INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

scorus Conference 2016

Indicators for territorial policies: closing data gaps by using traditional and new sources and methods

Lisbon | Statistics Portugal | 29 June - 1 July

Session 7:
The potential of open data and big data for territorial information

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

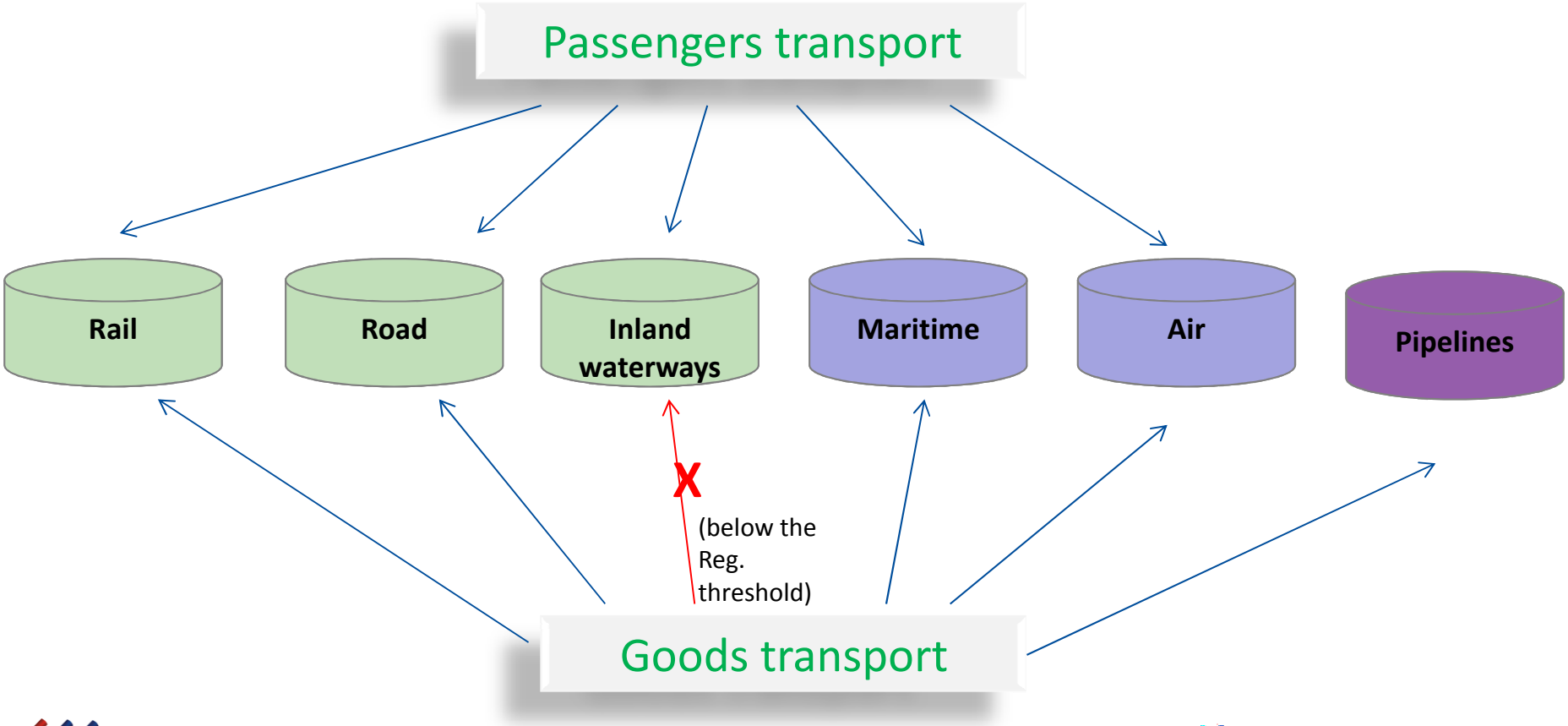
1st July 2016, Lisbon

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

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

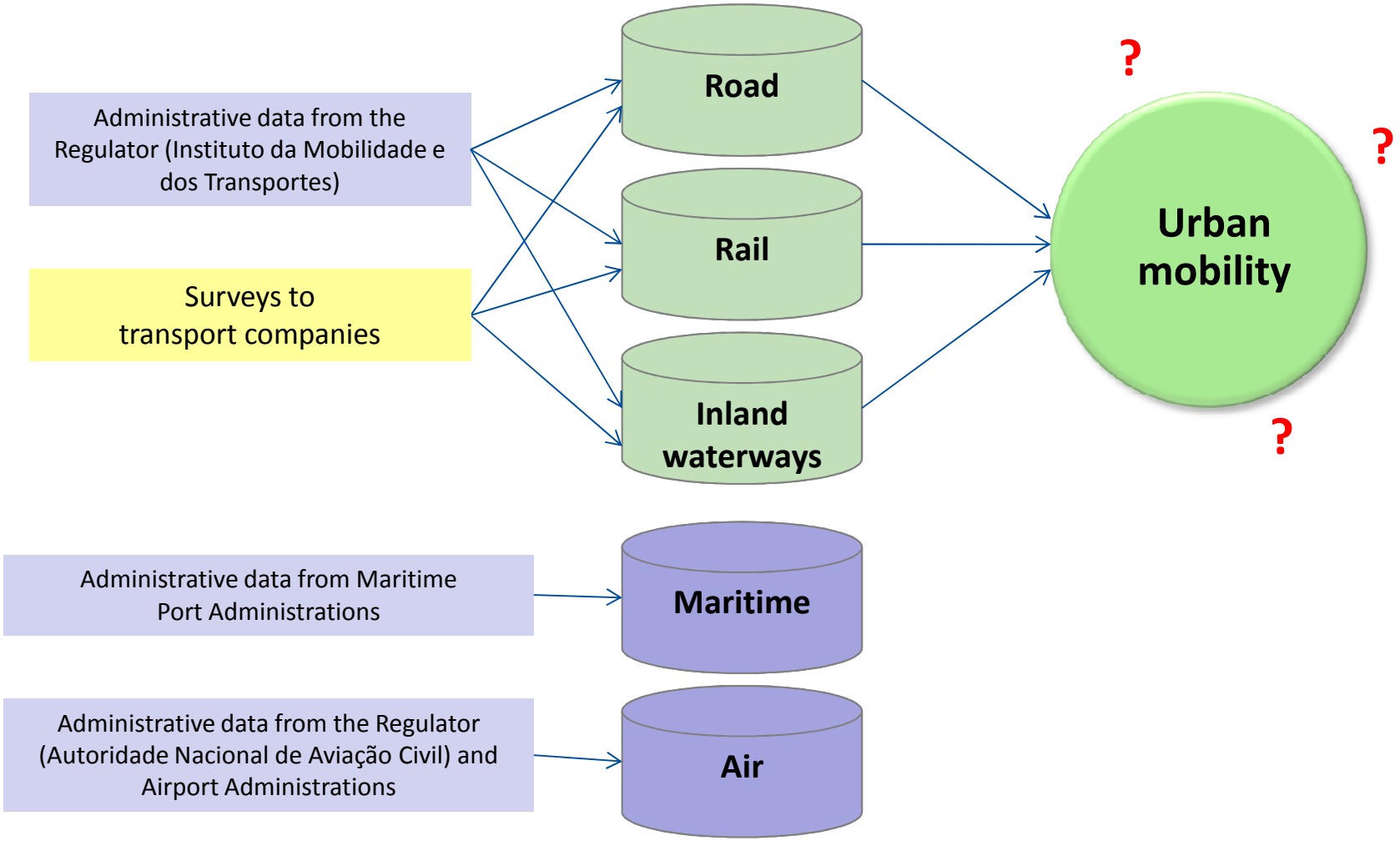
» Transport statistics in PT

Statistics Portugal -> responsibility for all national statistical production about Transports



