Sixth UNWTO International Conference on Tourism Statistics
MEASURING SUSTAINABLE TOURISM
Manila, Philippines, 21 – 24 June 2017

Tourism Statistics: Early Adopters Of Big Data?

Session 5. Producing Data on Sustainable Tourism

Christophe Demunter
Tourism Statistics Section
European Commission (EUROSTAT)

Abstract

The ubiquity of data revolutionises the world of official statistics. Citizens and enterprises leave a constant flow of digital footprints, voluntary or unintended. This data deluge can make it difficult to see the forest for the trees, but big data undeniably has a huge potential for many areas of statistics.

The arrival of big data also changes the statisticians' working environment. They no longer hold a monopoly to producing statistics but now compete with a wide range of data producers. Ignoring innovation will push statistical authorities out of the information market, an evolution that can jeopardise the critical role of independent, official statistics in any democratic debate.

Many sources of big data measure flows or transactions. Within the wide range of statistical domains, tourism statistics is in the frontline of big data related innovations of sources and methods. Indeed, tourism statistics tries to capture physical flows of people – as well as the accompanying monetary flows – and big data provides promising new sources of data and previously unavailable indicators to measure these flows (and stocks).

This paper gives an overview of the different sources of big data and their potential relevance to compile tourism statistics. The discussion includes the opportunities and risks that the use of new sources can create: new or faster data with better geographical granularity, synergies with other areas of statistics sharing the same sources, cost-efficiency, trust of users, partnerships with those organisations holding the data, access to personal data, continuity of access and output, quality control and independence, selectivity bias, alignment with existing concepts and definitions, need for new skills, etc.

The global dimension of big data and the transnational nature of companies or networks holding the data necessitate a discussion in an international context, even if legal and ethical issues often have a strongly local component.

Keywords: big data, tourism statistics, innovation

Table of contents

1 Big data and the seven V's

There are probably more articles attempting to define big data than articles actually applying big data sources and techniques. A definition of big data is out of scope of this paper and perhaps not even relevant as it is an ever-changing concept. To describe or delineate big data, authors have made reference to the 3 V's of big data: volume, variety and velocity (see for instance Laney (2001) and Beyer & Laney (2012)).

Briefly, the ***volume*** refers to the exploding quantity of data in terms of observations – in orders of magnitude of gigabytes, terabytes, petabytes (and soon exabytes or zettabytes?) - as well as variables observed; the ***velocity*** refers to how quickly these data are generated and their resolution in time. Several infographics depicting what allegedly happens in an internet minute or second circulate on the internet: every second 2.5 million e-mails are sent, over 50 000 Facebook updates are posted, over 60 000 Google searches are entered and 15 000 USD is spent online. By the time you read this, the above figures will for sure be outdated. However, the internet is not the only data source, mobile network operators store each second many gigabytes of information linked to the whereabouts and service usage of their subscribers, supermarkets gather a constant flow of cashier data, etc.

The ***variety*** refers to the many different types of data, often of an organic nature and not primarily designed to compile statistics [Groves (2011)], such as natural language textual data (e.g. social media posts), photos (e.g. posted on Instagram or Facebook), website logs, videos (e.g. camera surveillance), recordings or geo-coded data.

Besides these there "core" V's of big data, important other V's entered the debate in recent years: veracity, validity, volatility and value.

The **veracity** and **validity** touch upon the quality, the reliability and the usefulness of big data. The sheer volume of data and observations does not guarantee quality. On the contrary, the unwanted bias and noise in most big data sources are without any doubt among the more complicated challenges for statisticians (whereas volume and velocity are somehow under control thanks to technological innovations in IT infrastructure and storage capacity). Traditional quality measures and solutions applied to census data, sample survey data or panel data cannot easily be transposed to the new data sources characterised by non-probabilistic samples and unknown inclusion probabilities. Coverage of big data sources (e.g. not everyone has a mobile phone) or self-selection error (e.g. not everyone uses a mobile phone equally intensive) requires new methodological approaches to ensure quality and to ensure trust of data users and data producers. For a study with an overview of methods to address selectivity bias in big data sources, see European Commission (2017).

The **volatility** refers to how long the data remains relevant and how long it should be kept, keeping in mind the billions of impulses registered every second or keeping in mind the legal framework regarding the retention of personal data. This aspect is of particular importance for official statistics where the focus is on continuous data series rather than ad-hoc studies or one-off exercises.

The **value** of big data is twofold: firstly its value for statisticians as a potentially richer or timelier data source; secondly its value for businesses and policy makers in an era of data-driven decisions. A special case is the businesses or organisations holding the data. Data is a valuable, marketable asset and can make these stakeholders reluctant to grant access to the data they hold.

2 Big data in official statistics in the European Union

The strategic importance of big data for the European Statistical System (ESS) was recognised by the ESS Committee[1] (ESSC) in adopting in September 2013 the *Scheveningen Memorandum* [ESSC (2013)], which calls for an action plan for big data and official statistics to be addressed jointly by the ESS. As a follow-up, Eurostat created an internal task force on big data (TF BD) and a task force at the level of the ESS. The latter task force combines members from national statistical authorities, UN organisations, other European Commission services and scientific advisors. It produced a *Big Data Action Plan and Roadmap* [ESSC (2014)] for big data that was adopted by the ESSC in 2014.

