EUROPEAN TRANSPORT CONFERENCE 2017
Barcelona, Thursday 5 October 2017

Treatment and cleaning data in mobility surveys

Maite Perez-Perez, Manel Pons, Elisabeth Queralt and Jorge Cátedra

Barcelona Institute of Regional and Metropolitan Studies (IERMB), Cerdanyola del Vallès, Spain

OVERVIEW

1. Introduction
2. Mobility Surveys: EMEF case
3. Data cleaning
4. Results
5. Conclusions

1. INTRODUCTION

Treatment and cleaning data in mobility surveys

RESOURCE: mobility surveys

Sensitive Information

Journeys:

- How?
- When?
- How much?
- Where?
- Because?

Dynamic Information

- How are the individuals?
- Are the journeys consistent?

→ *e.g.: Active employees who do not perform trips for work*

Reliable, consistent and quality information

DATA CLEANING

APPLICATIONS: multiple analysis

iermb
Conseqüències econòmiques i
territorials de l'estació de La Sagrera

1. INTRODUCTION

Objectives

- Emphasize that **before analyzing** any type of **data**, these must **be cleaned** and verified **to avoid errors** in the analysis.

- Explain **the data cleaning technique**, especially application to **mobility surveys**.

- **Comprehensive method of quality control** in CATI interviews, where **the data cleaning** technique is **a built-in part**.

- **Mobility surveys** allow obtaining information and elements of analysis about the **mobility patterns** of the population and for this reason **the data** must be of **very good quality**.

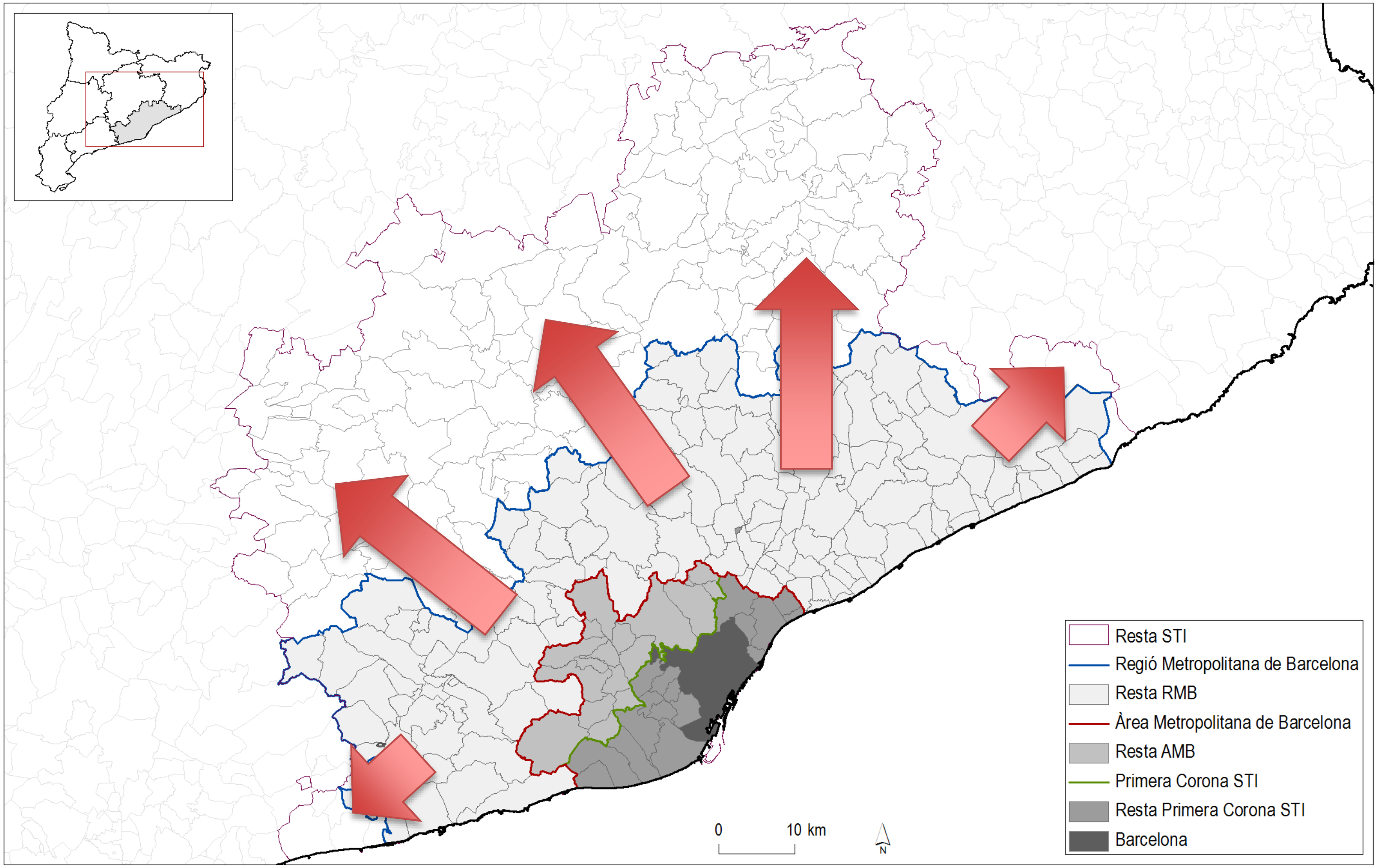
2. MOBILITY SURVEYS: EMEF CASE 2003-2016

