CDR Data vs. Long-Distance Travel Surveys

Maxim Janzen, Maarten Vanhoof,
Kay W. Axhausen, Zbigniew Smoreda

IVT, ETH Zurich
Open Lab, Newcastle University
Orange Labs, Paris

01 July 2016

Institut für Verkehrsplanung und Transportsysteme
Institute for Transport Planning and Systems

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Outline

1. Motivation

2. CDR Data
   - Description
   - Framework
   - Identifying Long Distance Tours

3. Validation
   - French National Travel Survey

4. Conclusion

Motivation

Long-Distance Travel

- Responsible for 35-50% of overall VMT.
- Need for models and simulations.
- Need for reliable data sources.

Motivation

Long-Distance Travel

- Responsible for 35-50% of overall VMT.
- Need for models and simulations.
- Need for reliable data sources.

Problem:

Long-distance travel surveys are limited:

- known to report low trip rates,
- number of observations is comparably low.

Alternative data sources are needed.

Mobile Phone Billing Data

The biggest data set available to researchers at Orange Labs.

Some facts:

- reports all GSM actions (originating/terminating calls/SMS) in Orange network
- for each action a Call Data Record (CDR) appears in the data
- users are anonymised
- covers the time period: 16 May 2007 till 15 October 2007
- in total 22.3 million customers
- in total 15.5 billion CDRs

Advantages and Drawbacks of CDR Data

Advantages:

- The amount of data is huge.
- The effort needed to collect the (raw) data is much lower than for surveys.

Drawbacks:

- The action frequency is low (back in 2007).
- Not precise, because just the position of (one of) the next towers is known.
- No travel purposes, modes etc. are available.
- No sociodemographic information is available.
- In this case: no roaming information.

Methodology - Framework

Approach:

1. Identify home locations.
2. Select customers (by home location).
3. Extract data for selected customers.
4. Reconstruct long-distance tours.
5. Store the tours.
6. Impute a tour purpose.
7. Compare results to survey results

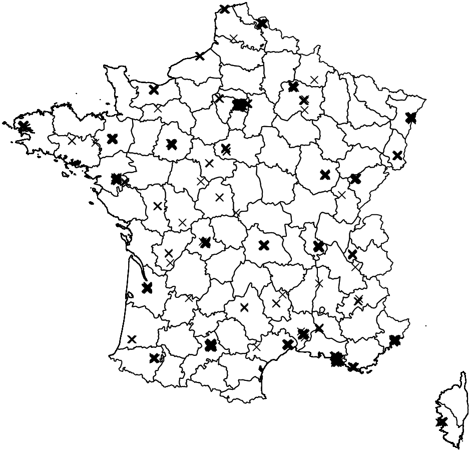
Methodology - Framework

Approach:

1. Identify home locations.
2. Select customers (by home location).
3. Extract data for selected customers.
4. **Reconstruct long-distance tours.**
5. Store the tours.
6. Impute a tour purpose.
7. **Compare results to survey results**

Selected Municipalities - Figure

14854 towers in 2977 distinct locations are considered



Selected Customers - Statistics

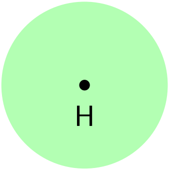
Population [*in* 1000]	Tracked Persons	Number of Communes
Paris	4953	1
200-900	19394	10
100-200	25294	13
50-100	9580	5
20-50	7461	4
10-20	7730	5
5-10	3190	5
1-5	1376	7
rural (< 1)	896	8
Total	79874	58

Identifying Long Distance Tours - Algorithm

CDR Long-Distance-Tour Reconstruction Algorithm

```
for all customers $C$ do
    $cdr\_set \leftarrow get\_cdr(C)$
    $order(cdr\_set, time)$
    for all $cdr \in cdr\_set$ do
        if not $next(cdr) \in UE(C)$ then
            new tour $t$
            while not $cdr \in UE(C)$ do
                $t \leftarrow t + cdr$
                $cdr \leftarrow next(cdr)$
            end while
            $tour\_set \leftarrow tour\_set + tour$
        end if
    end for
end for
```

