A data-driven approach to infer spatial characteristics and service reliability of public transport hubs

Ir. Menno Yap
Dr. ir. Niels van Oort
Dr. Oded Cats
Prof. dr. ir. Serge Hoogendoorn

M.D.Yap@TUDelft.nl
https://nielsvanoort.weblog.tudelft.nl/

May 22nd, 2017
Introduction (1)

• Public transport hubs have a central role in the network

• Public transport hub characteristics (analogy airports):
  o High connectivity (Pels, 2001)
  o Network centrality (Shaw 1993, Lohmann et al. 2009)
  o Limited number of hubs in network (Alderighi et al. 2005)
  o Concentration of different OD-passenger flows in time and space transferring via hub (Burghouwt, 2007)

• Hubs important in relation to passenger reliability
Introduction (2)

• Public transport reliability measures: from vehicle-based to passenger-based metrics
  o Punctuality
  o Regularity

• Passenger-oriented reliability measures: from trip to journey level; use of passive data sources
  o Additional passenger waiting time per line
  o Journey excess time

• Despite importance of hubs in affecting passenger reliability, no measures focusing specifically on hub reliability
Research goal

• Development of measures to quantify and compare hub reliability from a passenger perspective
  o Based on passive data sources
  o General applicable, independent of the case study network

• Research consists of three steps:
  o Infer spatial characteristics of potential hubs: which stops form a coherent cluster of transfer stops
  o Hub identification: which cluster of transfer stops concentrate substantial transfer flows in the network
  o Hub reliability: quantify and compare reliability of identified hubs

• Focus on urban public transport hubs only
Case study: network

- The Hague metropolitan area: ≈800,000 inhabitants
  - 2 light rail lines, 10 urban tram lines, 8 urban bus lines
  - 500 urban public transport stops (1650 Stop IDs), 8 train stations
  - ≈250,000 journeys per average working day (light rail + tram + bus)
  - 80% of these journeys by light rail / tram, 20% by bus
Case study: passive data sources

- Automated Fare Collection (AFC) data: entry-exit system

<table>
<thead>
<tr>
<th>Tap-in date + time</th>
<th>Tap-in stop-ID</th>
<th>Tap-in line</th>
<th>Tap-out date + time</th>
<th>Tap-out stop-ID</th>
<th>Trip-ID</th>
<th>Vehicle ID</th>
<th>Smart-card ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-3-2014 11:42:37</td>
<td>35309</td>
<td>6</td>
<td>4-3-2014 12:03:19</td>
<td>34997</td>
<td>3423</td>
<td>3050</td>
<td>81675688</td>
</tr>
<tr>
<td>4-3-2014 12:15:57</td>
<td>30091</td>
<td>18</td>
<td>4-3-2014 12:23:04</td>
<td>32857</td>
<td>6545</td>
<td>187</td>
<td>81675688</td>
</tr>
</tbody>
</table>

- Automated Vehicle Location (AVL) data

<table>
<thead>
<tr>
<th>Stop-ID</th>
<th>Trip-ID</th>
<th>Order-nr</th>
<th>Nominal arr</th>
<th>Realized arr</th>
<th>Nominal dep</th>
<th>Realized dep</th>
</tr>
</thead>
</table>

- Infer vehicle occupancy by integrating AFC+AVL data
- Stop data

<table>
<thead>
<tr>
<th>Stop-ID</th>
<th>RD x-coordinate</th>
<th>RD y-coordinate</th>
<th>Passenger stop name</th>
</tr>
</thead>
<tbody>
<tr>
<td>35309</td>
<td>81962</td>
<td>450867</td>
<td>Dr. H. Colijnlaan</td>
</tr>
<tr>
<td>30091</td>
<td>82188</td>
<td>455213</td>
<td>Central Station</td>
</tr>
</tbody>
</table>

- For this study: data used of 1 week (Nov 23 – Nov 27 2015)
Data processing: destination inf.