Methodology

- **Target:** Population of 16 y.o. and over, living in the territory analysed
- **Sampling:** probabilistic sample selected by stratified/bietapic cluster sampling, where the last unit is the individual.
- **Surveying technique:** CATI
- **Design weights:** based on stratification variables and post-stratification variables. Population distribution obtained from annual register of inhabitants by Idescat.
- **Questionnaire:** 3 content blocks
  - Description of journeys
  - Subjective mobility and opinion
  - Socio-demographic characterization of respondents

2. MOBILITY SURVEYS: EMEF CASE

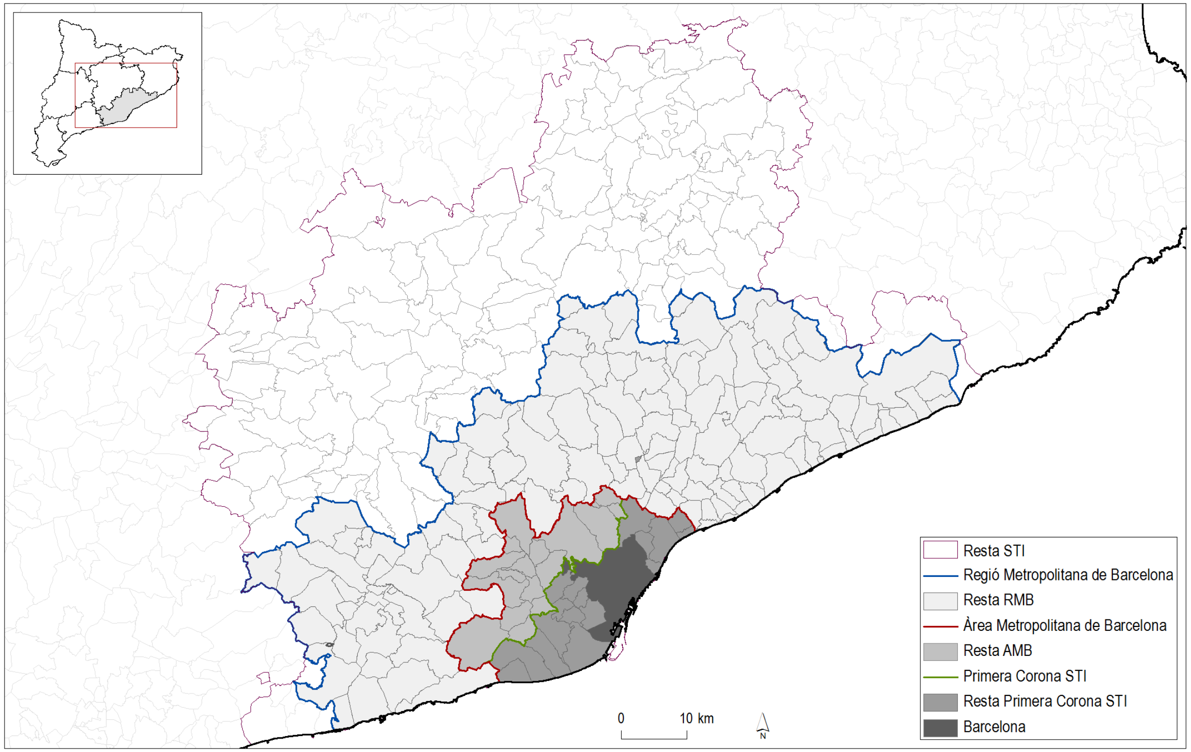
EMEF territorial scope increased from RMB to STI (2014) and almost to al province of Barcelona in 2016