The task force works on the implementation of the action plan. Eurostat launched several initiatives to explore the potential of big data and to identify its challenges. The ESSnet[2] Big Data was launched in March 2016, continuing till mid-2018. The basis of the ESSnet Big Data is a set of pilots run by national statistical institutes. Their purpose is to explore the potential of selected big data sources for the production of official statistics and the application of results to specific statistical domains. The ESSnet aims to generalise the findings of the pilots in in terms of methodology, quality and IT infrastructure for the future use of the selected big data sources from the pilots within the European Statistical System. These source-oriented pilot projects include webscraping (for enterprise characteristics and job vacancies), smart meters, mobile phone data and AIS data (automated tracking of ships).

Eurostat also launched a study on ethics, communication, legal environment and skills (expected at the of end 2017). Since 2016, the European Statistical Training Programme (ESTP) includes five dedicated 3 to 4 days' courses related to big data: introduction to big data and its tools, hands-on immersion on big data tools, big data sources (web, social media and text analytics), automated collection of online prices, advanced big data sources (mobile phone and other sensors).

Finally, Eurostat launched a series of in-house big data pilots with the purpose to build internal technical expertise and infer from its own experience the implications at strategic level for official statistics in general, for the ESS and for Eurostat and the European Commission.

The underlying objective of the above mentioned activities is to pave the way for introducing big data sources in the regular production process of official statistics.

Given the borderless nature of these new data sources - available on the worldwide web or held by organisations across the globe using somewhat comparable data structure definitions - international cooperation in statistics has never been more crucial, from developing strategies to access and handle data to troubleshooting methodological issues and disseminating trusted statistics addressing the needs of the 21st century users.

3 The many faces of big data: sources with potential for measuring tourism

The variety of big data was mentioned in the first chapter of this paper. Many sources measure human activity or

[1] European Statistical System Committee, composed of high-level representatives of Member States' national statistical institutes. For more information, see the ESS website.

[2] An ESSnet project consists of a network of several ESS organisations aiming to provide results that will be beneficial to the whole ESS.

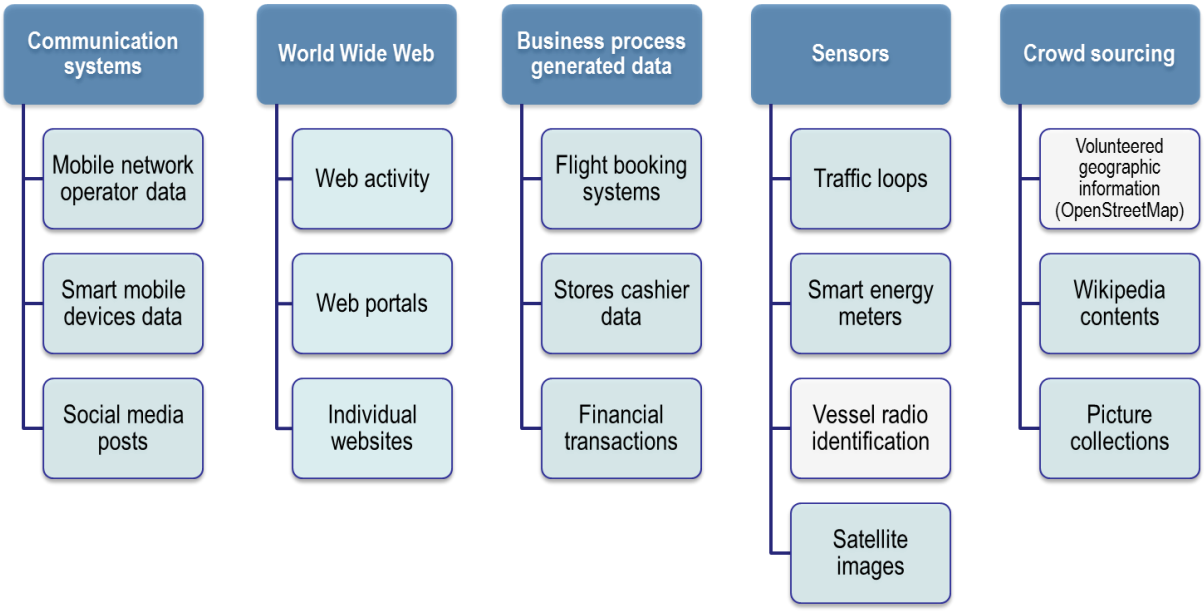
mobility, in other words flows of persons or transactions they make. Knowing that primary tourism statistics measures physical flows (and the corresponding monetary flows) of citizens, it doesn't come as a surprise that tourism statistics has been in the frontline of big data related innovations of statistical sources and methods. In this respect, tourism statisticians can help shape the future of information management, can benefit from experiments and share their experiences can become forerunners in the rethinking of statistical systems beyond the traditional methods based on sample surveys addressed to households or businesses.

The diagram in Figure 1 sketches an overview of the most commonly discussed sources of big data. Just like any other classification, individual items can be allocated to different groups, depending on the angle of view. The same is true for this taxonomy as sources are interrelated and multifaceted. For instance social media posts can be filed under 'communication systems' as well as under 'world wide web', Wikipedia is web based but also crowd sourced.

