Quality implications of the use of big data in tourism statistics: three exploratory examples

F. Cortina García[1], M. Izquierdo Valverde[2], J. Prado Mascuñano[3], M. Velasco Gimeno[4]

[1] *National Statistical Institute (INE), Madrid, Spain; fernando.cortina.garcia@ine.es*
[2] *National Statistical Institute (INE), Madrid, Spain; maria.izquierdo.valverde@ine.es*
[3] *National Statistical Institute (INE), Madrid, Spain; jesus.prado.mascunano@ine.es*
[4] *National Statistical Institute (INE), Madrid, Spain; maria.velasco.gimeno@ine.es*

Abstract
Tourism statistics is one of the subject areas which are being considered at present in the ESS as a potential field for the development of big data use in order to improve the relevance, opportunity and punctuality of the products offered under the quality standards of official statistics.
In Spain, data from traffic loops and traffic control cameras are already being used in the estimation of inbound tourists. The paper presents three pilot studies about the use of big data and the integration of multiple sources: credit cards, mobile phones and web scraping (to collect prices of package tours and of its components).

Keywords: big data, traffic loops, mobile phones, credit cards, package tours, net valuation, web scraping

1. Introduction

Tourism is one of the most dynamic industries in many economies. According World Tourism Organization (WTO)[1], it represents 10% of the world GDP, it generates one in eleven jobs and international tourist arrivals have increased 4% to 1.2 billion in 2015. Estimation of tourism flows come mainly from border crossing and accommodation statistics, and from household surveys to resident population. The interregional component of this phenomenon makes comparability an essential feature of the reliability of data, which has been developed within the frame of WTO and Eurostat manuals, guides and recommendations. Thus, these sources of primary information, which have been providing data for many years, have a strong methodological base and fulfil high quality standards.

[1] Source: International Tourism Arrivals infographics, WTO (http://media.unwto.org/content/infographics)

However, in a world where mobility has increased to its highest levels in few years and border controls have disappeared in neighbouring areas, such as the Schengen Space in Europe, border crossing surveys are becoming more costly and difficult to conduct, and many countries are looking for alternatives and complementary information.

In this context, data generated not from purely statistical sources but from events intimately linked to the tourism phenomenon appear as a source of information that can improve the relevance, opportunity and punctuality of the products offered under the quality standards of official statistics. Examples of these new data sources are registers from traffic loops and traffic control cameras capturing flows of vehicles, records of mobile phones travelling from one place to another, activity of credit cards during a trip, among others.

Eurostat identified the potential of this sources for tourism statistics and launched in 2012 a project on the use of mobile data. Access to this information was identified as one important barrier to make its use feasible (Eurostat, 2014). Since then Task Forces on Big Data have been launched both at European and at national level to coordinate different projects and initiatives. Spain participates in an ESSnet pilot project on the use of mobile positioning data for official statistics whose first objective is obtaining access to the data. In the meanwhile, some preliminary analysis have been carried out within the Spanish System of Tourism Statistics, which will be presented in the next sections.

2. Traffic loops and traffic control cameras

The first experience of INE using big data in tourism statistic is related to the task of building the frame of people crossing the borders by road. Due to Schengen Treat, there is no control over people that cross the border from France or Portugal to Spain (and vice versa).

The register of traffic loops provides the total number of vehicles that cross the border for each crossing-road, in both ways (going in and out of Spain) by hour and classifying the vehicles according their length (short, medium and large). This information is completed by the traffic control cameras that are installed in the border lines (both registers are managed by Traffic General Direction). It is a complete database of number plates of vehicles that come into our country. Combining both sets of big data

we can estimate the number of foreign vehicles that enter in Spain monthly broken down by vehicle nationality.

The next step to know the number of persons that come into our country is transforming the *Vehicles Frame* in a *Travelers Frame*. To get this aim, sample data of vehicles by type of vehicle, nationality (of number plate) and number of occupants per vehicle are collected. Using the collected information, an occupancy rate of vehicles, by type of vehicle and nationality is calculated. Mixing both data the *People Frame* mentioned before is calculated.