- Data cleaning (0.05 – 0.5% of daily transactions)
  - Delete system error transactions / unrealistic CoTime / missing trip ID

- Missing check-outs (1.4%): destination inference (Trépanier)
  - If \( m > 1 \) and \( j \neq m \), alighting location of \( j \) is closest to \( s_{p(j+1)k}^b \).
  - If \( m > 1 \) and \( j = m \), alighting location of \( j \) is closest to \( s_{p(j=1)k}^b \).
  - If \( m = 1 \), trip chaining is not possible: remove from dataset

- \( d_{walk} = \arg\max(\hat{s}_{pjk}^{a,c} - \hat{s}_{pjk}^{a,w}) \), \( d_{walk} \{d_{200}, d_{400} \ldots d_{1600}\} \): 400 Euclidean meter

![Validation of destination inference algorithm](image-url)
Data processing: transfer inference

- **State-of-the-practice:** \( t_{dp(j+1)k} \leq t_{apjk} + t_{t,max} \) (e.g. 35 min)
- **State-of-the-art:** alighting + boarding is transfer if:
  - \( l_{p(j+1)k} \neq l_{pj k} \rightarrow \text{what in case of short-turning, deadheading?} \)
  - If first vehicle run \( r_{lp(j+1)k} \) is taken after alighting \( \rightarrow \text{denied boarding?} \)
  - \( d(s^b_{p(j+1)k}, s^a_{pj k}) \leq d_{walk} \rightarrow \text{use intermediate PT on other network level?} \)
- **Improved transfer inference algorithm:** transfer if:
  - \( l_{p(j+1)k} \neq l_{pj k} \) or \( l_{p(j+1)k} = l_{pj k} \) if first run after alighting \( r_{lpj k} \) is taken
  - If first vehicle run \( r_{lp(j+1)k} \) is taken after alighting where \( q_{lr} < \text{capacity} \)
  - If first vehicle run is taken given intermediate level AVL data, \( d > d_{walk} \)
Spatial demarcation of potential hubs (1)

• Cluster transfer stops which form a coherent set of stops between which passenger transfer flows occur

• Clusters of transfer stops form potential hubs

• Determining clustering technique

<table>
<thead>
<tr>
<th>Technique Characteristics</th>
<th>K-means/ K-medoid</th>
<th>Hierarchical agglomerative clustering</th>
<th>DBSCAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-defined k</td>
<td>Pre-defined</td>
<td>Not pre-defined</td>
<td>Not pre-defined</td>
</tr>
<tr>
<td>Complete / partial</td>
<td>Complete</td>
<td>Complete</td>
<td>Partial</td>
</tr>
<tr>
<td>Exclusive / overlap</td>
<td>Exclusive</td>
<td>Exclusive</td>
<td>Overlap</td>
</tr>
</tbody>
</table>

• DBSCAN clustering technique applied
Spatial demarcation of potential hubs (2)

• Determination of distance measure DBSCAN:
  o Not distance based, but passenger-oriented: transfer flow based
    o \( F(i, j) = F(i, j) + F(j, i) \rightarrow F(j, i) = F(i, j) \rightarrow \) symmetric distance mat
    o \( F(i, j) = \max(F) - F(i, j) \rightarrow \) inversed, non-negative distance matrix

• Determination of DBSCAN parameters:
  o The neighborhood of a given radius \( Eps \) contains at least \( MinPoint \)
    o \( MinPoint: \) context-derived. Hub min. 2 Stop IDs \( \rightarrow \) \( MinPts = 1 \)
    o \( Eps: \) experiment values to check external validity \( \rightarrow Eps = max(F)-100 \)

Ester, Kriegel, Sander, Xu (1996)
Spatial demarcation of potential hubs (3)

• Resulting stops clustered by DBSCAN algorithm:
  o From 1650 StopIDs → transfers occurred between 910 StopIDs
  o 694 (76%) of these StopIDs is not clustered → ‘noise’
  o Remaining 216 (24%) StopIDs clustered in 62 clusters

![Histogram cluster size distribution]

• Resulting transfer flows clustered by DBSCAN algorithm:
  o Maximize within-cluster transfer flows / minimize between-cluster flows
  o 86% of all network transfer flows: within-cluster transfer flows
  o 98% of transfer flows from/to clustered StopIDs: within-cluster flows
Hub identification (1)