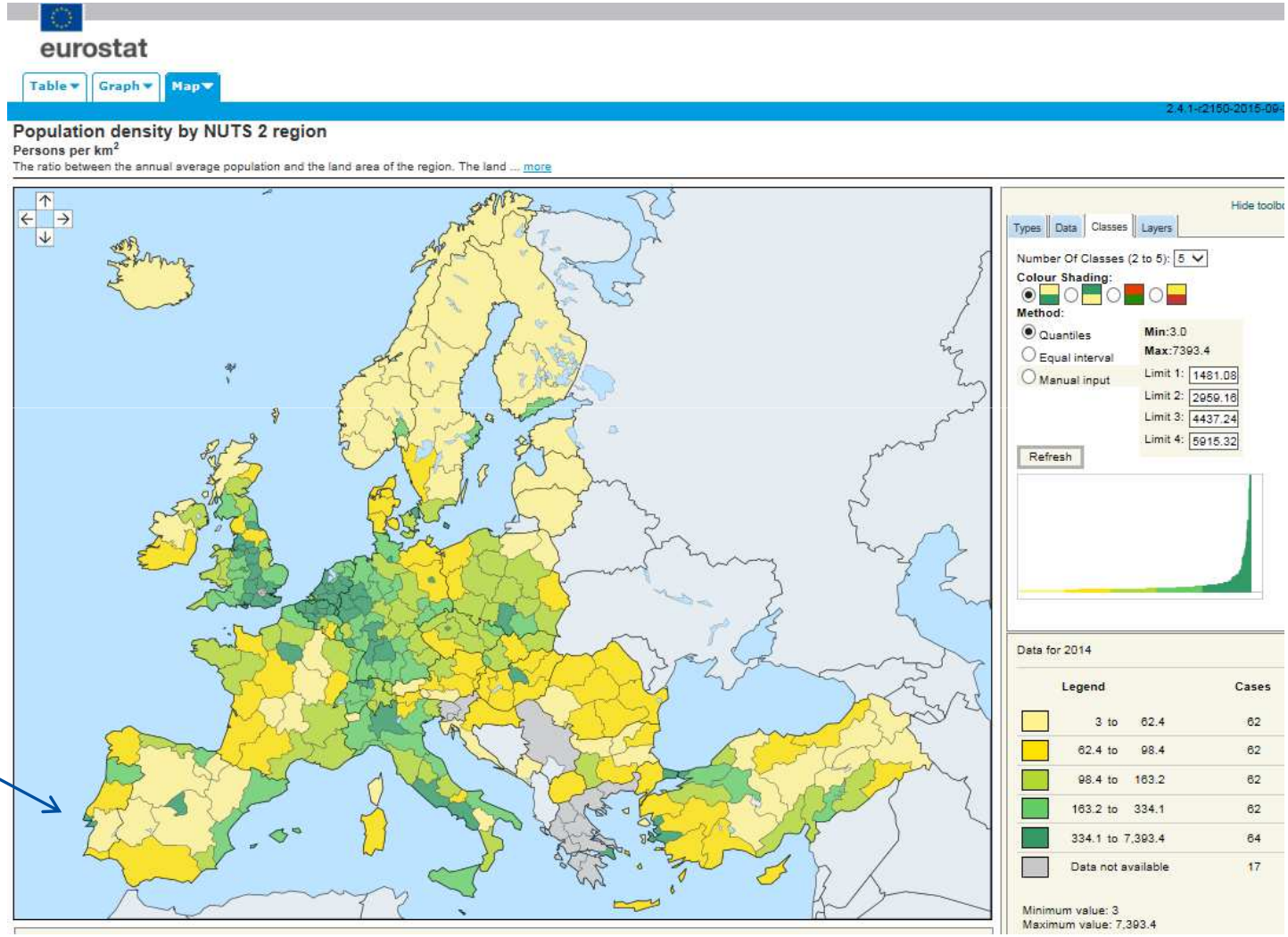
1. **Introduction**
2. The Lisbon M.A. and the ticketing system
3. Data details
4. Validation and imputation
5. Tables of results
6. Major challenges
7. Future applications

