OECD transport statistics meeting

1 – 2 April 2015, Paris

Item 4.2

Main outcomes of the feasibility study on the use of mobile positioning data for tourism statistics

Christophe Demunter
EUROSTAT – Unit G-3 "Short-term business statistics and tourism" & TF Big Data

eurostat

Aim of the project

- ⇨ Explore the possibilities and limits of using mobile positioning data stored by network operators for measuring tourism flows
- ⇨ Explore the possibilities and limits of using big data for official statistics
- ⇨ Getting answers to the many questions raised by big data "doubters"/"non-believers" (but also by "believers")

"What about those who don't use mobile phones?"

"I live near the border and sometimes connect to a foreign network!"

"Tourists buy foreign SIM cards when travelling, don't they?"

Origins of the project

- ⇨ Changing geo-political environment
- ⇨ Quickly evolving technology and large-scale adoption of tools/devices
- ⇨ Changing working environment of official statisticians

- ⇘ New technologies, new techniques, new sources and a new 'Zeitgeist' boost and stimulate a paradigm shift in official statistics

Who carried out the feasibility study?

A multidisciplinary, international consortium (DE, EE, FR, FI)

- National statistical institutes
- Tourism reseachers
- Academics
- Data scientists

Activities and reports

- ⇨ Five main project tasks
- ⇨ All reports are on the Eurostat website
  - Stock-taking (31 cases relevant for official statistics)
  - Feasibility of access
  - Feasibility of use - methodological issues
  - Feasibility of use - coherence
  - Opportunities and benefits
  - ↳ Consolidated report (34 pages)

Barriers to access

#1 Protection of personal data

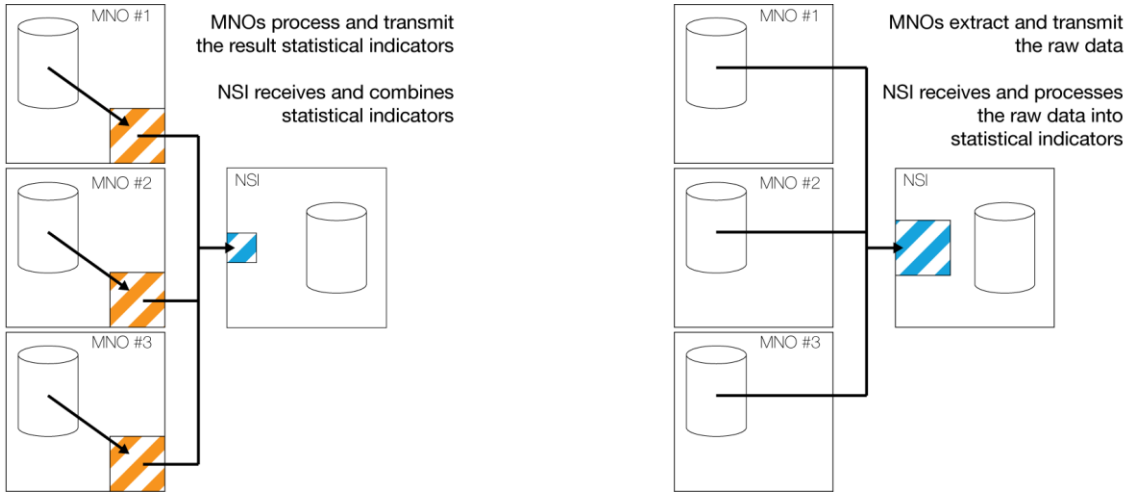
- Privacy issues linked to (legal) interpretation of concepts such as 'personal data', 'anonymised', etc.
- Fear of public opinion

☞ *Strong need for a less fuzzy legal environment at national & international level !*

Barriers to access

#2 Technological challenges

- Treatment of very large datasets
- Choice between a centralised or decentralised system



☞ *Complex but not impossible; not considered a hard barrier*

Barriers to access

#3 Financial and business related barriers

- Business secrets for Mobile Network Operators (MNOs)
- Public opinion & reputation
- Cost and burden for MNOs

☞ *Need for a mutually beneficial relationship to motivate or incite MNOs*

Improving access to mobile positioning data is
<u>the</u> main short term challenge
in order to pave the way for using of this source of big data!