- From 62 clusters of potential hubs: which clusters concentrate substantial transfer flows to be considered a hub

- Analogy airline industry to apply economic metrics \((\text{Costa et al. 2010; Rodriguez-Deniz et al. 2013})\)
  - Use Herfindahl-Hirschman Index (HHI) to calculate market concentration based on market share of cluster \(i\) \(P_i\):
    \[
    HHI = \sum_{i=1}^{l} P_i^2
    \]
  - Number of ‘effective’ market players (= hubs) \(n_e = HHI^{-1}\)

- Results case study network:
  - \(HHI = 0.0889, n_e = 11.3 \rightarrow 11\) hubs identified from 62 clusters of potential hubs
Hub identification

Cumulative distribution function

Lorenz curve
Gini coefficient = 0.745
(cluster market share = 100%)
Hub identification (3)

<table>
<thead>
<tr>
<th>Cluster ID</th>
<th>Hub name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Centrum / Spui / Kalvermarkt</td>
</tr>
<tr>
<td>2</td>
<td>Central Station</td>
</tr>
<tr>
<td>4</td>
<td>Station Hollands Spoor</td>
</tr>
<tr>
<td>28</td>
<td>Laan van NOI</td>
</tr>
<tr>
<td>32</td>
<td>Brouwersgracht</td>
</tr>
<tr>
<td>35</td>
<td>The Hague Market</td>
</tr>
<tr>
<td>36</td>
<td>Wouwermanstraat</td>
</tr>
<tr>
<td>40</td>
<td>Leyenburg</td>
</tr>
<tr>
<td>41</td>
<td>Leyweg</td>
</tr>
<tr>
<td>50</td>
<td>Herenstraat</td>
</tr>
<tr>
<td>61</td>
<td>Leidschenveen</td>
</tr>
</tbody>
</table>
Hub reliability (1)

• Hub-level passenger-oriented reliability indicators:
  
  o % transferring passengers missing their connection $Q_{mc}$ at hub $s_{hub}$

  $$Q_{mc} = \frac{\sum_L \sum_R q_{r_{l1}-r_{l2}} \cdot MC_{q_{r_{l1}-r_{l2}}}}{\sum_L \sum_R q_{t,r_{l1}-r_{l2}}} \quad \forall s_{hub} \in S_{hub}$$

  o Perceived journey excess time to due to lost connection at hub

  $$PJET_{mc} = \frac{\sum_L \sum_R q_{r_{l1}-r_{l2}} \cdot MC_{q_{r_{l1}-r_{l2}}} \cdot (T^a - T^s)}{\sum_L \sum_R q_{r_{l1}-r_{l2}} \cdot MC_{q_{r_{l1}-r_{l2}}}} \quad \forall s_{hub} \in S_{hub}$$

  o Societal unreliability costs due to lost connection at the hub

  $$C_{mc} = \sum_L \sum_R q_{r_{l1}-r_{l2}} \cdot MC_{q_{r_{l1}-r_{l2}}} \cdot (T^a - T^s) \cdot VOT \quad \forall s_{hub} \in S_{hub}$$

  with $MC = \begin{cases} 1 & \text{if } r_{l2}^a > r_{l2}^s \\ 0 & \text{if } r_{l2}^a \leq r_{l2}^s \end{cases}$
Hub reliability (2)

- Example hub reliability quantification: hub Leyweg

- Average: 5.3% lost connections with on average 12 minutes additional perceived journey travel time --> yearly societal costs ≈ €18,000