» Passengers transport statistics



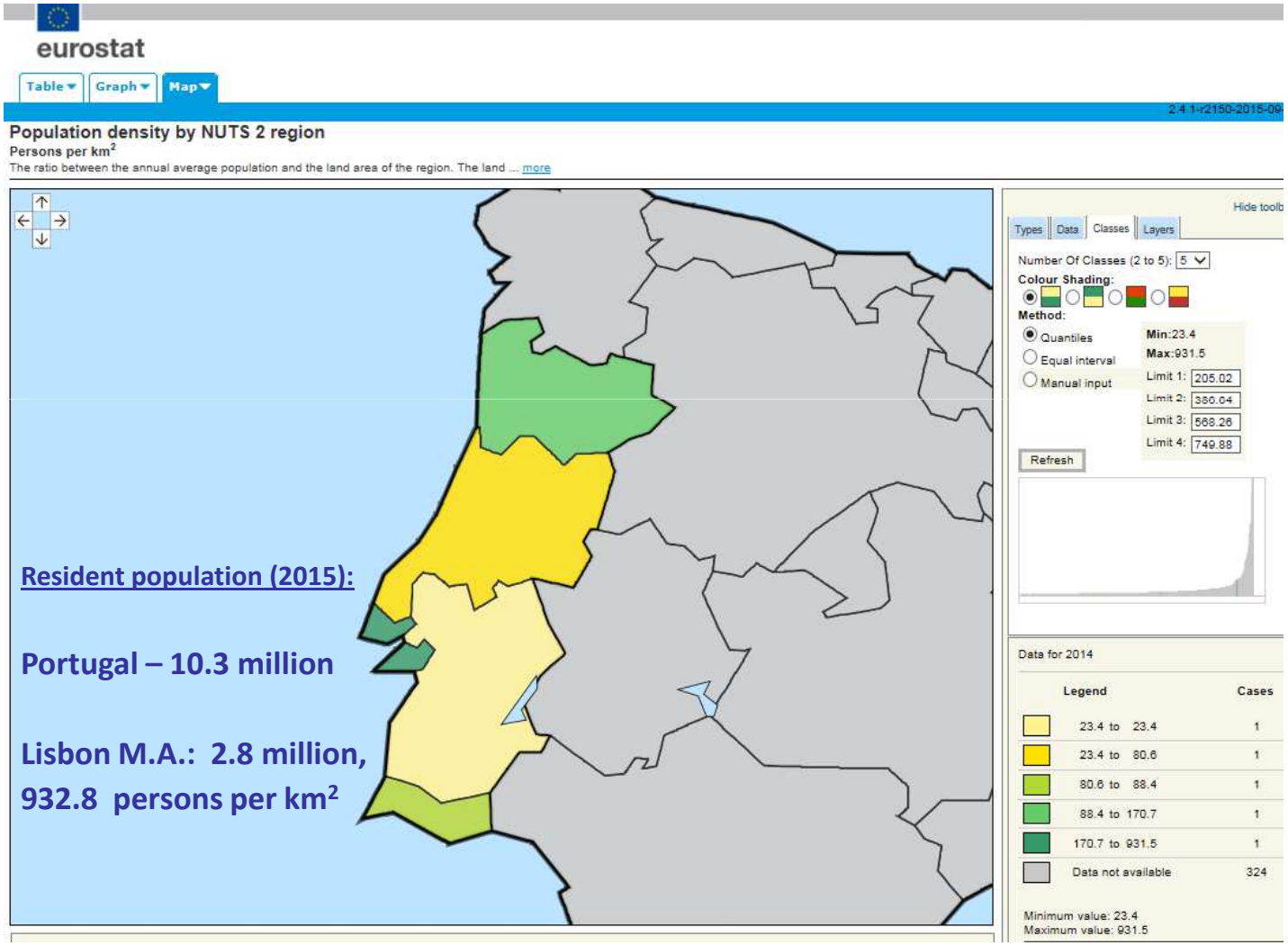
1. Introduction
2. **The Lisbon M.A. and the ticketing system**
3. Data details
4. Validation and imputation
5. Tables of results
6. Major challenges
7. Future applications

» The Lisbon metropolitan area ...



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 Lisbon metropolitan area



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 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**

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

» Transports in Lisbon metropolitan area

- Actually: under supervision of the 18 Municipalities, forming a regional transport authority named “Área Metropolitana de Lisboa”;

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.

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

- Contactless technology;
- Works with pre-charged cards;

- Several types of cards for different uses, personalized or for general use.

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, 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.

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 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.

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 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

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

- 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, ...).

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 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, …),

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 process stages
(2/3)

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.

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 process stages
(3/3)

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).

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

» 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].

