7.2. Big data in Statistics on Passengers Transport – a case study on Lisbon Metropolitan Area

1st July 2016, Lisbon

Statistics Portugal
Economic Statistics Department
Distributive trade, tourism and transport statistics unit
Rute Cruz Calheiros (rute.cruz@ine.pt)
Porfirio Leitão (porfirio.leitao@ine.pt)
Transport statistics in PT

Statistics Portugal -> responsibility for all national statistical production about Transports
« Passengers transport statistics

1. Introduction
2. The Lisbon M.A. and the ticketing system
3. Data details
4. Validation and imputation
5. Tables of results
6. Major challenges
7. Future applications

- Administrative data from the Regulator (Instituto da Mobilidade e dos Transportes)
- Surveys to transport companies
- Administrative data from Maritime Port Administrations
- Administrative data from the Regulator (Autoridade Nacional de Aviação Civil) and Airport Administrations

Urban mobility

Road
Rail
Inland waterways
Maritime
Air
The Lisbon metropolitan area...
The Lisbon metropolitan area

Resident population (2015):
- Portugal – 10.3 million
- Lisbon M.A.: 2.8 million, 932.8 persons per km²
The Lisbon metropolitan area
18 Municipalities

North side of river Tejo:
- Amadora
- Cascais
- Lisboa
- Loures
- Mafra
- Odivelas
- Oeiras
- Sintra
- Vila Franca de Xira

South side:
- Alcochete
- Almada
- Barreiro
- Moita
- Montijo
- Palmela
- Seixal
- Sesimbra
- Setúbal
Transports in Lisbon metropolitan area

- Actually: under supervision of the 18 Municipalities, forming a regional transport authority named “Área Metropolitana de Lisboa”;
  
  - In the past: central authority for transports in Lisbon (“Autoridade Metropolitana de Transportes de Lisboa”);
  
  - Road, inland waterways, light and heavy railway systems;
  
  - Public and private transport companies;
  
  - Consortium of the transport companies (named OTLIS) to manage data from the common ticketing system.
The ticketing system

- Contactless technology;
- Works with pre-charged cards;
- Several types of cards for different uses, personalized or for general use.
The ticketing system, complexity

- Several types of passes: single company pass, intermodal (by zones), combined operators;
- Several types of tickets to charge: single company rate, Lisbon city rate, zapping rate (by value);
- Special rates on board (only some operators);
- Special reduced rates (Social+; elder; retired; 4_18 years; Sub23 and children).
1. Introduction
2. The Lisbon M.A. and the ticketing system
3. Data details
4. Validation and imputation
5. Tables of results
6. Major challenges
7. Future applications

» The ticketing system, validation equipment

- In the entrance of road vehicles;
- In the entrance of ferries piers;
- In the entrance and exit of underground and light rail system stations;
- In the entrance and/or exit of heavy rail stations - but some stations with no physical barrier and/or no equipment on exit.
Data structure

Reflects the complexity of cards, passes, tickets and special rates.

Primary data structure:

- Serial Number – Serial number of the card
- Card Type – personalized user / universal user / multi-operator / single operator
- Title – more than 1.200 different types – time period, combination of operators, rates, discounts ...
- Date/hour – Date/hour of the interaction
- Operator – owner of the validation equipment
- Validation type – Entry or exit (when applicable)
- Stop Code – place of the interaction
- Line – Network line/segment (when applicable)

And also: separate tables with information about tickets/titles, cards and stop codes locations.
» Data volume (a)

1 working day:

- Number of real interactions: ~1.600.000 (month: ~ 43.000.000);
- Number of “missing” interactions (exits unknown): ~ 900.000 (month: ~ 25.000.000);
- Daily CSV file = ~200 Mb.