Feasibility of use - methodological issues

#1 'Universal' issues

- Data collection and compilation related: sampling design, stratification, calibration, etc.

#2 Issues that are inherent to mobile phone data

- Representativeness (systematic / sampling bias?) of the technique, assessment compared to traditional techniques for data collection?
  - ✓ e.g. structural bias: increase in trips *or* only increase in use?
  - ✓ overcoverage & undercoverage (> 1 SIM card ; foreign SIM card)
- Applying tourism statistics scope and definitions?
  - ✓ exclude flows within the usual environment, longitudinal data, …

☞ *Not more significant than similar shortcomings of 'traditional' sources*

Feasibility of use - methodological issues

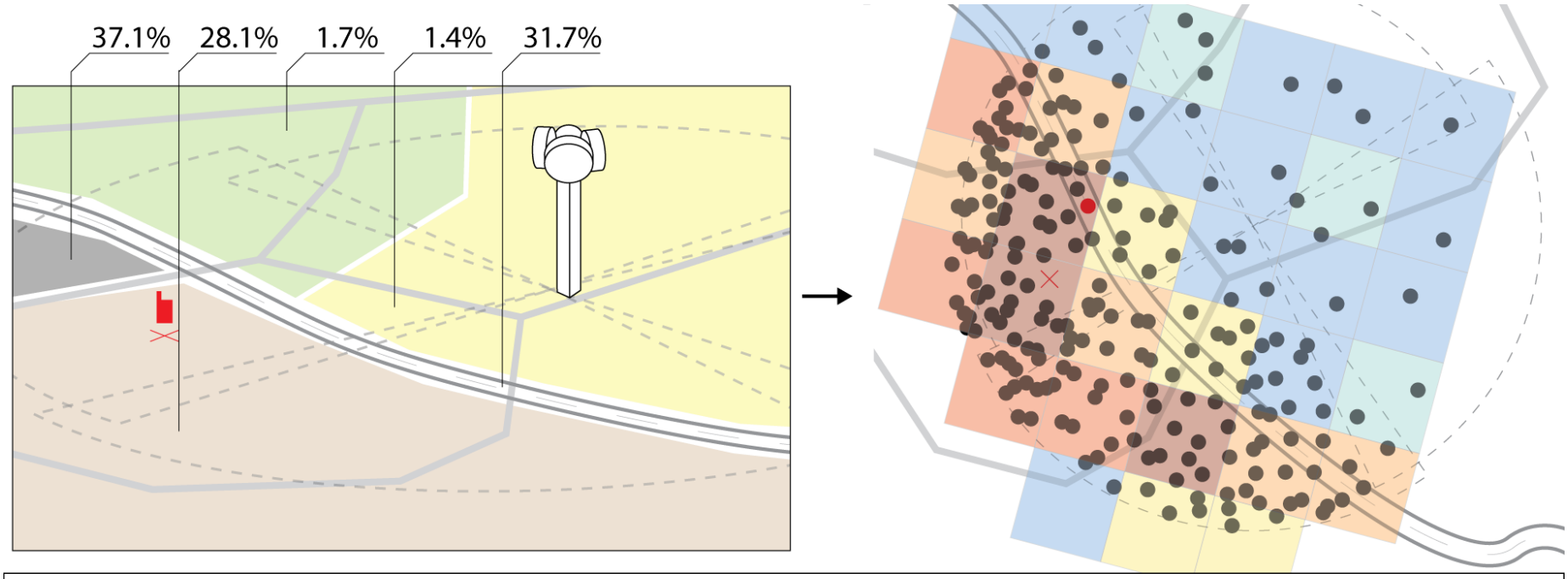
#3 Issues that are inherent to new technologies

- Continuity of data access
  - ✓ flexibility of changing the data requirements (e.g. new breakdown)
  - ✓ robustness of series if one or more MNOs drop out
  - ✓ contingency planning if *all* MNOs stop providing data
- Shifts in technology and consumer behaviour
  - ✓ new devices and their impact on the way people communicate
  - ✓ new services (e.g. relevance of Call Detail Records in 2020?)