Many of the sources listed are not new to (official) statistics: satellite images, scanner data or traffic loops have been used for a long time to feed geographic information systems, price statistics or transport statistics. The novelty is how to prepare the statistical systems for a large scale, generalised, integrated use of these new (and not-so-new) sources of information – notwithstanding many countries' experience with using administrative data.

This chapter briefly introduces the different sources with potential relevance for measuring tourism (highlighted in the scheme). While some sources are more promising or more widely used, others are included in the analysis only for reasons of completeness. A first glimpse at the scheme of classification in Figure 1 learns that links to tourism are omnipresent. Rather than collecting data, the tourism statisticians of the future will be connecting data from various sources into a modernised system of tourism information (see also Chapter 4 of this paper).

Figure 1: Taxonomy of big data sources



3.1 Communication systems

This group compromises the commonly used sources in big data experiments, making use of the digital footprint that people leave in their day-to-day communication, actively or passively. Within the mobile positioning data, a distinction is made between mobile network operator data and other data gathered via smart mobile devices. Furthermore this group includes social media posts.

Until now, mobile positioning data has been a focal source for big data research. Firstly, it exists in all countries (which doesn't mean it can be accessed or used in all countries…). Secondly, many promising studies or experiments are available. Thirdly, this source has a potential relevance to many different areas of statistics, enabling synergies.

3.1.1 Mobile network operator data

Data held by mobile network operators (MNOs) is perhaps the most commonly used big data source for measuring tourism flows. Many countries across the world have embarked on pilots and ad-hoc studies. The growing penetration of mobile phone use (approaching or exceeding 100%) and the dropping roaming rates in certain parts of the world (in particular the European Union) make the analysis of the whereabouts of mobile phone use a highly relevant source for analysing the presence and movements of tourists. Besides the information on presence and flows, derived information can also contribute to more precisely determining the usual environment, for instance by determining social networks on the basis of call history (who calls who).

Since the pioneering work of Ahas et al. (2008) in exploring the use of mobile phone data for statistics (in particular tourism statistics), nearly ten years ago, the constellation has tremendously changed. Up to now, experiments with using mobile phone data were largely limited to the use of call detail records (CDR) – basically administrative information gathered for billing purposes. A comprehensive overview of this source, and the methodological issues, opportunities and weaknesses was reported in the Eurostat *Feasibility study on the use of mobile positioning data for tourism statistics* (European Commission (2014a)).

On the one hand, changed behaviour of mobile phone users is more and more affecting the relevance of call detail records (alternative non-SIM based messaging services, alternative voice or video call systems), which necessitates auxiliary data to assess the selectivity bias of this source, and to correct/calibrate for this bias. See for instance the work carried out by the Italian statistical office in assessing the use pattern of mobile phones in a tourism context, showing that tourist will use their mobile phone on 90% of domestic trips but only on 71% of outbound trips (Dattilo et al. (2016)).

On the other hand, mobile network operators are shifting to the use of other data sources available within their network infrastructure, in particular signalling data. Such network probing systems offer a much better temporal granularity (and indirectly also a better geographical granularity since the increased number of observations will capture more changes in location at cell level). These systems capture all signalling events, billable and non-billable. The amount of useful signalling events is up to ten times higher as compared with CDRs (De Meersman et al. (2016)). In case of one Belgian mobile network operator Proximus (see Seynaeve & Demunter (2016)), network detects the position of a device minimum every three hours (unless the device is switched off). For devices with data 'on', this drops to approximately 1 hour. In practice, through usage of the phone for calls, messages or data, devices are observed with a much higher frequency. During daytime hours, 7 out of 10

devices are observed after one hour during a given timeframe; 1 out of 3 devices are detected within 15 minutes. The mix depends on the actual usage and on the technology (e.g. 4G devices are typically giving more location points than 2G devices).

Mobile network operator data is a scholar example of how one source can serve multiple statistical domains simultaneously. On the basis of MNO data, information can be derived on the present population and the usual place of residence (population statistics). Movements away from the place of residence are relevant for mobility statistics. And irregular, infrequent, further away movements can reveal information on trips made outside the usual environment; as such tourism statistics is interested in the noise that can be observed in the data. In this respect, a key challenge is the correct delineation of the usual environment on the basis of the data (and not on the basis of the subjective opinion of the survey respondent).

Notwithstanding the evident relevance of MNO data for tourism statistics, getting access to the data for research purposes or for producing statistics unfortunately remains the main barrier to a widespread use of this source.

3.1.2 Smart mobile devices data

A second group of mobile phone data comes from the geo-positioning data from the device, from its activity sensors or from installed apps and the information these apps record. As such, this group can be extended to include other smart mobile devices, for instance tablets.

The geo-positioning data and information from activity sensors stored on the device can include very relevant information for analysing mobility, in particular tourism movements. Pattern recognition can detect the usual environment of the user of the device and the movements outside this usual environment, i.e. tourism trips. This source can in particular be interesting in mixed-mode data collection where respondents are selected via traditional sampling design methods for social statistics but report part of the data semi-automated on the basis of the logs of the device (with additional information entries on e.g. purpose of the trip or expenditure made). In the European Union, experiments are ongoing to such mixed-mode surveying in time use surveys or household budget surveys – two domains related to tourism demand surveys.