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 stages according to the transport title and time period

Monday — **Fictional data**

Transport title / Ticket	Time period							
	00:00/06:29	06:30/09:29	09:30/11:59	12:00/13:59	14:00/17:29	17:30/19:29	19:30/23:59	Total
12	10	11	12	13	14	15	16	91
23	30	33	36	39	42	45	48	273
123	90	99	108	117	126	135	144	819
24H	270	297	324	351	378	405	432	2.457
72H	810	891	972	1.053	1.134	1.215	1.296	7.371
Animal	2.430	2.673	2.916	3.159	3.402	3.645	3.888	22.113
48h ticket	7.290	8.019	8.748	9.477	10.206	10.935	11.664	66.339
Single ticket	21.870	24.057	26.244	28.431	30.618	32.805	34.992	199.017
BUC	65.610	72.171	78.732	85.293	91.854	98.415	104.976	597.051
Total	98.410	108.251	118.092	127.933	137.774	147.615	157.456	895.531

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)

Friday — **Fictional data**

Transport operator	Time Period							
	00:00/06:29	06:30/09:29	09:30/11:59	12:00/13:59	14:00/17:29	17:30/19:29	19:30/23:59	Total
Transport operator A	11	22	33	44	55	66	77	308
Transport operator B	33	66	99	132	165	198	231	924
Transport operator C	99	198	297	396	495	594	693	2.772
Transport operator D	297	594	891	1.188	1.485	1.782	2.079	8.316
Transport operator E	891	1.782	2.673	3.564	4.455	5.346	6.237	24.948
Transport operator F	2.673	5.346	8.019	10.692	13.365	16.038	18.711	74.844
Total	4.004	8.008	12.012	16.016	20.020	24.024	28.028	112.112

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 stages by means of transport and title type

Monday, 00:00 / 06:29 — **Fictional data**

Means of transport	Title type					
	Total	Type A	Type B	Type C	Type D	Type E
Total	**275.512**	**1.497**	**10.940**	**9.449**	**230.963**	**22.663**
Heavy Rail Transport	**17.771**	**999**	**1.887**	**6.443**	**7.443**	**999**
Transport operator A	8.219	444	888	5.555	777	555
Transport operator B	9.552	555	999	888	6.666	444
Light Rail Transport	**10.552**	**333**	**7.777**	**1.221**	**666**	**555**
Transport operator C	4.776	222	3.333	666	333	222
Transport operator D	5.776	111	4.444	555	333	333
Road transport	**233.762**	**99**	**1.221**	**777**	**222.777**	**8.888**
Transport operator E	7.164	55	555	444	555	5.555
Transport operator F	226.598	44	666	333	222.222	3.333
Inland waterway	**13.427**	**66**	**55**	**1.008**	**77**	**12.221**
Transport operator G	8.842	11	22	999	33	7.777
Transport operator H	4.585	55	33	9	44	4.444

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)

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

Wednesday, 24 hours — **Fictional data**

Origin \ Destination	Municipality A	Municipality B	Municipality C	Municipality D	Municipality E	Municipality F
Municipality A	1,000	1,010	1,020	1,030	1,041	1,051
Municipality B	1,100	1,000	1,010	1,020	1,030	1,041
Municipality C	1,200	1,212	1,000	1,010	1,020	1,030
Municipality D	1,300	1,313	1,326	1,000	1,010	1,020
Municipality E	1,400	1,414	1,428	1,000	1,000	1,010
Municipality F	1,500	1,515	1,530	1,545	1,561	1,000

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)

Average number of stages, by transport card and ticket type, by day of the week and time period of the first journey of the day

Monday, 24 hours — **Fictional data**

Transport card / Tickets	Ticket type					
	Total	No discount	Special rate A	Special rate B	Special rate C	Special rate D
Total	**2,952**	**3,574**	**2,541**	**3,754**	**1,974**	**2,745**
L1	**1,111**	3,351	2,222	-	-	2,766
L12	**3,333**	3,741	1,111	-	-	2,633
L123	**2,222**	3,541	1,111	-	-	2,654
12	**4,444**	-	3,333	2,122	-	-
123	**1,123**	-	2,222	-	1,871	-
23	**1,100**	-	4,321	-	-	-

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

» 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.

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**

» 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;

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**

» 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, …

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**

» 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.

INSTITUTO NACIONAL DE ESTATÍSTICA
STATISTICS PORTUGAL

scorus Conference 2016

Indicators for territorial policies: closing data gaps by using traditional and new sources and methods

Lisbon | Statistics Portugal | 29 June - 1 July

Obrigada pela vossa atenção!

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 (porfirio.leitao@ine.pt)