This is the general schema that is carried out to get this basic information to estimate the number of foreign visitors (tourist and same-day visitors) that come to Spain every month by road. In this case we don't have to face problems related to different definitions used in the register of traffic loops and traffic control cameras, vehicles crossing border is the counted unit in both registers. But the coverage of these sets of data sometimes is not exactly the same, due to technical problems that are being solved.

Tracking an anonymized number plate through the camera registers database will allow new studies about same-day visitors.

3. Mobile positioning data

Mobile phones connect to cell towers with a defined geographical coverage. Mobile phones connected to the network generate events that are recorded in a database associated to the cell phone ID. Tracking an ID in the events database gives information about mobility of the cell phone.

These events can be classified in two categories: active and passive:

- Active events: those generated when the subscriber makes or receives a phone call, sends or receives a text message, or when he switches on or of the device.

- Passive events: those generated when the telephone is not active. Location of inactive cell phones is known through `location updates' sent by the network. Passive events can be generated randomly (when the telephone changes from one LAC – group of cell towers controlled by the same base controller- to other) or periodically, every four hours.

The combination of both kind of events is especially relevant in the case of international tourists, because it allows to analyse a much wider population. At the moment of this project, the system employed to obtain positioning data used both active and passive events generated in the layers 2G and 3G.

For the first approximation to the use of mobile phones positioning data, an ad-hoc extraction from the events database of one of the most important MNO operators in Spain was defined for analysis. The objective was measuring the number of tourists both residents and non-residents and their average stay, broken down by region of destination (NUTS 2) and region/country of origin. We compared the results obtained with those derived from official statistics.

In the case of residents, data were provided only for august 2014. For non-residents, data for august 2013 are also available, making possible comparisons over time.

The first step was to identify tourists within the whole events database. To define them, the international accepted definitions were adapted to the possibilities of the database.

Tourism is defined in the regulation 692/2011 on European tourism statistics as the activity of visitors taking a trip to a main destination outside their usual environment, for less than a year, for any main purpose, other than to be employed by a resident entity in the place visited. Usual environment is the geographical area, not necessarily a contiguous one, within which an individual conducts his regular life routines.

Tourism includes trips with overnights stays and same-day visits. This study focuses only in trips with overnight stays.

An overnight stay is defined in this project as a stay of more than 24 hours in a region (NUTS 2) of destination. In the case of residents, if the region of destination is also the one of origin, the stay must take place in a municipality (LAU 2) different of the one of residence.

In practice, the standard definition has been adapted to consider as a tourist every mobile phone staying at least two consecutive days in a region of destination, the stay comprising an eight-hour period between 22:00 hours of the day of arrival and 08:00 of the next day.

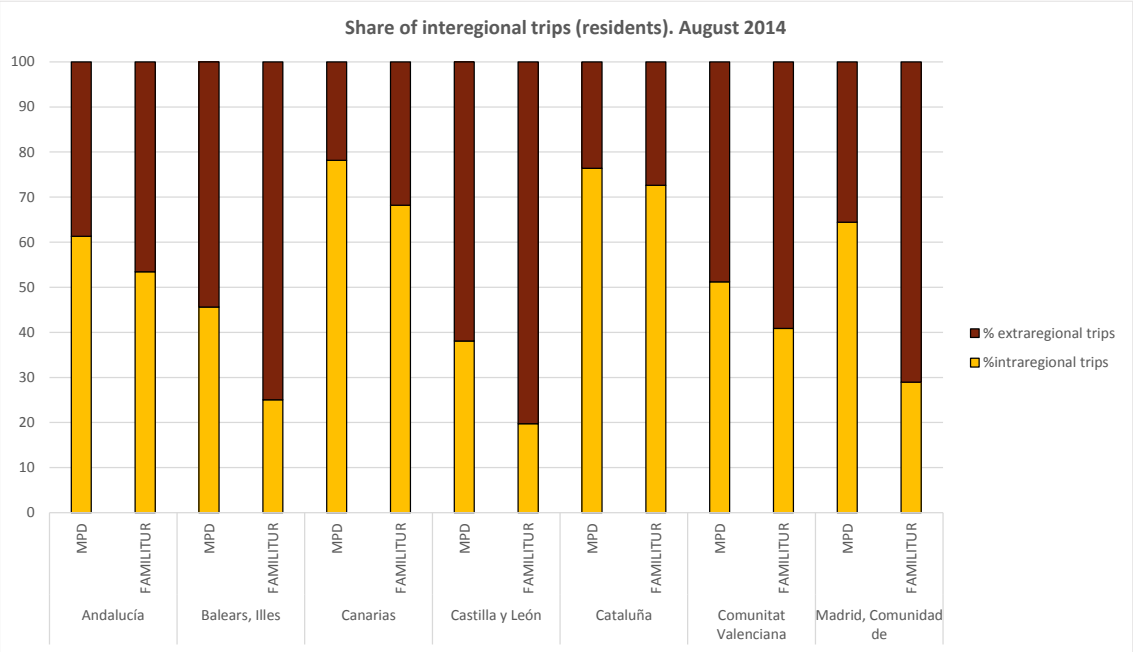
For resident mobile phones, residence is assigned empirically, taking into account the different places where the cell phone has made an overnight stay (between 0:00 and 8:00) in the last six months.