LD Tour Reconstruction



Legend

H - Home anchor,

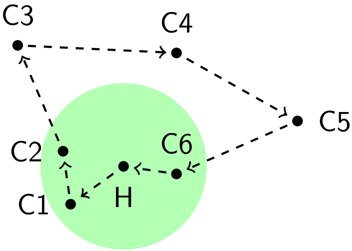
C1...C6 - CDR positions,

🟢 - User environment,

⤻ - Reconstructed tour,

⇢ - Real world tour

LD Tour Reconstruction



Legend

H - Home anchor,

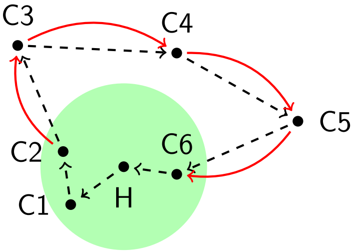
C1...C6 - CDR positions,

🟢 - User environment,

→ - Reconstructed tour,

⇢ - Real world tour

LD Tour Reconstruction



Legend

H - Home anchor,

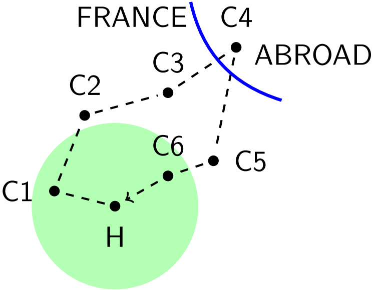
C1...C6 - CDR positions,

🟢 - User environment,

→ - Reconstructed tour,

-→ - Real world tour

Problem I - International Tours



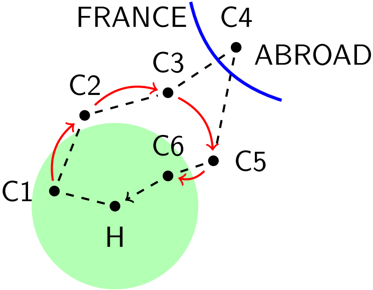
Legend

H - Home anchor, C1...C6 - CDR positions,

🟢 - User environment, ↷ - Reconstructed tour,

-→ - Real world tour

Problem I - International Tours



Legend

H - Home anchor,

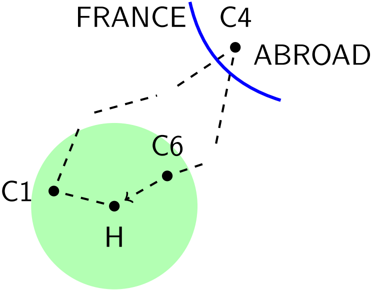
C1...C6 - CDR positions,

🟢 - User environment,

↷ - Reconstructed tour,

⇢ - Real world tour

Problem I - International Tours



Legend

H - Home anchor,

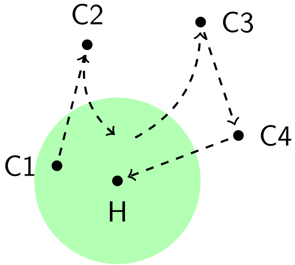
C1...C6 - CDR positions,

🟢 - User environment,

↷ - Reconstructed tour,

⇢ - Real world tour

Problem II - Merging two Tours



Legend

H - Home anchor,

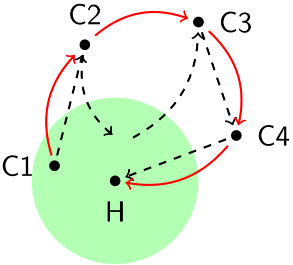
C1...C6 - CDR positions,

🟢 - User environment,

→ - Reconstructed tour,

⇢ - Real world tour

Problem II - Merging two Tours



Legend

H - Home anchor,

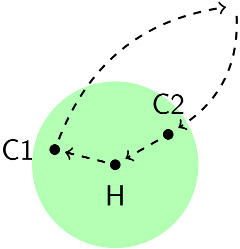
C1...C6 - CDR positions,

🟢 - User environment,

→ - Reconstructed tour,

⇢ - Real world tour

Problem III - Missing a Tour



Legend

H - Home anchor,

C1...C6 - CDR positions,

🟢 - User environment,

⤻ - Reconstructed tour,

⇢ - Real world tour

Results

Main Question:
CDR Data = Survey Data ?

French National Travel Survey

Enquête Nationale Transports et Déplacements (ENTD)

- performed every 10-15 years:
  1967, 1974, 1982, 1994, 2008
- we focus on last one: April 2007-April 2008 (6 waves)
- cooperation of a large number of actors, including ministries (CGDD, DGAC, RDG, DRAST, DSCR, DGITM), INSEE, Ifsttar, the Directorate of Tourism, SNCF, RFF, CCFA, FFSA, ADEME, IFEN, EDF, FIU.
- the goal is the analysis of
  1. regular and local mobility,
  2. vehicle fleet and its uses,
  3. **long-distance mobility**.