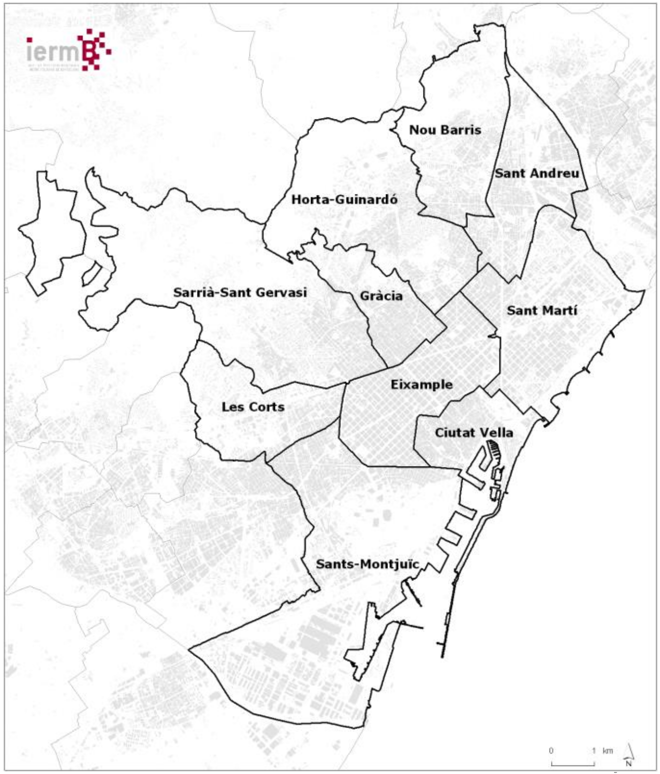


2. MOBILITY SURVEYS: EMEF CASE

Complexity of territorial stratification

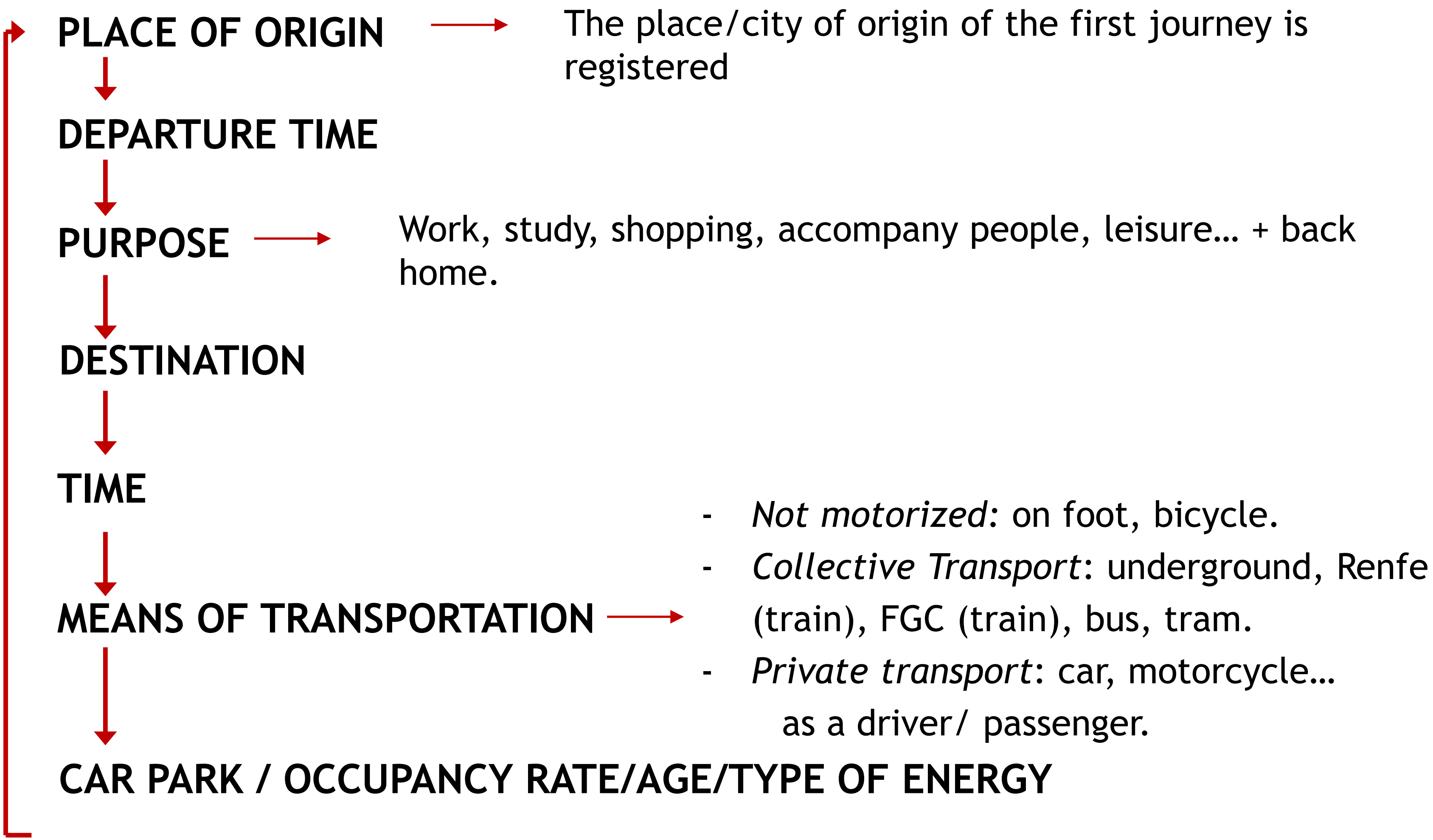
Territorial stratification is complex in sampling considering the posterior territorial analysis:





2. MOBILITY SURVEYS: EMEF CASE

Questionnaire of journeys



PLACE OF ORIGIN → The place/city of origin of the first journey is registered

DEPARTURE TIME

PURPOSE → Work, study, shopping, accompany people, leisure… + back home.