In their article on the use of GPS-based surveys for travel demand analysis, Vij & Shankari (2015) conclude 'that passively collected GPS-based surveys may never entirely replace surveys that require active interaction with study participants'. Indeed, while the first have the potential to produce more accurate, more detailed information on the number of movements, frequency, distances, etc., the latter will remain essential to complement the analysis with information on mode of transport, purpose of trips, expenditure, etc. However, this observation holds for most of the big data sources discussed in this paper.

3.1.3 Social media posts

Intended or unintended, people leave their digital footprint when using social media. Posts can be an information source on people's movement and behaviour.

Although the relevance for measuring tourism flows is obvious, this source faces a number of important methodological barriers, in particular related to the selectivity bias: the inclusion probability or likelihood that an

individual or event will be observed is highly correlated with the intensity of activity (namely the frequency of posting on social media). This limits the usefulness to detecting short term trends rather than volume information or longitudinal trends.

Other challenges include the absence of socio-demographic information (although profiling exercises are ongoing, testing whether the socio-demographic status can be detected within the data) and the continuity of the data source. The latter is caused by the fact that players come and go: ten years ago MySpace could have made a good source, today no one can predict what the weight of Facebook will be five or ten years from now.

3.2 World Wide Web

The next paragraphs focus on the internet as a data source, notwithstanding the fact that certain sources mentioned elsewhere in this chapter could also be filed under the current 'world wide web' heading.

3.2.1 Web activity

This group includes the traces left behind through search engines (e.g. Google Trends data), webpages visited (e.g. Wikipedia page views) or the traffic of websites.

Web activity can give an indication of topics of interest. Searching information on tourism destinations or page views of Wikipedia articles related to destinations can have predictive power for estimating tourism flows. Obviously, interest shown via search queries does not always lead to a visit or a purchase, but a correlation can be assumed. Separating tourism-related web activity from other web activity is a challenge: not everyone using the search term "Paris, France" will be interested in actually travelling to Paris (but is perhaps looking for information on French politics). However, refined analysis (e.g. destination names in combination with search terms such as "hotel" or "metro") could increase the correlation with tourism visits. Even if this source may face difficulties to produce volume data or absolute numbers, it can be useful starting point for estimating breakdowns that are otherwise difficult to collect (e.g. tourism activities, by looking at the weight of searches for "cruise" or "golf" or "gastronomy").

'Wikipedia contents' is mentioned under the heading 'Crowd sourcing' in the taxonomy presented in Figure 1. However, a derived source, namely *page views* of Wikipedia articles can be a proxy of visits to a destination, measured through the traveller's web activity. Signorelli et al. (2016) evaluated the use of page views as a source for identifying factors that drive tourism and whether these data can predict tourism flows. In the course of 2017, Eurostat will launch a project exploring how Wikipedia page views of (tourist) places of interest can help to disaggregate annual or national level tourism statistics into more detailed (and user relevant) infra-annual or regional series. For this purpose, an inventory of places of interest is not sufficient but an indicator of intensity of visits (reflecting tourist presence) is essential – page views could possibly be a useful distribution key (depending on the importance of the time-lag between page view and actual visit – if correlated at all – and the impact of repeat visits to the same destination on the likelihood of looking up preparatory information on e.g. Wikipedia).

3.2.2 Web portals

Portals are characterised by structured data and an interface allowing to access and consult a (dynamic) database. Typical examples include Tripadvisor.com, Booking.com or Airbnb.com. In general, the data is obtained by running a multitude of queries (webscraping).

Researchers have used webscraping to analyse the collaborative economy (e.g. Inside Airbnb analysing the offer and occupancy of properties rented out via Airbnb) or to use listings of attractions on Tripadvisor (and feedback and satisfaction levels of those attractions) to better understand visitors' preferences and behaviour (see for instance Almeida de Oliveira & Abrantes Baracho Porto (2016)).

Scraping portals has produced some information on the offer (e.g. number of hotels or apartments) while obtaining information on the actual occupancy (nights spent) proves to be more complicated (see for instance Schmücker et al. (2016)). A practical problem is that websites can detect and block the bots scraping the information, which means this source can be of dubious reliability for a longer-term perspective (but nevertheless useful for ad-hoc analysis). In the context of official statistics, the use of bots may be problematic as it might not be acceptable that statistical offices 'go rogue' on the portals but should do it with either the agreement of the company operating the portal, or with legal backing (e.g. from the statistical legislation). On the other hand, direct agreements with the companies behind the portals may jeopardise the independence and objectivity of statistical offices, whereas unidentified bots can better avoid possible manipulation by the company with the intent to influence the statistics.

The European project on big data (ESSnet, see Chapter 1) includes scraping of job portals to produce job vacancy statistics.

3.2.3 Individual websites

Contrary to the web portals discussed above, websites are rather static than dynamic. Businesses or organisations create contents for their stakeholders (customers, fans, etc.). In this case, data is obtained by extracting the contents from the html source code and transforming, clustering this into meaningful information for further analysis.

Taking tourist accommodation as an example, websites can give information on the activity status of establishments and their location, on the number of rooms and bed places available and on standard prices.

3.3 Business process generated data

Many enterprises produce a constant flow of data through their regular business processes. The next paragraphs discuss how this data can be relevant for measuring tourism.

3.3.1 Flight booking systems

Air travel leaves a trace via the booking and reservation systems of airline companies or transaction processors

such as Amadeus.