In the case of residents, when we compare mobile positioning data (MPD) with survey data (FAMILITUR), Table 1 shows a quite similar distribution of trips among the regions of destination. Main differences are found in Madrid, where the percentage of tourists identified from MPD is three times bigger than in the survey. Being Madrid a city with a big metropolitan area, this could be indicative of the need of a more accurate definition of usual environment. In fact, taking into account the residence of the trip, Figure 1 below shows that data from MPD present a higher proportion of intraregional trips than the survey in all the regions represented, Madrid with the highest difference.

Table 1. Distribution of trips by destination (residents) August 2014

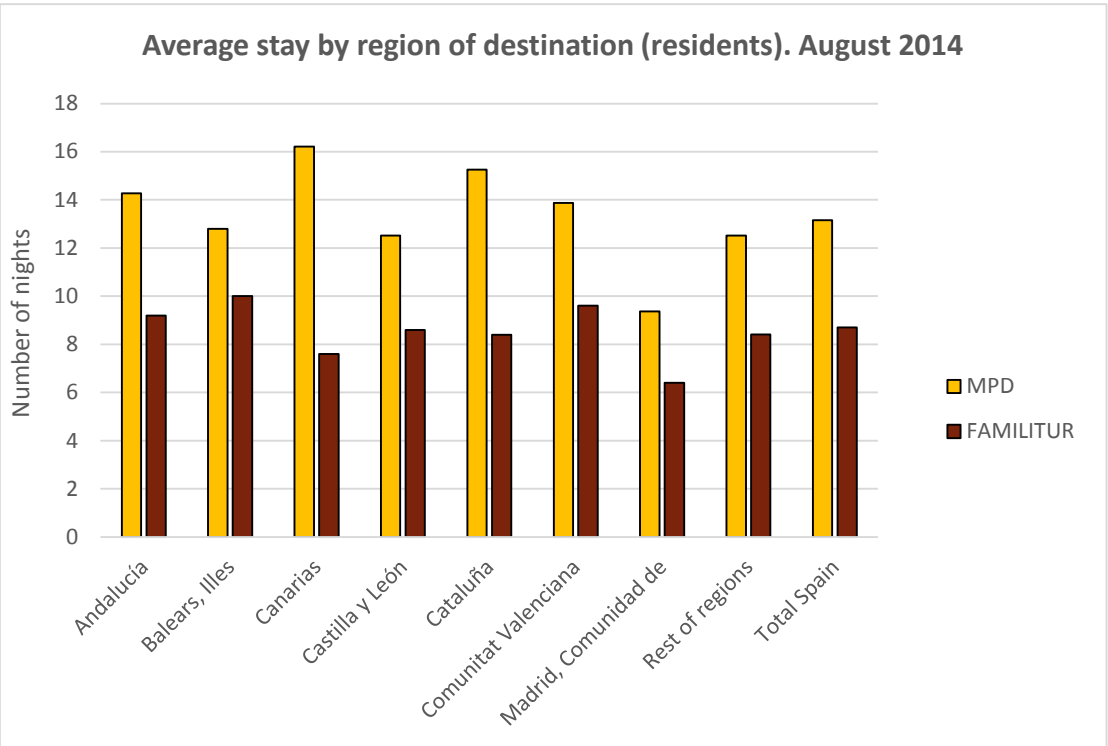
	% trips MPD	% trips FAMILITUR
TOTAL	100,0	100,0
Andalucía	15,8	19,1
Illes Balears	2,8	2,2
Canarias	3,1	3,6
Castilla y León	8,7	10,1
Cataluña	14,8	13,4
Madrid, Comunidad de	10,7	3,4
Comunitat Valenciana	12,7	15,6
Rest of Regions	31,5	32,6