ENTD 2008

In total

- 20’178 households and
- 44’958 individuals.

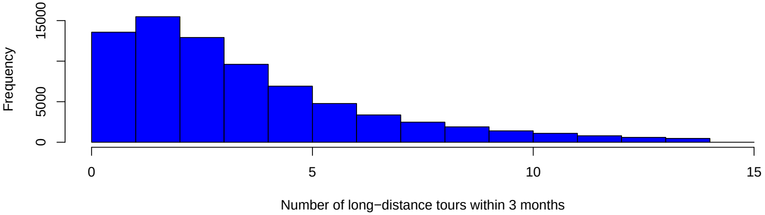
18’632 (representative) were chosen for LD questionnaire.

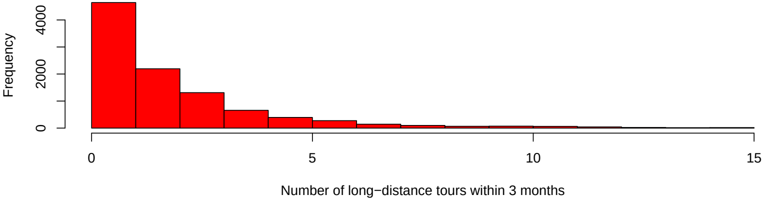
- 10’095 persons did a LD tour in previous 13 weeks.
- 5’670 persons did a LD tour in previous 4 weeks.
- 18’718 LD trips in 4 weeks form
- 8’505 LD tours, which were
    - 7’623 within France,
    - 6’978 in France and longer than 80km from home and

Results - Mobile Persons

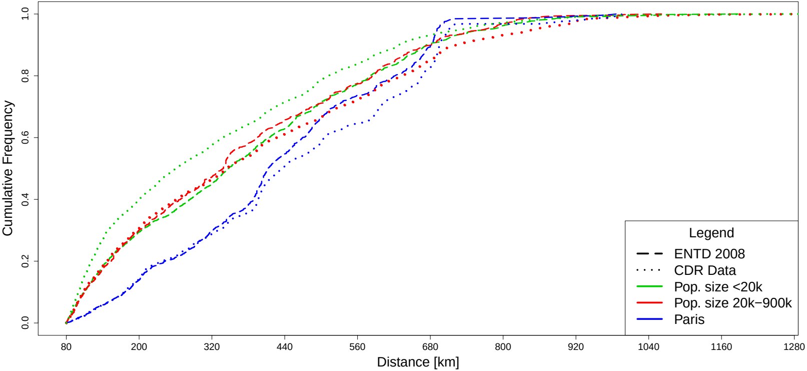
Data	Tracked Interval	Surveyed Persons	Mobile Persons	Mobile Share	Selected for analysis
CDR	30 days	1’388’941	814’381	58.6%	79’874
ENTD	28 days	18’632	4’796	25.7%	4’796
ENTD	91 days	18’632	8’743	46.9%	8’743