↳ *bigger exposure to exogenous factors makes close monitoring and constant innovation essential conditions for using big data in official statistics*

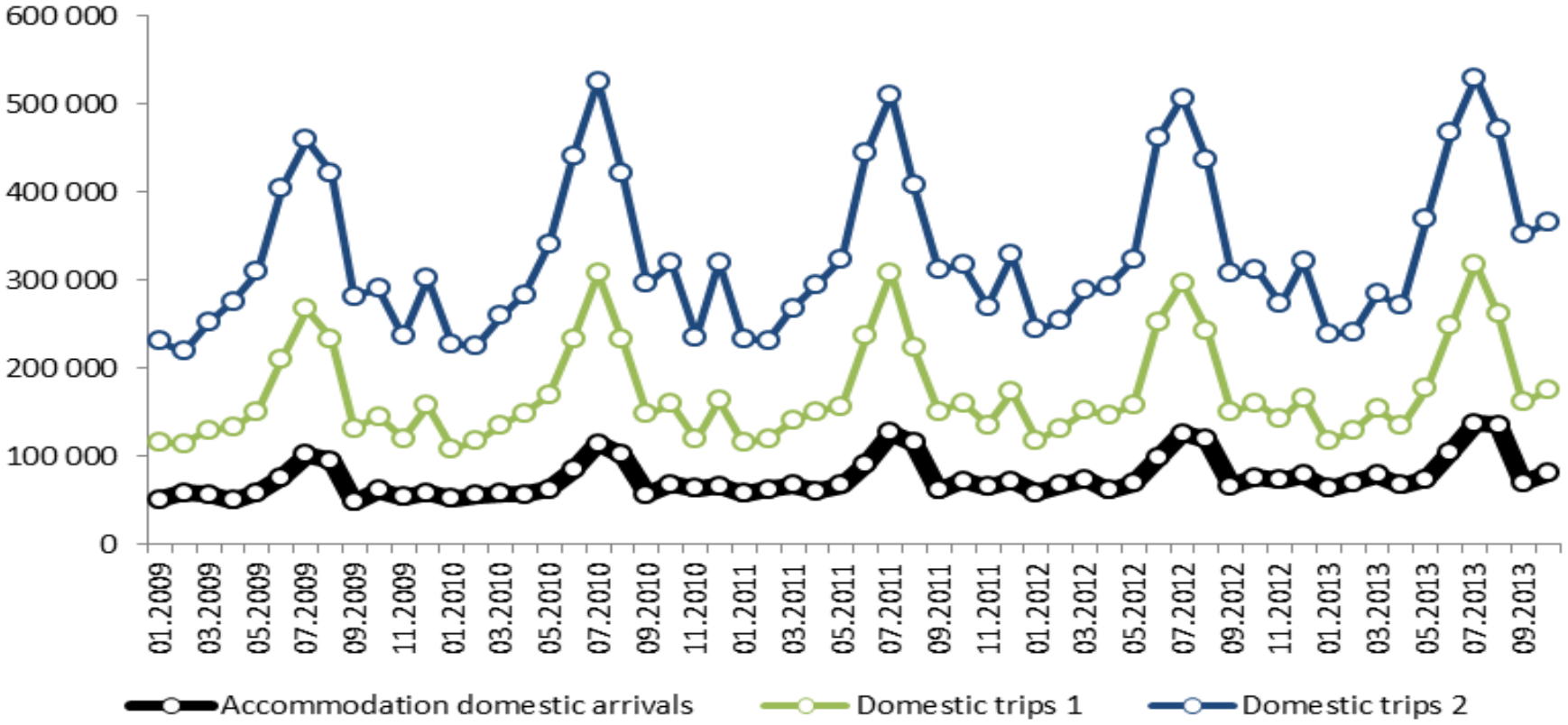
Feasibility of use - methodological issues