Figure 1. Share of interregional trips by destination (residents). August 2014



Big differences are found for the variable average stay (Fig. 2). In aggregate terms, mobile data show an average of 13.5 nights for residents' trips to a destination in Spain, while survey data estimate is 8.7 nights. Once again, the definition of tourist and usual environment seem to be the cause of this discrepancies.

Figure 2. Average stay by region of destination (residents). August 2014



Comparing MPD for non-resident mobiles with survey data (FRONTUR-EGATUR), the distributions of tourist by country of origin present slight differences (Table 2). The most important countries are United Kingdom, France and Germany. In both cases Germany gets the third position, but UK and France exchange their ranking.

Table 2. Distribution of international tourist by country of origin August 2014

	% tourist MPD	% tourist FRONTUR
TOTAL	**100,00**	**100,00**
Germany	16,24	14,22
France	18,52	21,88
Holland	8,69	3,69
Italy	6,34	7,18
Portugal	4,24	3,41
United Kingdom	17,09	23,52
USA	1,29	1,62
Rest of the w orld	27,59	24,48

Analysing the average stay the differences are significant (Table 3), always much higher the estimation of the survey, just the opposite situation that the resident analysis. For non-residents, usual environment is not expected to be a general problem, although border areas and residents with would require a separate analysis. This low averages from MPD could be explained by the fact that different legs of the same trip for the survey are considered as different trips in MPD.

Table 3. Average stay by country of origin August 2014

	MPD	EGATUR
TOTAL	**5,8**	**9,8**
Germany	7,0	10,2
France	4,9	9,4
Holland	5,6	11,3
Italy	6,1	8,1
Portugal	4,4	6,3
United Kingdom	6,2	9,8
USA	5,3	11,7

4. Data from credit cards

Another source of information being explored is data recorded by the electronic payment system of one of the most important banks in Spain. In the case of residents, we analyse registers of all payments made by the bank clients in every point of sales terminal (POS) and ATM extractions with an entity card. Only cash payments and those made with a card of any other entity are out of scope of the study. For non-residents, available information comes from payments or extractions in POSs or ATMs in the BBVA network, so we have a more partial vision of their activity in Spain.

As in the previous case, aggregated results with a high level of detail were provided following INE's indications to obtain the information. Direct work with the database was carried out by the bank.

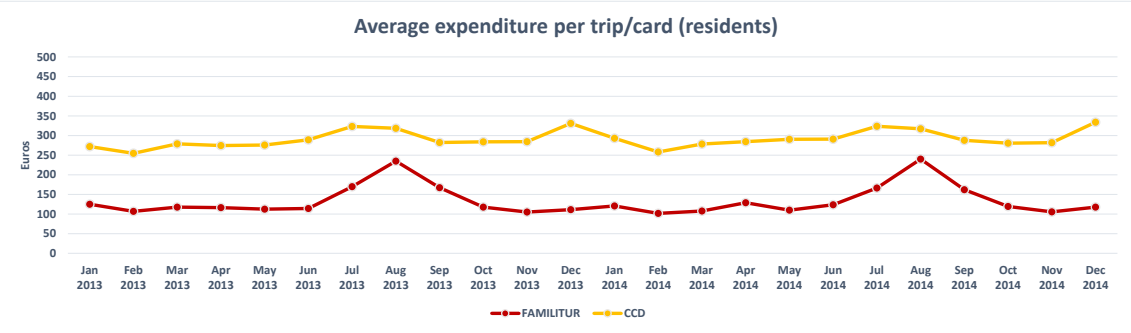
Residence of the card holder is available based on the information provided by the client. POS' are geolocalised. In the study, every payment in a municipality different from the one of declared residence of the card holder has been considered as tourism expenditure. Besides, POS are associated with an economic activity so that expenditures can be broken down in different categories: travel agencies, food and beverages, accommodation, shopping, recreation, transport, cash withdrawals and other expenditures.

