</td>
<td>1(125)</td>
</tr>
<tr>
<td>1003</td>
<td>2014-08-04</td>
<td>red</td>
<td>0</td>
<td>1</td>
<td>0(3)</td>
<td>1(125)</td>
</tr>
<tr>
<td>1004</td>
<td>2014-08-04</td>
<td>red</td>
<td>0</td>
<td>1</td>
<td>0(3)</td>
<td>1(125)</td>
</tr>
<tr>
<td>1005</td>
<td>2014-08-04</td>
<td>red</td>
<td>0</td>
<td>1</td>
<td>0(3)</td>
<td>1(125)</td>
</tr>
<tr>
<td>1006</td>
<td>2014-08-04</td>
<td>green</td>
<td>0</td>
<td>1</td>
<td>0(3)</td>
<td>1(125)</td>
</tr>
<tr>
<td>1007</td>
<td>2014-08-04</td>
<td>green</td>
<td>1</td>
<td>1</td>
<td>1(3)</td>
<td>1(125)</td>
</tr>
<tr>
<td>1008</td>
<td>2014-08-04</td>
<td>green</td>
<td>0</td>
<td>1</td>
<td>0(151)</td>
<td>1(125)</td>
</tr>
<tr>
<td>1009</td>
<td>2014-08-04</td>
<td>green</td>
<td>0</td>
<td>1</td>
<td>0(153)</td>
<td>1(141)</td>
</tr>
<tr>
<td>1010</td>
<td>2014-08-04</td>
<td>green</td>
<td>0</td>
<td>0</td>
<td>0(153)</td>
<td>0(141)</td>
</tr>
<tr>
<td>1011</td>
<td>2014-08-04</td>
<td>green</td>
<td>0</td>
<td>0</td>
<td>0(153)</td>
<td>0(141)</td>
</tr>
<tr>
<td>1012</td>
<td>2014-08-04</td>
<td>amber</td>
<td>0</td>
<td>0</td>
<td>0(153)</td>
<td>0(141)</td>
</tr>
<tr>
<td>1013</td>
<td>2014-08-04</td>
<td>amber</td>
<td>0</td>
<td>0</td>
<td>0(153)</td>
<td>0(141)</td>
</tr>
<tr>
<td>1014</td>
<td>2014-08-04</td>
<td>red</td>
<td>0</td>
<td>0</td>
<td>0(1)</td>
<td>0(139)</td>
</tr>
<tr>
<td>1015</td>
<td>2014-08-04</td>
<td>red</td>
<td>0</td>
<td>0</td>
<td>0(1)</td>
<td>0(139)</td>
</tr>
<tr>
<td>1016</td>
<td>2014-08-04</td>
<td>red</td>
<td>0</td>
<td>0</td>
<td>0(1)</td>
<td>0(0)</td>
</tr>
<tr>
<td>1017</td>
<td>2014-08-04</td>
<td>red</td>
<td>0</td>
<td>0</td>
<td>0(1)</td>
<td>0(0)</td>
</tr>
<tr>
<td>1018</td>
<td>2014-08-04</td>
<td>red</td>
<td>0</td>
<td>0</td>
<td>0(5)</td>
<td>0(0)</td>
</tr>
<tr>
<td>1019</td>
<td>2014-08-04</td>
<td>red</td>
<td>0</td>
<td>0</td>
<td>0(1)</td>
<td>0(0)</td>
</tr>
<tr>
<td>1020</td>
<td>2014-08-04</td>
<td>red</td>
<td>0</td>
<td>0</td>
<td>0(5)</td>
<td>0(0)</td>
</tr>
<tr>
<td>1021</td>
<td>2014-08-04</td>
<td>red</td>
<td>0</td>
<td>0</td>
<td>0(5)</td>
<td>0(0)</td>
</tr>
</tbody>
</table>
Observing detector occupancy %

Occupancy of detectors B0-1, B0-2 in junction309

- Amber
- Green
- Red

Time: Aug 04 00:00 to Aug 05 00:00

Variable:
- dt_1
- dt_5

SCHOOL OF INFORMATION SCIENCES

UNIVERSITY OF TAMPERE
What to use the traffic light data for?

• We can e.g. test the success of the "late bus priority policy"
  • If a bus is late (delayed), then the green light is extended for it, if it is close enough to the traffic lights.
• In the cases when extended green is given, we can check how many times the bus got through (success) and how many times it did not (failure).
Conclusions

• A lot of attention has been paid to taxi data.
  • Taxis may cover the city better, but buses have a schedule!
• Bus data can be used to analyze bus traffic, create more reliable services, increase the information to travelers, and to analyze traffic in general.
• From analysis to on-line prediction and services.
• More information from the buses (pictures, video, e.g) will improve the traffic analysis and services.
More information

- E. Betekhtina, J. Nummenmaa, P. Syrjärinne, "Prediction of Successful Bus Connection Based on Bayesian Analysis“, submitted for publication

- Web page: trafficdata.sis.uta.fi
- Email: jyrki.nummenmaa@uta.fi
Our team

University of Tampere:
• Jyrki Nummenmaa
• Paula Syrjärinne
• Peter Thanisch
• Timo Nummenmaa
• Elena Betekhtina
• Chenlu Wang
• Mykola Andruschenko

Tampere Univ. Of Technology:
• Ari Visa
• Riitta Kerminen
• Yibin Qian

Sichuan University:
• Lei Duan
• Yaowei Hu
• Xia Wu