DESTINATION

TIME

MEANS OF TRANSPORTATION →

- *Not motorized*: on foot, bicycle.
- *Collective Transport*: underground, Renfe (train), FGC (train), bus, tram.
- *Private transport*: car, motorcycle… as a driver/ passenger.

CAR PARK / OCCUPANCY RATE/AGE/TYPE OF ENERGY

3. DATA CLEANING

Data cleaning: why?

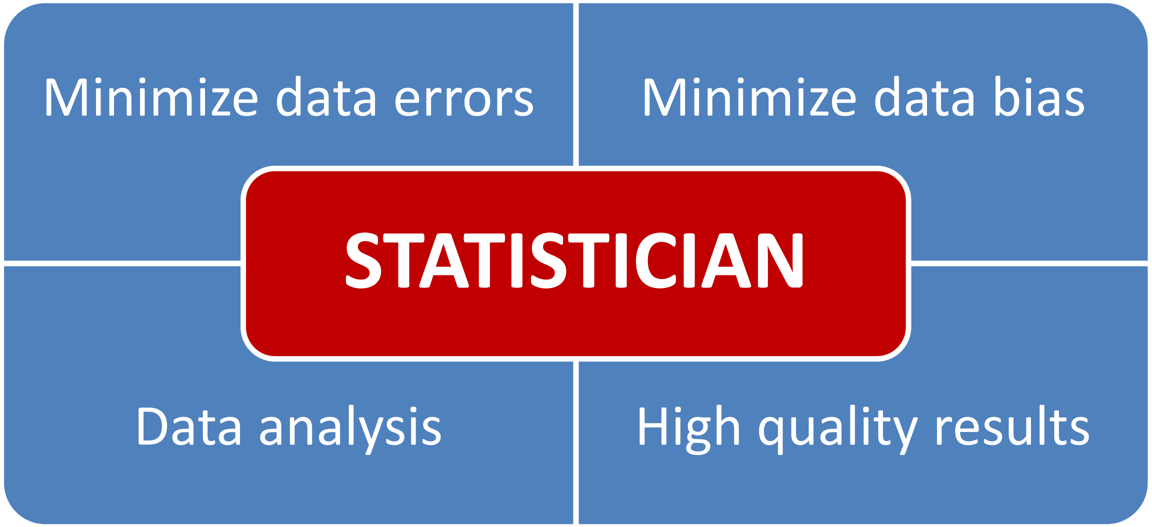
- All analysis of the information requires that the data be of quality, otherwise:

  GIGO: Garbage In, Garbage Out

- Usually some investigators don't perform the necessary checks on the data collected through questionnaire/fieldwork and performed statistical analysis of the 'gross' data directly without thinking about the multiple problems, bias, errors and inconsistencies to be found (Rial et alt, 2001)
  - Errors due to the measuring instrument used for collection
  - Data recording errors
  - Response coding errors
  - No-response errors
  - Inconsistencies of relationships between variables

3. DATA CLEANING

Data cleaning: for whom?