While this source is by default incomplete (only covering air travel, and within air travel not including all airline carriers – e.g. low-cost companies tend to be underrepresented), the data can be usual for specific destinations (in particular islands that are largely visited by plane) or as auxiliary information for tourism demand surveys where more remote destinations are typically poorly covered via trips data coming from sample surveys.

3.3.2 Stores cashier data

Tourists leave a digital trace of their stay via purchases made in local retail stores. Seasonal fluctuations in turnover (or in types of products sold) can be a proxy for seasonality in tourism activity in the region or destination. New ways of measuring seasonality at a local, destination level are crucial to better understand the impact of tourism and the sustainability of tourism.

Furthermore, electronic payments in stores could serve as a source for estimating TSA (tourism satellite accounts) tourism ratios for the retail industries on the basis of the share of cards uses all year round at the point-of-sale versus cards used for a short period only). Information on the issuing bank can give auxiliary information for estimating the country of origin of tourists.

3.3.3 Financial transactions

Decades before the concept 'big data' was introduced, tourism statistics set their mind on using payment card data for measuring tourism and travel (in the Balance of Payments terminology). With many systems of tourism statistics – and big data sources outlined elsewhere in this chapter - focusing on physical flows (accommodation statistics, border counts), payment card data is the missing link to monetary information.

The quality of the data improved over recent years, e.g. the possibility to distinguish between foreign purchases via e-commerce versus point-of-sale transactions, the possibility to detect the economic activity according to ISIC on the basis of the merchant code). However, despite the huge potential that this source can offer, few experiments take place (due to the sensitivity of the data). Burson & Ellis (2014) developed a methodology to use electronic card transaction data for measuring and monitoring regional tourism in New Zealand.

Notwithstanding the applications this source can offer, some "built-in" issues can limit the usefulness. In a society which is not cashless (in general all countries…), an increase in the volume and value of transactions observed does not necessarily translate into increased tourism figures because it can be caused only due to a substitution effect (new adopters of card payments, as replacement for cash payments). In order to produce meaningful estimates for absolute values, auxiliary information on how people use cards is desirable (see also the discussion on mobile phone data in paragraph 3.1.1)

3.4 Sensors

Sensors monitor movement of persons, land use, consumption of commodities or resources, etc. Many of these systems can, as a by-product, give relevant information for measuring (sustainable) tourism.

3.4.1 Traffic loops

Traffic counting is not new and has been used for many years in tourism statistics. In the past, traffic counting was rather quick and dirty in the context of border surveys, but automation opens new perspectives.

Statistics Netherlands (2015) published its first statistics purely based on big data in the area of transport statistics, more specifically traffic intensity statistics, based the total of counts performed each minute of vehicles crossing the more than 20 000 traffic loops on Dutch motorways. Statistics Netherlands perceived as an advantage that "results are more quickly available, more up to date and more detailed". The 115 billion observations in this exercise (corresponding to 80 terabytes) exceeded more than seven times the amount of data generally processed by the entire statistical office during a year…

3.4.2 Smart energy meters

Electronic devices recording energy consumption are more and more installed in private homes as well enterprises.

So far, experiments have mainly focused on population statistics, but applications for tourism statistics are self-evident. Tourists can be seen as a temporary population in the destination city, region or country.

Smart meters installed in holiday homes can detect on the basis of the usage pattern whether a given dwelling is likely to be a holiday home (a segment of accommodation that is often not represented in registers), for instance on the basis of energy consumption concentrated in weekends or typical holiday periods. Once identified as a holiday home, the energy consumption in subsequent periods can be used to estimate the occupancy, i.e. the number of nights actually spent. At a more macro-level, fluctuations in energy consumptions measured via smart meters can monitor seasonality with a much better temporal (and geographical) granularity.

3.4.3 Satellite images

Although of limited direct relevance for measuring tourism, satellite images can contribute to monitoring land use in endangered tourism areas or the urbanisation of natural heritage, for instance evolution of building construction in popular coastal areas.

3.5 Crowd sourcing

People do not only leave digital footprints but also actively generate information that can be a data source in itself for measuring human mobility, including the subcategory of tourism related movements. The next paragraphs discuss two particular cases of user-generated contents with relevance for tourism statistics.

3.5.1 Wikipedia contents

The relevance of Wikipedia page views as a source was discussed above in paragraph 3.2.1. Not only web

activity but also the underlying web contents can be relevant for tourism statistics. Detailed information on the location of sites, attractions, destinations (Wikipedia pages are often geo-located) can help to inventorise points of interest in a given country, region, destination.

Experiments are ongoing at Eurostat to use information on points of interest retrieved from Wikipedia as a key to improving geographical granularity of data on accommodation capacity (also comparing Wikipedia with other sources on points of interest, for instance collections made available by producers of traffic and navigation products). Aspects studied include the possible bias, for instance not all tourists will look for any information at all (beach and sun holidays) or repeat visitors are less likely to look up general information on attractions as compared to first-time visitors to the destination. The intensity of using Wikipedia will also differ depending on the age group or country (according to Eurostat data on ICT usage for 2015, 56% of the internet users in the European Union used the internet to consult wikis – ranging from 28% in Latvia to 82% in Luxembourg).

Hinnosaar et al. (2015) analysed the relationship of content availability on Wikipedia and choices of tourism destinations, including the causality of the relationship. Positive correlations were found, but with limited statistical significance – most probably because Wikipedia contents is only one of the many sources of information for tourists and because availability of information is only one of the driving factors in tourists' decision making process (alongside with cost, distance, etc.).