Results - Histogram: LD Tour Rates

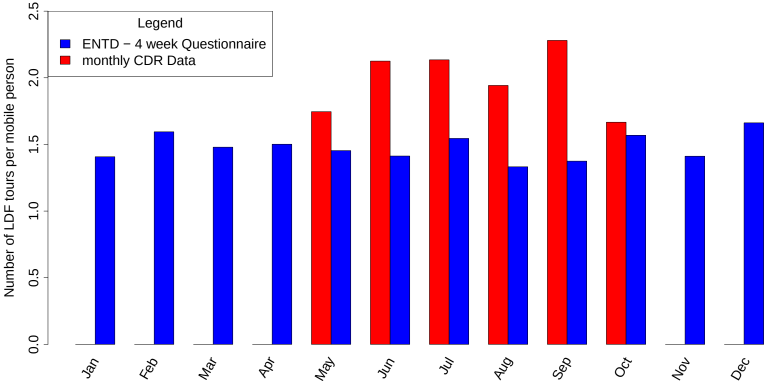




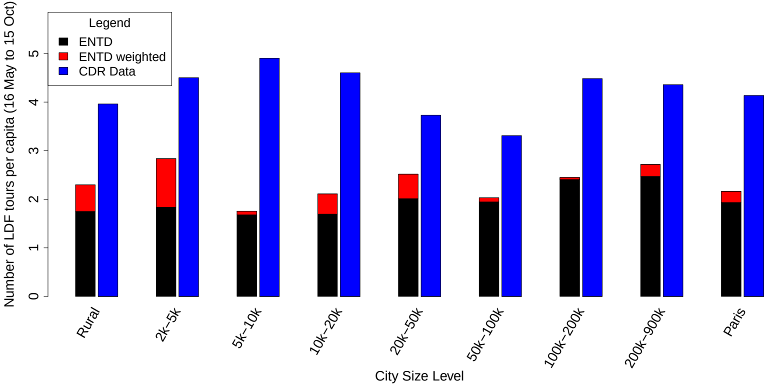
Results - Tour Distance Distribution



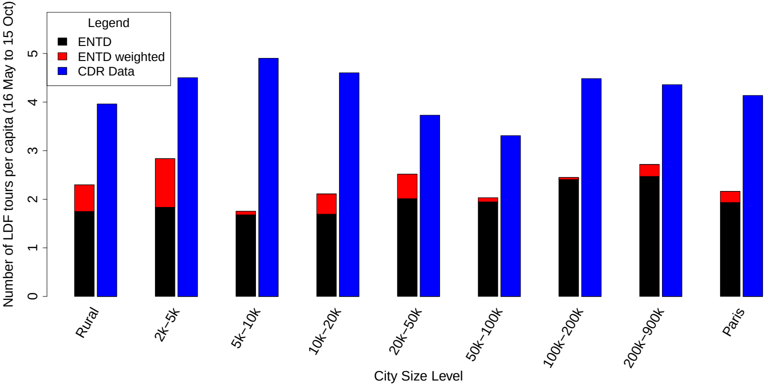
Results - Tour Frequency for Mobile Persons



Results - Tour Frequency per Capita



Results - Tour Frequency per Capita



Reference Interval	CDR data 5 months	ENTD 4 weeks	ENTD 13 weeks	ENTD weighted 1 year
Tours in 5 months per capita	4.27	2.25 (52.7%)	1.96 (45.9%)	2.36 (55.3%)

Discussion - Limitations

(Our) CDR data has limitations:

1. Selection of customers might be biased
   (frequent callers are more likely to be chosen)
2. Computation of home locations.
3. No Roaming/International tours
4. Spatial inaccuracy.
5. Frequency of CDR data points.

Discussion - Limitations

(Our) CDR data has limitations:

1. Selection of customers might be biased (frequent callers are more likely to be chosen) ⇒**small effect**
2. Computation of home locations. ⇒**small effect**
3. No Roaming/International tours
4. Spatial inaccuracy.
5. Frequency of CDR data points.

Discussion - Limitations

(Our) CDR data has limitations:

1. Selection of customers might be biased (frequent callers are more likely to be chosen) ⇒**small effect**
2. Computation of home locations. ⇒**small effect**
3. No Roaming/International tours ⇒ **we excluded international travel**
4. Spatial inaccuracy.
5. Frequency of CDR data points.

Discussion - Limitations

(Our) CDR data has limitations:

1. Selection of customers might be biased
   (frequent callers are more likely to be chosen) ⇒**small effect**
2. Computation of home locations. ⇒**small effect**
3. No Roaming/International tours
   ⇒ **we excluded international travel**
4. Spatial inaccuracy.
   ⇒ **Good enough for Long-Distance Travel**
5. Frequency of CDR data points.

Discussion - Limitations

(Our) CDR data has limitations:

1. Selection of customers might be biased
   (frequent callers are more likely to be chosen) ⇒**small effect**
2. Computation of home locations. ⇒**small effect**
3. No Roaming/International tours
   ⇒ **we excluded international travel**
4. Spatial inaccuracy.
   ⇒ **Good enough for Long-Distance Travel**
5. Frequency of CDR data points.
   ⇒ **The results provide a lower bound**

Conclusion

Main Result

Mobile phone data suggests that long-distance tour frequency is **twice as high** as in the National Travel Survey results

Conclusion

Main Result

Mobile phone data suggests that long-distance tour frequency is **twice as high** as in the National Travel Survey results

Result is a lower bound

1. Low CDR frequency.
2. Assumption that people that are not mobile in June are not mobile at all.

Conclusion

Main Result

Mobile phone data suggests that long-distance tour frequency is **twice as high** as in the National Travel Survey results

Result is a lower bound

1. Low CDR frequency.
2. Assumption that people that are not mobile in June are not mobile at all.

Conclusion

There is a big need of alternative data collection methods!

Thank You!