Minimize data errors

Minimize data bias

STATISTICIAN

Data analysis

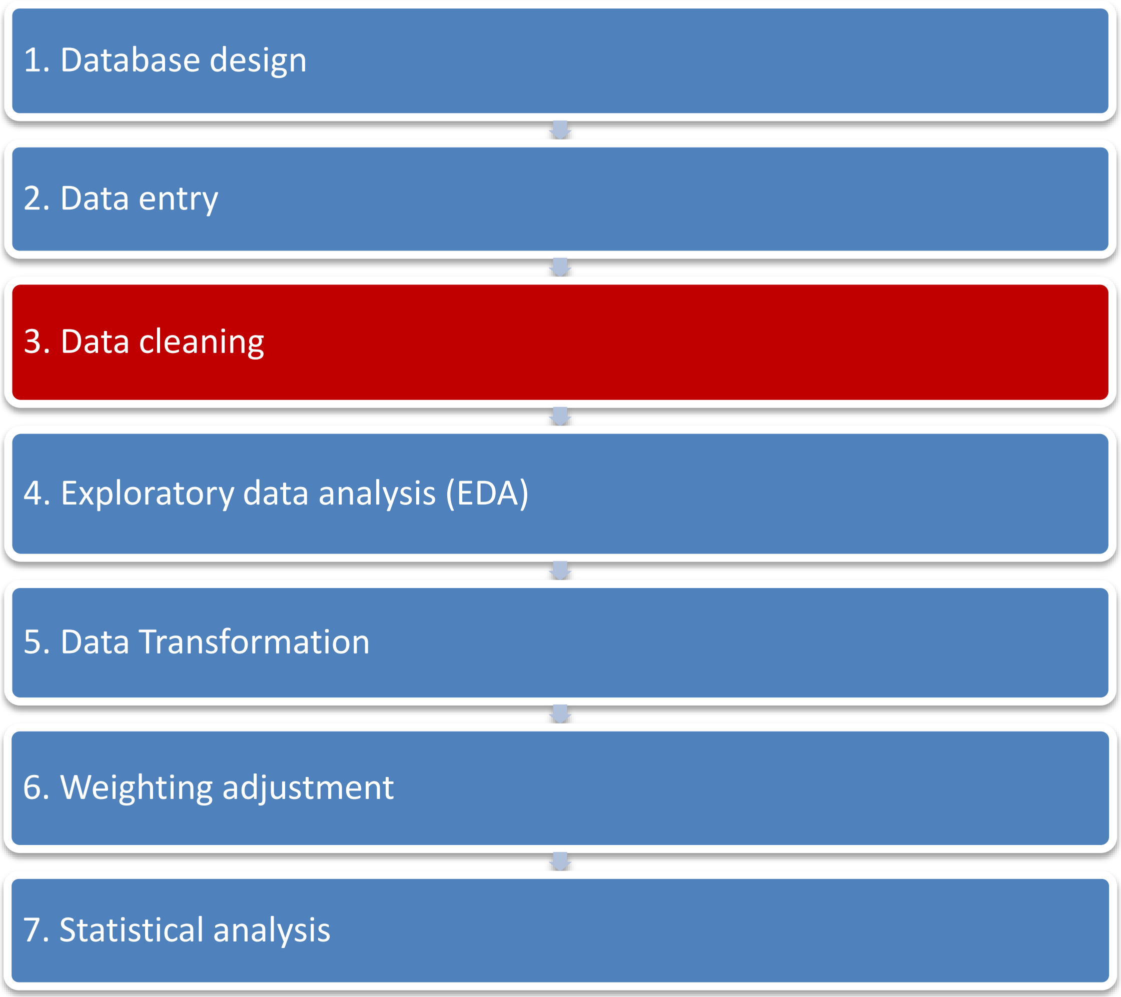
High quality results

- The process of obtaining information and data must be controlled at all times.
- The importance of the statistician‘s work lies in his ability to design a program debugging data which are obtained through a questionnaire, fieldwork or surveying technique.

3. DATA CLEANING

Statistical data processing in a research

Adaptation from Melià, 1990



1. Database design

2. Data entry

3. Data cleaning

4. Exploratory data analysis (EDA)

5. Data Transformation

6. Weighting adjustment

7. Statistical analysis

3. DATA CLEANING

Data cleaning in mobility surveys

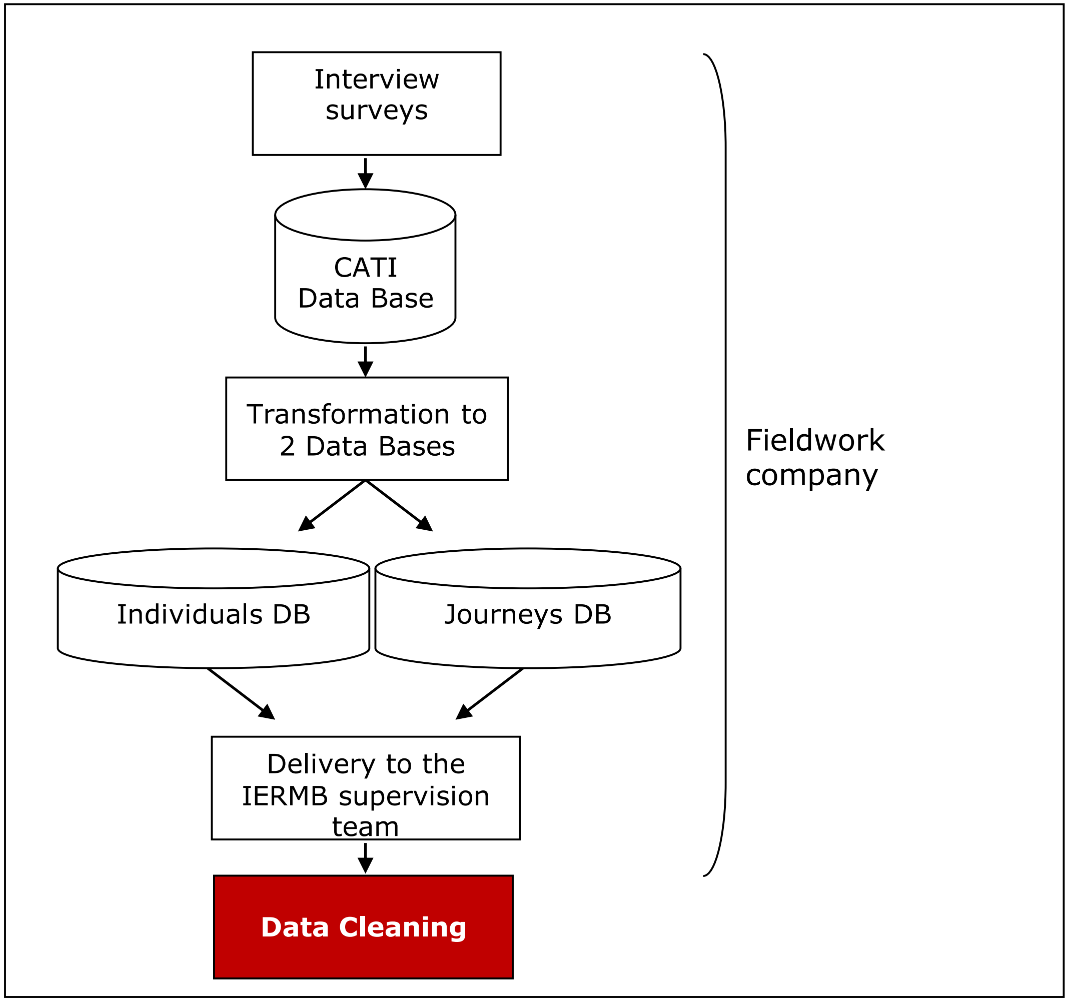
- OBJECTIVES:
  - **Corrective and preventive actions** to ensure the stability and simplification of the process at all times
  - **Detect and correct** potential bias, inconsistencies and errors in the data obtained directly from interviewee
  - **To answer to the objectives and to satisfy the analyzes** of the research hypotheses with data of the highest quality.