3.5.2 Picture collections

Travelling and taking pictures go hand in hand for many tourists. Since a decade, people share pictures online rather than via printed photo albums. The smart devices used to take the pictures typically log the location and time stamp. Although the above comment on bias also applies to picture collections (e.g. repeat visitors versus first-time visitors), tourist publicly disclosing pictures online are generating data.

Already ten years ago, Girardin et al. (2008) examined the potential of such digital traces to uncover the presence and movement of people in a city. The relevance of this data for tourism (and urban) planning was stressed in this study: "information about who populates different parts of the city at different times can lead to the provision of customised services (or advertising), accurate timing of service provision based on demand (e.g. rescheduling of monument opening times based on the presence of tourists), and, in general, more synchronous management of service infrastructures".

4 The impact of big data on a system of tourism statistics

This chapter summarises the insights gained from the previous chapter on big data sources with potential for tourism statistics into a sketch of a future tourism statistics system. How can different sources of big data complement each other and interact with data obtained from more traditional sources such as surveys or administrative data? What are the main outstanding gaps and methodological challenges?

A system of tourism statistics is not a matter of OR, but AND. While the current International Recommendations for Tourism Statistics IRTS2008 (UN/UNWTO (2015)) as well as Eurostat's Methodological Manual on Tourism

Statistics (European Commission (2014b)) focus mainly on surveys as a data source[3], the new sources outlined in Chapter 3 open entirely new perspectives on improving and enriching the existing system of tourism statistics, making it timelier and closer to the user needs. Tourism statisticians' skills will shift from (mainly) designing surveys to (also) putting the building blocks together and solving the multi pieced puzzle.

4.1 Evolution of the system of tourism statistics in the years to come

For many decades - if not since Quetelet organised the first international statistical conference in 1853 - official statistics relied on surveys and censuses. The exploration of alternative sources, in particular those held by other public authorities (so-called administrative data), has been paving the way for revolutionizing the way official statisticians work: namely to shift from being pure data collectors to becoming data connectors, assessing the relevance and methodological quality of a varied range of input sources and putting the pieces of the puzzle together to obtain a powerful information system.

In the lifecycle for the coming years, three stages can be distinguished.

In the short term, traditional surveys (household surveys, business surveys for the accommodation sector) remain the main input for primary tourism statistics, but big data sources slowly become important sources of auxiliary information (see Figure 2a).

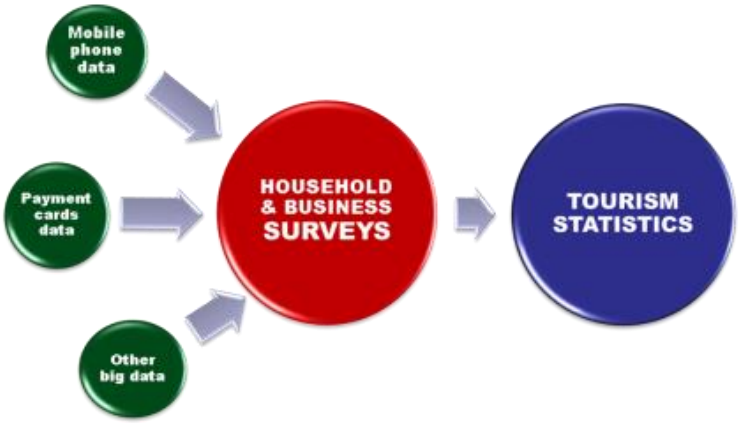
In the mid term, the weight of surveys is likely to decrease in favour of big data. In parallel the weight of new sources grows in a more integrated system (see Figure 2b). Traditional household and business surveys are no longer be the main filter but one of the many sources feeding the system of tourism statistics.

In the longer term, surveys are gradually (partially) replaced by new sources (see Figure 2c). Notwithstanding the data deluge, big data can for the time being not cover all aspects of tourism information. New sources will give insights on tourist flows (and expenditure?), with a revolutionary temporal and geographical granularity, but this information will be complementary to data collected via smaller-scale surveys. Indeed, information on socio-demographic characteristics of the traveller, the purpose of the trip, the means of transport of accommodation, etc. is difficult to retrieve from big data. This phase is expected to give users of tourism statistics timelier and more cost-efficient data. Moreover, the current data will be enhanced with previously unavailable indicators or breakdowns (hence the bigger 'pie' in Figure 2c). The latter can be very relevant for measuring sustainable tourism, an area of research where the absence of local, destination-level information or information for a specific (short) period has for many decades been a barrier to measuring the impact on tourism on the environment, on the economy, on the labour market and on local communities in a meaningful and methodologically sound way.

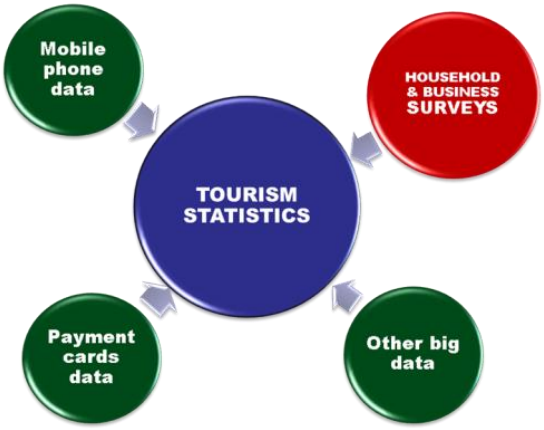
[3] Note that IRTS2008 mentions explicitly "other data sources such as credit card records" in the context of measuring tourism expenditure (para 4.30); the EU legislation opens the door to using – besides 'surveys' or 'appropriate statistical estimation procedures' – 'other appropriate sources, if these are appropriate in terms of timeliness and relevance.