Monthly data are available since January 2013 to December 2014 both for residents and non-residents.

In the case of residents, when comparing average expenditures per trip, Figure 3 shows higher values for the credit cards' data (CCD) series. Coverage of expenditure is not the same in both sources: official statistics measure expenditures made during the trip and those made for it before it takes place, while CCD should reflect only expenditures made in destination, thus, during the trip. Besides, CCD do not include neither cash payments nor those made with other cards. Consequently, CCD average expenditures were expected to be lower than survey results. Such a difference must be due to methodological causes. Further analysis by type of expenditure might provide a clue for this discrepancies and also an empirical determination of the place of residence of the card holder should provide better results.

Another aspect we observe in Figure 3 is that seasonality seems to be softer in the CCD series. One reason underlying this different pattern could be the fact that in official statistics a trip is assigned to the month of finalization of the trip, and the expenditures made during or for the trip as well, while credit card registers may be assigned to the real date in which the payment takes place.

Figure 3. Average expenditure per trip/card (residents). August 2014



5. Net valuation of package tours: quality limitations of multisource methods and big data approach

One of the specific aspects of the Tourism Satellite Account is the treatment of package tours. Unlike the central framework of the National Accounts, the package tour should be unpacked and should not be treated as a product itself but as the sum of its components. Each of the components of a package tour, including the value of the service offered by the tour operator and travel agency, is considered to be purchased directly by visitors.

The estimation of each of the components of the package tours is complex for several reasons:

- Information on the costs or prices is generally very sensitive and difficult to provide by informants.

- In the case of inbound tourism, tour operators are usually non-resident companies from which is very difficult to obtain information (they are not required by their national laws to respond to questionnaires from foreign institutions).

- In general, tour operators negotiate a set price with suppliers for various products in often very difficult to differentiate the price of each; for example hotels they are paid a cost that includes accommodation, breakfast, sometimes internet, etc.

- The estimation of the percentage of the tourist package that belongs to the home economy and the percentage to be counted in the destination economy.

Considering all of the above limitations, the National Statistics Institute of Spain is developing a pilot exercise through techniques of "web scrapping" to get price information for the products offered by tour operators through the web, whether individually (transport, accommodation, etc.) or together (tourist packages) in order to obtain a cost structure of the components of the package, and to improve the quality standards of the estimations. To do so a couple of tour operators operating on the network have been selected and their travel offer to one or more destinations (Canary Islands) will be analyzed for a fixed period (one week) and other similar characteristics. Hopefully the enormous possibilities offered by the Internet and the use of big data provide information to the so-called "unbundling" of tourist packages.

6. Conclusions

The new sources of information are really promising as they can provide accurate and punctual information about the tourism phenomenon, allowing a more detailed geographical analysis than those permitted nowadays by official statistics.

Nevertheless, important and coordinated efforts have to be made by statistical authorities and data providers to obtain results with the quality standards actually achieved by current official statistics.

The examples presented in this paper try to show that an in - depth conceptualization exercise should be made in first place to identify the phenomenon to be measured and second to assure comparability over time and between countries. For testing the more adequate definitions and parameters, assessing impacts of changes and monitoring the consistency of the decisions finally adopted, official statisticians need to have big control of the original databases, how are they processed, every assumption made, etc. with the highest level of detail. Of course, they must be aware of any incidence occurred

in the systems and its possible implications. In summary, if direct access to the databases is not possible, detailed metadata should be delivered with the information requested and fluent communication with data providers is essential during the whole process, but especially in this initial phases.

7. References

EUROSTAT, Ahas R., Armoogum J., Esko S., Ilves M., Karus E., Madre JL., Nurmi O., Potier F., Schmücker D., Sonntag U., Tiru M. (2014), Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics, Consolidated Report, Publications Office of the European Union.

Regulation (EU) No 692/2011 of the European Parliament and of the Council of 6 July2011 concerning European statistics on tourism.

United Nations, World Tourism Organization, Eurostat, OECD (2010), Tourism Satellite Account: Recommended Methodological Framework 2008, United Nations publication.