Example: probabilistic geographical distrubition to improve the positioning (as compared to the location of the antenna)



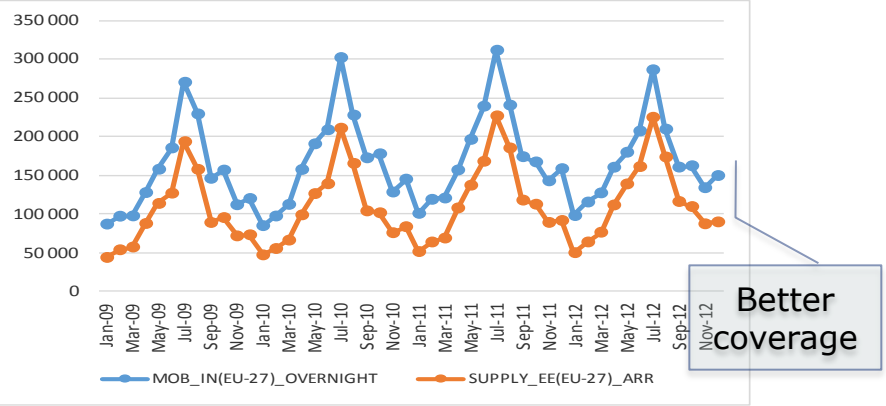
Feasibility of use - methodological issues

Example: effect of using different administrative borders on the delineation of the usual environment



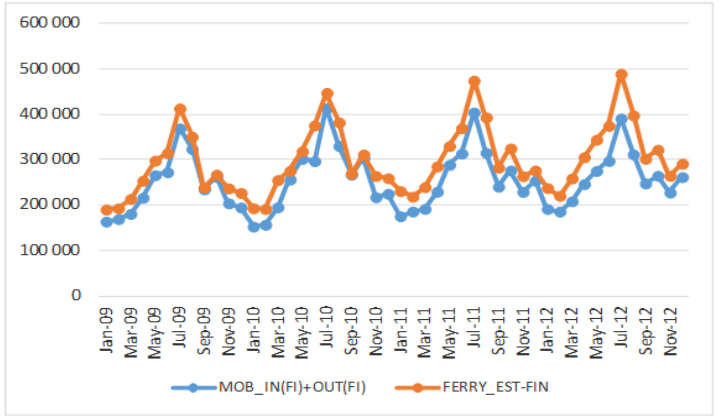
Feasibility of use - coherence

Inbound overnight trips (vs. accommodation statistics)

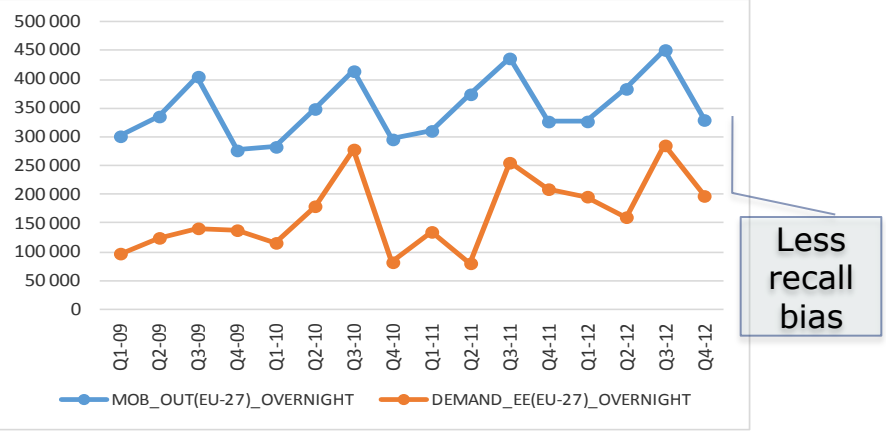


Better coverage

Inbound, outbound overnight trips (vs. ferry passengers data)