Figure 2: Evolving system of tourism statistics

a. short term scenario



b. mid term scenario



c. longer term scenario

4.2 Ultimate target: regular production of mixed-source official statistics

Currently, many pilot projects are ongoing and statistical authorities start releasing so-called experimental

statistics[4] based on innovative sources. However, the final aim of the work is to transform the system of tourism statistics (or statistics in general) into a data factory using many input sources to serve simultaneously many output needs.

The feasibility of using big data is being explored at length; external sources start being used as auxiliary information for quality checks or for calibration[5] (see also Figure 2a). Following this, a next step (ongoing) is that big data is used to fill current data gaps and to produce (experimental) 'flash estimates'. Slowly but steadily, big data will partially (!) replace the traditional sources or surveys. Eventually, a rethinking, in a user-oriented way, of the system of tourism statistics will be necessary, fully taking account of the opportunities offered by integrating big data sources (and other types of smart data).

Indeed, the currently produced tourism statistics are often a product of the available sources ten or twenty years ago, rather than the concrete user needs. A "zero based budgeting" or "start from scratch" approach will be a necessary condition if we want to avoid that the statistical system stays with one leg in the 20th century (while emerging competitors don't…). This will be the final step, the evolution should become a revolution.

The two paragraphs below use some of the insights gained in Chapter 3 to tentatively reflect on future ways of combining sources.

4.2.1 Case 1: data on flows

Mobile network operator data is an obvious source to measure tourism flows. Call detail records and/or signalling information enhancing the coverage and completeness takes advantages of the footprint mobile phone users leave behind when they travel. This data can give geographical and temporal detail (destination, weekends) previously not available to users. The estimates should however be adjusted using auxiliary information to compensate for built-in biases (see also Chapter 5): flight reservation data to better cover more remote destinations where mobile phones may be under-used while travelling, credit card data, traffic counts, smart meters.

Specific variables such as purpose of the trip, composition of the travel party, expenditure will still need to be estimated from other sources, in particular sample surveys. However, new technologies can also lead to better ways of data collection, for instance combining automatically grabbed data on the movements from the respondents' phone or operator with follow-up questions presented via an app or pop-up screen to collect the remaining variables of interest.

---

[4] On 8 June 2017, Eurostat opened of a new section on its website, dedicated to experimental statistics. Eurostat's Acting Director-General, Mariana Kotzeva said at that occasion that "this is a major step forward for Eurostat; we now give access to Eurostat's innovation and development work to better respond to our users' needs and we are deliberately asking for feedback on these statistics and, through this site, expect an increased dialogue with users and the scientific community." Experimental statistics are compiled from new data sources and methods. For example, for the first time Eurostat is estimating price changes in the food supply chain, from farm to consumer. Another example, relevant to tourism, is the use of Wikipedia as a new source to produce statistics on the visits to UNESCO World Heritage Sites (see also paragraph 3.2.1. of the current paper) - this is to measure not only the popularity of the sites but also the public's 'cultural consumption'.

[5] Statistics Austria, for instance, carried out experiments with using payment cards data to check plausibility of travel statistics (and vice versa) and to generate a full geographical breakdown for tourism and travel statistics (whereas the inbound accommodation data is limited to 60 countries of origin).

4.2.2 Case 2: expenditure

In the case of expenditure, payments card data is an obvious source. Point of sale (POS) transactions can give information on the products or services purchased (via the merchant code a link to the economic activity is available, e.g. accommodation, transport, retail). The amount of ATM withdrawals can help to estimate the cash payments at the destination (however cash brought into the country of destination can induce a bias, in particular for shorter trips during which no local cash withdrawal may be needed). Filtering of non-tourist related transactions with foreign entities is essential; in this respect distinguishing between e-commerce and POS transactions seem to be generally possible. Breakdowns by expenditure type, could be obtained from retail cashier data or – again – from (smaller!) traditional surveys or mixed-mode data collections via respondents' smartphones (see also above).

5 Risk and constraints

The potential benefits of big data were outlined above in Chapter 3 and 4: improved overall quality, better timeliness, better geographical granularity, new indicators previously unavailable, synergies with other areas of statistics (namely using many sources for many purposes in one statistical ecosystem) leading to better coherence and comparability. All of the opportunities mentioned here are in particular relevant for the measurement of sustainable tourism where the past (and current) statistical methods don't succeed in addressing the detailed user needs.

Although cost-efficiency is often mentioned as a major advantage of using big data sources, their cost-cutting opportunities should not be overestimated. Survey fieldwork is a major cost driver for statistics, but handling of large volumes of data also comes at a price (note that the cost structure will also depend on the distribution of tasks between the entity holding the data and the national statistical office). Knowing that existing data collection cannot be fully replaced, the use of big data will not necessarily lead to a reduction in the number of processes or in overall workload for NSIs because new data will be processed in parallel with the existing system.