<table>
<thead>
<tr>
<th>Arriving line</th>
<th>Departing line</th>
<th>Lost transfer flow</th>
<th>Total transfer flow</th>
<th>$Q_{mc}$ (%)</th>
<th>$P_{JET_{mc}}$ (min)</th>
<th>$C_{mc}$ (€ / year)</th>
</tr>
</thead>
<tbody>
<tr>
<td>21</td>
<td>23</td>
<td>16</td>
<td>318</td>
<td>5%</td>
<td>13</td>
<td>1450</td>
</tr>
<tr>
<td>21</td>
<td>25</td>
<td>6</td>
<td>269</td>
<td>2%</td>
<td>3</td>
<td>146</td>
</tr>
<tr>
<td>23</td>
<td>21</td>
<td>26</td>
<td>477</td>
<td>5%</td>
<td>15</td>
<td>2664</td>
</tr>
<tr>
<td>23</td>
<td>25</td>
<td>108</td>
<td>1344</td>
<td>8%</td>
<td>10</td>
<td>7784</td>
</tr>
<tr>
<td>25</td>
<td>21</td>
<td>16</td>
<td>441</td>
<td>4%</td>
<td>18</td>
<td>1415</td>
</tr>
<tr>
<td>25</td>
<td>23</td>
<td>46</td>
<td>1253</td>
<td>4%</td>
<td>15</td>
<td>4136</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>218</td>
<td>4102</td>
<td>5.3%</td>
<td>12.3</td>
<td><strong>€18,000</strong></td>
</tr>
</tbody>
</table>
Hub reliability (3)

- Yearly societal costs due to hub unreliability at all hubs (accounting for 86% of all transfers) for case study network: €386,000

<table>
<thead>
<tr>
<th>Cluster ID</th>
<th>Hub name</th>
<th>$Q_{mc}$ (%)</th>
<th>$P_{JET mc}$ (min)</th>
<th>$C_{mc}$ (€ / year)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>Central Station</td>
<td>3.6%</td>
<td>13.3 min</td>
<td>€ 114,000</td>
</tr>
<tr>
<td>4</td>
<td>Station Hollands Spoor</td>
<td>5.2%</td>
<td>11.9 min</td>
<td>€ 84,000</td>
</tr>
<tr>
<td>1</td>
<td>Centrum / Spui / Kalvermarkt</td>
<td>5.1%</td>
<td>12.1 min</td>
<td>€ 80,000</td>
</tr>
<tr>
<td>40</td>
<td>Leyenburg</td>
<td>3.6%</td>
<td>12.9 min</td>
<td>€ 23,000</td>
</tr>
<tr>
<td>41</td>
<td>Leyweg</td>
<td>5.3%</td>
<td>12.3 min</td>
<td>€ 18,000</td>
</tr>
<tr>
<td>50</td>
<td>Herenstraat</td>
<td>5.5%</td>
<td>12.1 min</td>
<td>€ 15,000</td>
</tr>
<tr>
<td>35</td>
<td>The Hague Market</td>
<td>3.1%</td>
<td>13.7 min</td>
<td>€ 15,000</td>
</tr>
<tr>
<td>61</td>
<td>Leidschenveen</td>
<td>2.1%</td>
<td>24.3 min</td>
<td>€ 11,000</td>
</tr>
<tr>
<td>28</td>
<td>Laan van NOI</td>
<td>4.3%</td>
<td>10.7 min</td>
<td>€ 10,000</td>
</tr>
<tr>
<td>32</td>
<td>Brouwersgracht</td>
<td>2.2%</td>
<td>12.8 min</td>
<td>€ 8,900</td>
</tr>
<tr>
<td>36</td>
<td>Wouwermanstraat</td>
<td>1.1%</td>
<td>14.0 min</td>
<td>€ 6,700</td>
</tr>
</tbody>
</table>
Conclusions & further research

• Conclusions:
  o Generic, data-driven methodology developed
  o To identify urban public transport network hubs
  o To quantify and compare hub (un)reliability
  o To express hub unreliability in monetary terms $\rightarrow$ SCBA

• Further research:
  o Incorporate hub connectivity / complexity explicitly in hub identification
  o Incorporate perceived in-vehicle time due to crowding as consequences of hub unreliability in $PJET_{mc}$ and $C_{mc}$
  o Incorporate hub unreliability in explaining passenger route choice
A data-driven approach to infer spatial characteristics and service reliability of public transport hubs

Ir. Menno Yap
Dr. ir. Niels van Oort
Dr. Oded Cats
Prof. dr. ir. Serge Hoogendoorn

M.D.Yap@TUDelft.nl
https://nielsvanoort.weblog.tudelft.nl/

May 22\textsuperscript{nd}, 2017