Erath A.L., P. Fourie, L. Sun, S Ordóñez, A. Chakirov (2015) From Big Data to Smart Data: developing a MATSim model that runs on Smart Card Data, IVT, ETH Zurich, Zurich, 7th August 2015
From Big Data to Smart Data: developing a MATSim model that runs on SmartCard Data

Seminar
IVT
ETH Zurich
7th August 2015
The potential of data driven transport planning
1. Traffic flows result from the movement of single vehicles and individual people and their interactions.

2. Accounts for dynamics that arise from dwell processes, overcrowded vehicles and congestion induced delays on links.

3. MATSim allows for big simulation scenarios with millions of agents and hundreds of public transport services.
Turning Big Data into Smart Data

Data → Models → Simulation → Insight
MATSim – Multi Agent Transport SIMulation

- Open source software
- Under constant development
- Well documented at [www.matsim.org](http://www.matsim.org)
- Can simulate scenarios with millions of agents

Engages a learning cycle to find (stochastic) equilibrium
Data
Singapore CEPAS system – quick facts & stats

Closed system on train and bus – distance-based charging
Adult, student, senior citizen cards
Senior citizens: 25% discount; students: 50% discount; both pay flat fare after 7.2km (so sometimes don't bother tapping out)
360+ bus services, 3000 buses, 5 heavy rail lines (153km, 104 stations), 3 light rail lines in new towns
10+ different bus types, information on deployment from bus spotter website.

Key statistics for 8 April 2013
  4,138,780 journeys
  5,675,986 stages (61% bus)
  40,224,444 person-km (36% bus)
  105,260 stages with no tap-out
Using real demand to simulate public transport

**Travel demand directly derived from smart card transactions**

- Transactions recorded on Tuesday, 22\textsuperscript{nd} April 2011
- 4 Mio journeys, 5.7 Mio stages
- Adult, student, senior citizen cards
- Boarding stop (journey level)
- Boarding time
- Alighting stop (journey level)

**Travel supply**

- 362 bus lines
- 4 MRT lines
- 5 light rail services
- 1 Mono rail
- Each with detailed description of deployed vehicles
CEPAS smart card data (bus) to MATSim events

From this:

To this:
Extracting operational schedule from smart card data

Ridership vs transaction times, colored by stop ID

Removing GPS errors

Two plots of ridership vs transaction times, colored by stop ID (top) and transaction ‘speed’ (bottom).

Note how high ‘speed’ transactions (pink to red) correspond to incorrect stop IDs in the top plot.
Removing GPS errors

Ridership vs transaction times, colored by stop ID (top) and transaction ‘speed’ (bottom).

Note how high ‘speed’ transactions (pink to red) correspond to incorrect stop IDs in the top plot.
Interpolation of missing dwell events

Two plots of bus arrivals (bars above zero) and departures (below zero), colored by stop ID.

The top plot shows dwell events before interpolation, the bottom includes interpolated dwell events, with a zero duration.

The travel time allocated between interpolated dwell events is proportional to the free speed travel time between stops.
Interpolation of missing dwell events
## Error rates for 8th April 2013, bus only

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of dwell events identified</td>
<td>1,143,619</td>
</tr>
<tr>
<td>Number of dwell events interpolated</td>
<td>466,723</td>
</tr>
<tr>
<td>Number of card transaction</td>
<td>6,775,855</td>
</tr>
<tr>
<td>Stop ID not part of route</td>
<td>156,437</td>
</tr>
<tr>
<td>Transaction dropped due to speed</td>
<td>87,352</td>
</tr>
<tr>
<td>Dwell event transaction dropped</td>
<td>148,162</td>
</tr>
<tr>
<td>Share of transactions dropped</td>
<td>5.8%</td>
</tr>
</tbody>
</table>
Models
Boarding and alighting process

Boarding/alighting flow and instantaneous on-board passengers (single decker, low floor).

Passenger activity time model

Boarding and alighting process

Statistical model

\[ t = \{ \beta B + a(\max(on - cr,0)), aA \} \]
Bus types in Singapore

Standart type, single floor
Low floor or with step
1 boarding and alighting door each
83-85 pax capacity

Double decker
Low floor or with step
1 boarding and alighting door each
131 pax capacity

Articulated
With step
1 boarding and 2 alighting doors
132 pax capacity
Variability of activity time for different bus types

**Boarding**

![Graph showing boarding variability](image)

**Alighting**

![Graph showing alighting variability](image)

- Articulated bus with two doors
Variability of stop-to-stop travel time
Simplifying the network

- Road network
- Stop to stop link
- Bus lines
- Dwell link
Accounting for travel time variability

- Road network
- Stop to stop link
- Dwell link
- Travel time distributions by time of day
Modelling stop to stop travel times