- HOW: **data cleaning to the first 24-48 hours** after the completion of the interview, so that **solves 99% of the inconsistencies and errors detected**
  - **minimize** the time between fact and memory, **bias produced by the 'telescope effect'**
  - Currently, recording interviews **minimizes inconveniences** to the survey respondents

3. DATA CLEANING

Data cleaning in mobility surveys

- The fieldwork technicians divide the CATI data base information according to the characteristics of the data:
  - Individuals DB
  - Journeys DB
- The IERMB supervision team subsequently do consistency analysis and data cleaning.



3. DATA CLEANING

Data cleaning in mobility surveys

During the phases of data collection and data cleaning, the **supervision team has the support of a statistician** whose main functions:

- Making the data cleaning script for individuals data base and journeys data base
- To give support the external supervisors in the data cleaning procedure, both via telephone and in person.
- To give support fieldwork companies in incidents that happen during the collection of data from interview surveys

3. DATA CLEANING

Data cleaning procedure: IERMB.JIT

- **JIT: Just in Time o Toyota method**. Objectives:
  - To reduce the management cost
  - Do not produce on a forecast basis but in real time and with real data

- **Adaptation** of the method: IERMB.JIT:
  - Minimize delivery time: delivery of the data between 24 and 48 hours.
  - Control of the production: controlling the production allows to organize quickly.
  - Zero tolerance to errors: nothing is allowed to analyze without the security of being able to do it without bias or errors.
  - Minimize technical stops: tray to prevent any problem paralyzes the fieldwork production.

3. DATA CLEANING

Data cleaning procedure: IERMB.JIT

Four important conditions:

1. Making by the IERMB data cleaning team with ad-hoc training to use the scripts.

2. Use an specific database format according to the definition provided by the IERMB coordination team.

3. The data cleaning is carried out immediately, between 24 and 48 hours after the interviews to eliminate the bias of the telescope effect.

4. Each database (individuals and journeys) has its own data cleaning script.

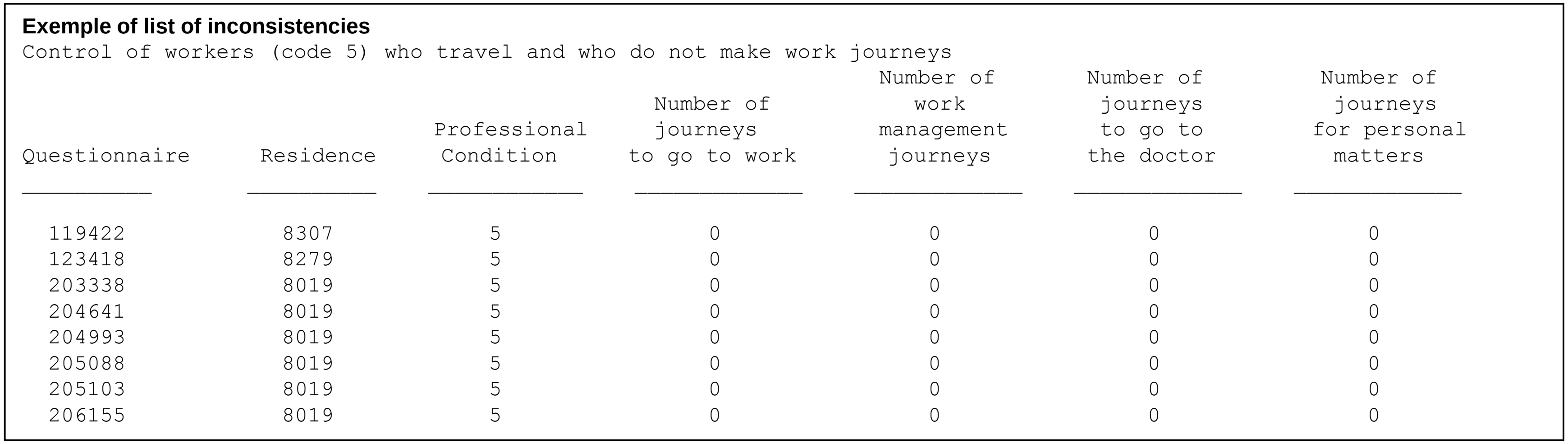
3. DATA CLEANING

Inconsistencies control procedure

- **Control of ‘out of range’ codes**
- **Control of the relationships between the responses of different variables**
  - Filters control
  - Control of simple logical relationships
  - Control of complex logical relationships
- **Control of relations between the two databases (individuals and journeys)**