(a) Raw data, interactions with the system, before error corrections and imputation of missing entries or exits; based only on the main companies (excluding some road companies from the suburbs)
1. Introduction
2. The Lisbon M.A. and the ticketing system
3. Data details
4. Validation and imputation
5. Tables of results
6. Major challenges
7. Future applications

» Data\(^{(a)}\), interactions breakdown by modes

- Road: 27% (before exit imputations)
- Inland waterways: 3% (before exit imputations)
- Heavy rail: 21% (before partial exit imputations)
- Underground and light rail: 49%

\(^{(a)}\) Raw data, interactions with the system, before error corrections and imputation of missing entries or exits; based only on the main companies (excluding some road companies from the suburbs)
Validation

Primary validation:

- Check and correct anomalies generated in individual companies data or during the data import process (blank data, incomplete data, misinformation, ...);

- Detect and eliminate outliers and non applicable cases:
  - Station workers,
  - Other non transport users (with dozens of daily interactions, such as beggars and pickpockets, ...).
» Data process stages
(1/3)

1. Split the data by operator;

2. For each operator, design and implementation of unique procedures for data validation and processing, such as:
   • imputation of missing interactions (for each corresponding entry validation must exist an exit validation),
   • creating missing steps within each stage (changing lines in underground, for instance, which are not registered),
   • elimination of redundant interactions (consecutive entries and exits in the same station, ...),
3. Rejoin of the data;

4. Checking the consistency on a user basis, based on each card serial number:
   - daily views, adjustment of “beginning” and “ending” times between the successive daily stages (maladjusted system clocks, ...),
   - incoherent stages eliminated.
5. **Construction of 3 different micro-data tables:**
   - Set of sub-stages (unique transport movement with no change of vehicle),
   - Set of stages (movement within a transport mode with possible unregistered change of vehicle),
   - Set of trips (succession of stages, derived from the sequential tracking of each card throughout the day).
Basic methodological principles to adopt

- Concepts for sub-stages, stages and trips;
- Definitions of:
  - Outliers,
  - Minimum time gap between stages, by mode (to evaluate clocks mismatch),
  - Maximum time gap between stages (to define the beginning of the next trip), for each mode of transport and considering the period of the day,
  - Conditions to imputation of commuting trips [assuming that the end (unknown) of the first trip is the beginning (known) of the last one].
Examples of results
(fictional data)

Distribution of stages according to the transport title and time period

<table>
<thead>
<tr>
<th>Monday</th>
<th>Fictional data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time period</td>
<td>00:00/06:29</td>
</tr>
<tr>
<td>Transport title</td>
<td>12</td>
</tr>
<tr>
<td>Ticket</td>
<td></td>
</tr>
<tr>
<td></td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>16</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

1. Introduction
2. The Lisbon M.A. and the ticketing system
3. Data details
4. Validation and imputation
5. Tables of results
6. Major challenges
7. Future applications
### Examples of results
(fictional data)

#### Distribution of trips according to the operator and time period (beginning)

<table>
<thead>
<tr>
<th>Friday</th>
<th>Transport operator A</th>
<th>Transport operator B</th>
<th>Transport operator C</th>
<th>Transport operator D</th>
<th>Transport operator E</th>
<th>Transport operator F</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time Period</td>
<td>00:00/06:29</td>
<td>06:30/09:29</td>
<td>09:30/11:59</td>
<td>12:00/13:59</td>
<td>14:00/17:29</td>
<td>17:30/19:29</td>
<td>19:30/23:59</td>
</tr>
<tr>
<td>-----------</td>
<td>---------------------</td>
<td>---------------------</td>
<td>---------------------</td>
<td>---------------------</td>
<td>---------------------</td>
<td>---------------------</td>
<td>----------------</td>
</tr>
</tbody>
</table>
### Examples of results

(fictional data)

<table>
<thead>
<tr>
<th>Means of transport</th>
<th>Title type</th>
<th>Fictional data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>Type A</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>275.512</td>
<td>1.497</td>
</tr>
<tr>
<td><strong>Heavy Rail Transport</strong></td>
<td>17.771</td>
<td>999</td>
</tr>
<tr>
<td>Transport operator A</td>
<td>8.219</td>
<td>444</td>
</tr>
<tr>
<td>Transport operator B</td>
<td>9.552</td>
<td>555</td>
</tr>
<tr>
<td><strong>Light Rail Transport</strong></td>
<td>10.552</td>
<td>333</td>
</tr>
<tr>
<td>Transport operator C</td>
<td>4.776</td>
<td>222</td>
</tr>
<tr>
<td>Transport operator D</td>
<td>5.776</td>
<td>111</td>
</tr>
<tr>
<td><strong>Road transport</strong></td>
<td>233.762</td>
<td>99</td>
</tr>
<tr>
<td>Transport operator E</td>
<td>7.164</td>
<td>55</td>
</tr>
<tr>
<td>Transport operator F</td>
<td>226.598</td>
<td>44</td>
</tr>
<tr>
<td><strong>Inland waterway</strong></td>
<td>13.427</td>
<td>66</td>
</tr>
<tr>
<td>Transport operator G</td>
<td>8.842</td>
<td>11</td>
</tr>
<tr>
<td>Transport operator H</td>
<td>4.585</td>
<td>55</td>
</tr>
</tbody>
</table>
### Examples of results
(fictional data)