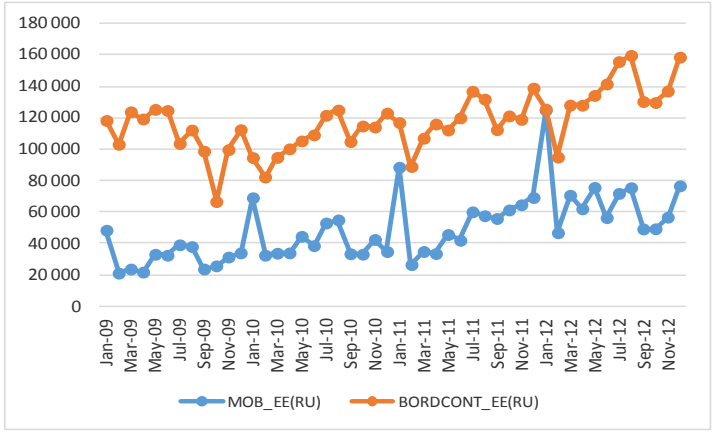


Outbound overnight trips (vs. household survey data)



Less recall bias

Inbound overnight trips (vs. border control data)



Strengths and weaknesses

⇨ Relevance

+ Completeness: better coverage, larger scope
+ New statistics, new indicators, breakdowns previously not available (e.g. finer granularity of space and time)
– Lack of socio-demographic variables and some domain-specific variables (purpose of trips, expenditure, …)

⇨ *explore mixed-mode solutions (e.g. large samples based on big data + smaller follow-up survey to collect domain-specific information)*

⇨ Timeliness

+ Increased integration and automation leads to better timeliness, up to near-real-time data (but impact on the cost!)

Strengths and weaknesses

⇨ Accuracy

+ Absence of non-response
+ Absence of memory effects or recall bias
– Some overcoverage and undercoverage issues
– Measurement error (# observations vs. precision of location/duration)

Strengths and weaknesses

⇨ Coherence and comparability

+ Good coherence with existing series
+ Synergies with related domains

  (BoP travel, transport and urban mobility, population, etc.)

  ⇨ *join forces, across domains and internationally*

+ Use of joint algorithms leads to better comparability across domains (and over time)
+ Additional calibration source for 'traditional' data

Strengths and weaknesses

⇨ Cost and burden

+ Elimination of direct respondent burden
+ Elimination of traditional data entry (important error source!)
+ Possibly more cost-efficient than traditional surveys
– Piloting and implementation cost (start up), regular production cost
– Possibly parallel processes (big data / traditional data) in a first phase
– New skills needed
– Dependency on external data providers (in casu MNOs)

Strengths and weaknesses : conclusions

At present, mobile positioning data cannot replace current statistics but can give complementary and/or faster results

However… official statisticians have to think out of the box and leave their comfort zone

- The existing scope and definitions are – besides user needs – based on the available sources and methodologies at the time of development
  → Do not repeat but do better !
- Use of big data necessitates a ***revolution*** of the mindset rather than a simple ***evolution*** !
  → Re-thinking indicators, zero-base user need analysis instead of incremental changes in the existing frame

Et maintenant?

- ⇨ Multi-country and multi-domain project in the pipeline
  - Given that getting access is a critical factor, the number of domains analysed and assessed should be maximised: e.g. population, balance of payments (travel), **transport and urban mobility**, tourism
  - Involve several countries, possibly two-speed approach
- ⇨ Use of data stored by Mobile Network Operators
  - Call detail records and data detail records
- ⇨ Expected output
  - Partnerships with MNOs
  - Studying data structures and defining data access standards
  - Testing data compilation and assessing quality

Thanks for your attention

christophe.demunter@ec.europa.eu