3. DATA CLEANING

Inconsistencies control procedure



Exemple of list of inconsistencies

```
Control of workers (code 5) who travel and who do not make work journeys
                                                             Number of      Number of       Number of
                                              Number of        work          journeys        journeys
                              Professional     journeys      management       to go to      for personal
Questionnaire    Residence     Condition     to go to work    journeys       the doctor       matters

_____________   ___________   ____________   _____________  _____________  _____________  _____________

   119422           8307           5               0              0              0              0
   123418           8279           5               0              0              0              0
   203338           8019           5               0              0              0              0
   204641           8019           5               0              0              0              0
   204993           8019           5               0              0              0              0
   205088           8019           5               0              0              0              0
   205103           8019           5               0              0              0              0
   206155           8019           5               0              0              0              0
```

- Review each record and the inconsistencies detected in it
- Review if there are observations about the inconsistency analyzed and in case there was no immediate solution, listen to the recorded interview.
- In the case of not finding the solution to the recorded interview, perform the redial
- Make the corresponding corrections about the inconsistency
- 2nd cleaning: possible errors are not detected or appear as a result of solving previous errors
- Inconsistency without solution: depending on the severity, the interview is removed from final database.

3. DATA CLEANING

Sample control

- The control of the sample is one of the basic processes in any survey study.

- Targets:
  - **To detect drifts** in the sample → corrective measures in the bias
  - Detect **strata that are more difficult** to interview: the more mobile, older people
  - **Avoid / minimize production queues**

- **CATI methodology:** much easier, quicker and more economical, allowing production to fit and balance well with the initial sample design

3. DATA CLEANING

Second phase: Fusion, "final data cleaning" and preparation of databases.

- Merge of daily database: individual DB and journeys DB.
- Run of "final data cleaning" procedure
  - Objectives:
    - Control possible numbers of questionnaires repeated.
    - Encode those open answers that are pending
    - Control the points checked in the first consistency analysis and the annotations in the diaries in regard to "not serious" inconsistencies.
- Carry out of an exploratory data analysis (through frequencies and descriptions) to observe:
  - The variables are well coded and labeled
  - The data complies with the filters
  - Extreme Values ...

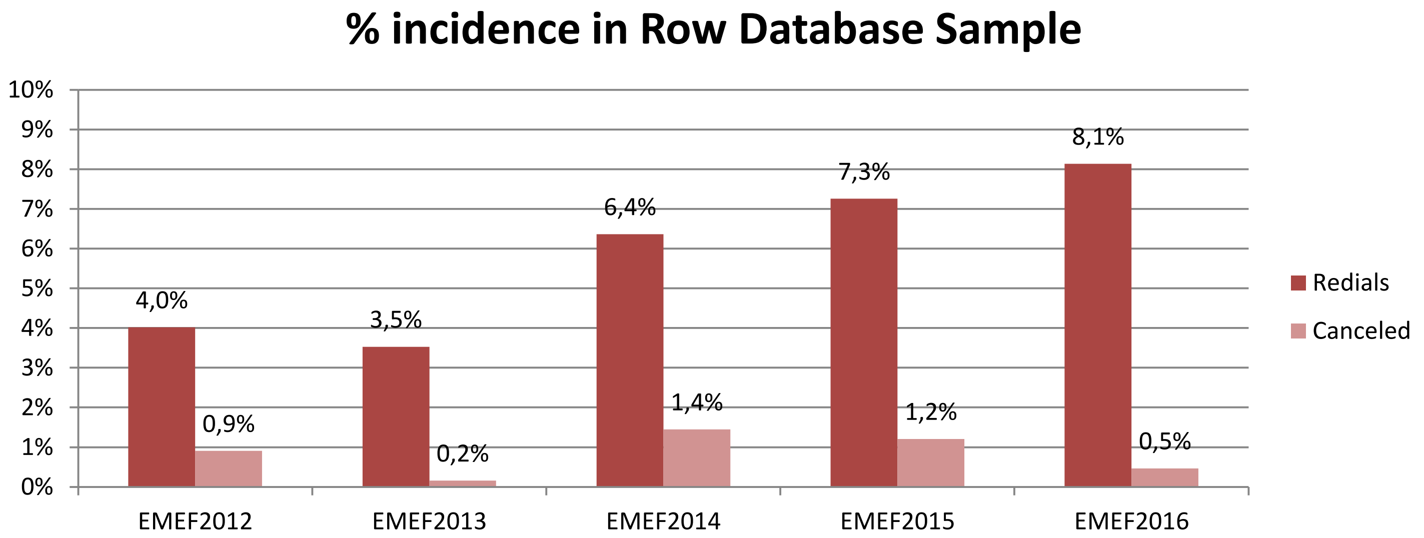
3. DATA CLEANING

“Final data cleaning“, really?