#### O/D matrix of the average number of stages per journey

<table>
<thead>
<tr>
<th>Wednesday, 24 hours</th>
<th>Fictional data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Origin</strong></td>
<td>Municipality A</td>
</tr>
<tr>
<td>Municipality A</td>
<td>1,000</td>
</tr>
<tr>
<td>Municipality B</td>
<td>1,100</td>
</tr>
<tr>
<td>Municipality C</td>
<td>1,200</td>
</tr>
<tr>
<td>Municipality D</td>
<td>1,300</td>
</tr>
<tr>
<td>Municipality E</td>
<td>1,400</td>
</tr>
<tr>
<td>Municipality F</td>
<td>1,500</td>
</tr>
</tbody>
</table>
Examples of results

(fictional data)

<table>
<thead>
<tr>
<th>Transport card / Tickets</th>
<th>Ticket type</th>
<th>Monday, 24 hours</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>No discount</td>
</tr>
<tr>
<td>Total</td>
<td>2,952</td>
<td>3,574</td>
</tr>
<tr>
<td>L1</td>
<td>1,111</td>
<td>3,351</td>
</tr>
<tr>
<td>L12</td>
<td>3,333</td>
<td>3,741</td>
</tr>
<tr>
<td>L123</td>
<td>2,222</td>
<td>3,541</td>
</tr>
<tr>
<td>12</td>
<td>4,444</td>
<td>-</td>
</tr>
<tr>
<td>123</td>
<td>1,123</td>
<td>-</td>
</tr>
<tr>
<td>23</td>
<td>1,100</td>
<td>-</td>
</tr>
</tbody>
</table>
Challenges

- Hardware requirements for Big Data (dedicated servers);
- Up-to-date software: advanced powerful data base management system, advanced data mining tools, other big data suitable statistical tools;
- Secure data transfer between the provider and the NSI;
- Advanced user skills (dedicated programming, database design and managing, communications and network, statistical expertise...);
- Dependency from transport operators and its administrative authority.

1. Introduction
2. The Lisbon M.A. and the ticketing system
3. Data details
4. Validation and imputation
5. Tables of results
6. Major challenges
7. Future applications
» **Strengths and opportunities**

- Exhaustive (close to) data – sampling only by choice;
- Rigorous date-time information, also for origin/destination when available;
- Possibility of tracking each card – longitudinal data along time;
- Full urban mobility picture on public transport operators.
Future applications (1/3)

Potential developments:

- Partial substitution of surveys on passenger transport, although:
  - only demand/occupation variables (not supply),
  - no estimation for fraud,
  - regional delimitation can collide with broader data (national) provided by companies to the NSI,
  - very resources consuming;

- Detailed table of results usually not provided by transport operators;
Future applications (2/3)

- Public dissemination of results that can enlighten citizens and decision makers about urban mobility;

- Ad hoc studies about impact on transport network of:
  - weather phenomena,
  - large public events,
  - network interruptions (strike/accident/operational failures,...),
  - social behavior/demographic changes,
  - new services or operators, ...
Future applications (3/3)

- Due to the detail of each origin/destination, possibility to elaborate accessibility indicators in connection to population data;

- If address data related to each card is accessible, possibility to estimate individual vehicle use (considering the first/last interactions of the day);

- Need to have collaboration from the transport authority to understand the transport systems and to obtain the required data.
Thank you!

www.ine.pt
Statistics Portugal
Economic Statistics Department
Distributive trade, tourism and transport statistics unit
Rute Cruz Calheiros (rute.cruz@ine.pt)
Porfírio Leitão (porfírio.leitao@ine.pt)