Derive from Smart Card Data records travel times between stops

Each observed travel time between two subsequent stops constitutes one observation

Independent variables to be either derived from smart card data or GIS data, but do not require any other data source (e.g. traffic flow)

Static variables

- Availability of bus lane
- Number of intersections
- Number of left/right turns
- Curviness
- Deviation from crowfly distance
- Number of traffic lights
- Intersection density

Time-dependent variables

- Boarding/alighting activities in 500m radius
Modelling stop to stop travel times

\[ R^2 = 19.75\% , \text{ metrics are normalised to sum 100\%} \]
Residuals by time of the day
Simulation
Key assumptions

Fixed demand

Demand concentrated at the bus stop

Simulation of public transport only

Simplistic description of transfer characteristics

Routing according to parameters of stated preference survey
Validation

Bus speed

Transfer times

Trip duration (Bus)

Journey duration all modes

MATSim
EZ Link

MATSim
EZ Link

N = 1525445  Bandwidth = 0.5234

N = 1410694  Bandwidth = 60

N = 880230  Bandwidth = 60

Access, egress times removed from matsim bus times
Case studies

Evaluation of new services and routes:

- How can new network designs improve reliability and tackle overcrowding?
- How many passengers will be attracted by a new service?

Simulation and analysis:

- A full day simulated to steady-stated conditions in just about 40 minutes.
- Leverage on off-the-shelf business analytic software for interactive analysis (plus Senozon Via, of course)
The reliability of a long bus line

Time-space diagramm of bus line with 94 stops
Simulating and evaluating a line split
The effect of splitting the line
Reliability before and after line split

Before line split

After line split

Time of day

Stop number

Occupancy

0

128
Some data visualisation fancyness
Reliability: Excess waiting time along line

**Excess Waiting Time - East to West**

- **Original line**
- **After split**

![Graph showing excess waiting time along line with stop numbers and time in seconds.](Image)
Conclusion

From Big Data to Smart Data
- Use Big Data to understand the underlying operational patterns of public transport operations
- Agent-based simulation to model the inherent dynamics of public transport operations

A tool to evaluate alternative service provision
- How do different fare collection alternatives impact service quality?
- Which bus type for which line?
- How to adjust service provision in case of road works or congestion?
- How is the performance of alternative network designs?

Further research
- Inferring activity purposes and locations
- Integration of induced demand
The team to make it happen

Pieter Fourie
PhD student
Operations Research

Dr. Alex Erath
Deputy PI, project manager

Lijun Sun
PhD student
Data scientist

Sergio Ordonez
PhD student
Computer Scientist

Artem Chakirov
PhD student
Electric Engineer

Prof. Dr. Kay Axhausen
PI
Thank you!

www.futurecities.ethz.ch
www.ivt.baug.ethz.ch
www.matsim.org
@alex_erath
The (non-) issue with spatial autocorrelation
Other smart card data research at FCL

Understanding Metropolitan Collective Encounter Patterns (Lijun Sun et al.)
- The familiar stranger phenomenon on the bus
- Published in PNAS http://www.pnas.org/content/early/2013/07/31/1306440110

Efficient detection of contagious outbreaks in massive metropolitan encounter networks (Lijun Sun et al.)
- Using smart card data to develop efficient detection of disease spreaders
- Published in Science Reports http://www.nature.com/srep/2014/140606/srep05099/full/srep05099.html

Study of bus service reliability in Singapore using fare card data (Lijun Sun et al.)
    api_user_id=151611&download=XM4VQFC8

Models of Bus Boarding/Alighting Dynamics and Dwell Time Variability
- Sun, Lijun, Alejandro Tirachini, Kay W. Axhausen, Alexander Erath and Der-Horng Lee (2014). ‘Models of

Estimation of revealed preference route choice models to account for crowdedness (Tirachini, Sun, Erath)
- People travel in the wrong direction to secure a seat
- Paper to be submitted the coming days

Activity Identification and Primary Location Modelling based on Smart Card Payment Data for Public Transport
(Chakirov, Erath)
    api_user_id=151611&download=RQZWFXBZ

Stay tunes for more at http://www.futurecities.ethz.ch/module/mobility-and-transportation-planning/
Case study 2: adding a new bus line

Residential new town
- Tidal demand patterns
- Issues with overcrowding during peak hours

New bus line:
- 26 stops
- 10km
- Loop from MRT A to MRT B and back
Predicted vs actual ridership

Ridership per line and time of day

- Line 1
- Line 2
- New Line

Simulation [2011]

Reality [2013]
Crowding heatmap: Line 1

Time of day [h]

Direction

BEFORE

empty  full

AFTER

empty  full
Crowding heat map: Line 2

BEFORE

AFTER

Time of day [h]

Direction

empty   full

empty   full
Crowding heat map: New line

Direction

MRT A

MRT B

MRT A

empty  full