- Small inconsistencies appear that require subsequent modifications to the databases:
  - I.E: in the study of EMQ 2006 (106,091 individuals and 406,366 journeys) this can be accentuated due to their complexity in the number of variables and the volume of records
    - In the journey database, approximately 20 modifications were made (less than 5 out of every 100,000 records).
    - In the individuals database the number of modifications was null.
  - This indicates the high degree of efficiency and reliability of the process of analysis of consistency and data cleaning that was carried out throughout all fieldwork.
- But… This efficiency should not be a limit to carry out future controls:
  - I.E: Control the population indicators and other important indicators that affect the mobility of individuals such as those obtained from the Active Population Survey (EPA).

4. RESULTS

Data cleaning

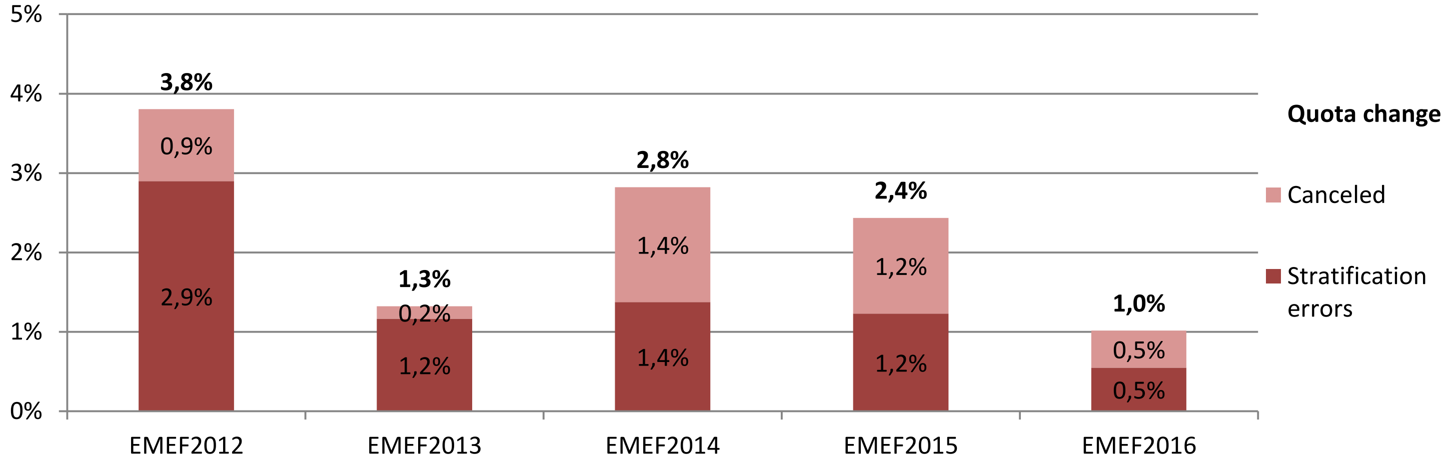


Row DB sample	6,521	6,346	9,600	9,606	9,646
Valid sample	6,462	6,336	9,461	9,490	9,601

- EMEF2014: EMEF territorial scope increased from RMB to STI
- EMEF2016: EMEF sampling methodology changed : until 2015 random calls were made and since 2016 census records are used from official statistical office of Catalonia (Idescat)

4. RESULTS

Data cleaning



% Quota change in Row DB Sample

Row DB sample	6,521	6,346	9,600	9,606	9,646
Quota changes	248	84	271	234	98

- EMEF2016: EMEF sampling methodology changed : until 2015 random calls were made and since 2016 census records are used from official statistical office of Catalonia (Idescat) and the stratification errors was reduced (n=53)
- Hypothesis: to use census records minimize stratification errors because each sample unit is nominal and it is known age, gender and residence.

5. CONCLUSIONS

Data cleaning: ‘investment’ for the analysis

- **Fieldwork is not just an information ‘gathering’ operation that is delegated to a fieldwork company**. There **can not be inconsistencies nor errors**.
- **Do data cleaning at the same** time as when the fieldwork is **done allows for accomplish the calendars and the research planning** carried out a priori by researchers, in such a way that **no more costs are generated** from the estimated ones from the beginning.
- Aim of data cleaning: having primary data of high quality allows for high quality analyzes: **Quality In Quality Out** (not Garbage In Garbage Out)
- The **methodology is adaptable**
- **The data** obtained through a fieldwork **must be treated** in parallel along the same **to minimize** the **errors** and **biases** that are inherent to many factors.
- **Win-win concept:** the initial rejection of fieldwork companies that their work can be audited by an external team can be easily defeated when it will be shown that this control can improve internal processes and take advantage of them for other studies.

Treatment and cleaning data in mobility surveys

Thank you very much for your attention!

Maite.Perez@uab.cat
Manel.Pons.Sanvidal@uab.cat

iermB
INSTITUT D'ESTUDIS REGIONALS
I METROPOLITANS DE BARCELONA