Coming back to the benefits and risks, there are two sides to the medal. Facing all the expected benefits, there is a range of challenges and barriers that need to be tackled. This chapter takes a closer look at the more negative side of the big data story. In this respect, it worth reminding the reader of Gartner's hype cycle: following a *peak of inflated expectations* and a *trough of disappointment*, a *slope of enlightment* will follow, resulting in the end in a *plateau of productivity*. Different sources of data are in different stages of the cycle.

When discussing risks and constraints, new sources are typically in a 'defensive' position. Results of pilots are compared with existing data – somewhat arrogantly labeled 'the ground truth'. To fully adhere to the scientific method, statisticians need to make a critical assessment of the current methodology (and even use new sources to do so). Having only 90% penetration rate of mobile phones in the population (and even less intense use when travelling?), is an issue that needs to be assessed and solved – but what about the tourism demand surveys using CATI data collection on the basis of landline registers where less than half of the population has a landline nowadays or what about dramatically falling response rates in surveys, sometimes below 50%, or significant bias

due to the memory effect[6]?

The discussion in this chapter wants to discuss the risks and constraints of new sources[7], but putting them in perspective compared with the shortcomings in the sources and methods that have been used by generations of statisticians in the past.

5.1 Access … and continuity of access

Some of the sources listed in Chapter 3 are 'open' but many types of big data are characterised by the fact that the data is held by private companies.

For different reasons, those entities holding the data may be reluctant to share their data with statistical offices: legal barriers (e.g. protection of personal data), internal data monetisation projects, business secret or fear of negative impact on the public opinion ("big brother is watching you"). Setting up partnerships is a key critical success factor; the governance of such partnerships needs to be a balanced win-win for all stakeholders involved. Alternatively, authorities can take regulatory measure to open privately held data for statistical purposes, giving national statistical offices access to the relevant new data sources.

Getting access to the data is not the end of the story. Compared with other data producers, official statistics have as a unique selling proposition its robustness and continuity of series ("we have data for 2016, but also for 2006 and 1996 and will also have it for 2026"). Relying on external data sources – whereas statistical authorities typically controlled the production chain from the questionnaire design phase until the final dissemination – creates a major critical risk with a likely dramatic impact on the trust users put in the statistical system. Instead of a production system with many tens of thousands suppliers (namely responding households and business), NSIs may face an oligopolistic model in which a handful of suppliers dominate the (data) market.

5.2 Alignment of concepts and definitions

Whereas surveys are designed with the sole purpose of gathering data, most new data sources are of a more organic nature. These sources potentially include relevant information, but it takes some digging and defining of algorithms on the databases before useful data can be extracted as input for producing statistics.

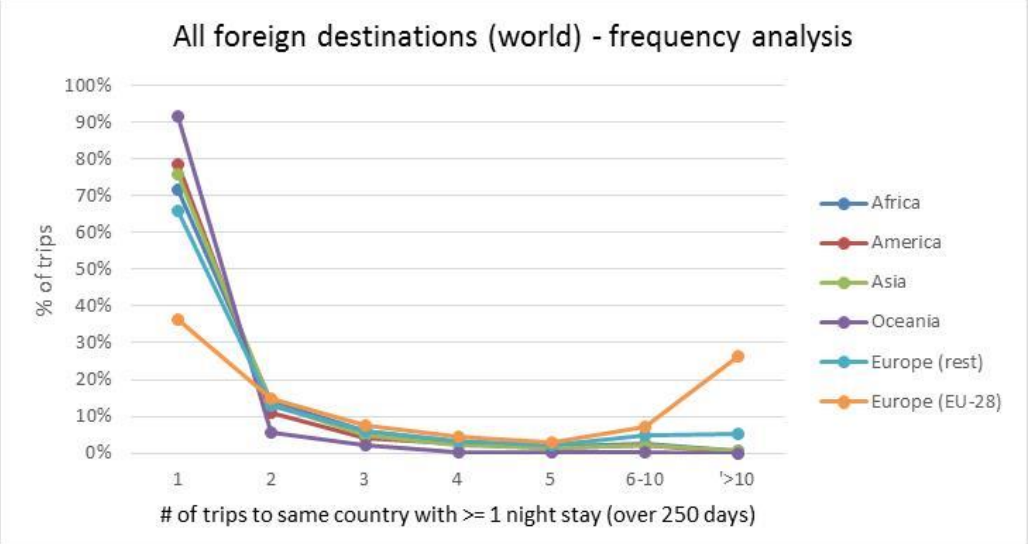
Not all movements away from home equal tourism activity. Algorithms to determine the usual place of residence and the usual environment are essential steps to identify a tourism trips when for instance using mobile network operator data. The subjective opinion of the responded is replaced with parameters defining the distance, the frequency and the duration of the movements observed in the subscribers' whereabouts. The choice of how many times a subscriber can be observed in a given destination (country) before the destination is considered part of the usual environment - and the length of the reference period considered – will directly impact the estimate of the

[6] Spain (Instituto de Estudios Turisticos (2008)) estimated the recall bias or memory effect of respondents in tourism demand surveys to cause a 15 to 20% underestimation of the number of trips made.

[7] Additional insights into the risks of big data sources can be gained from the stakeholder analysis carried out by Eurostat (Wirthmann et al. (2016)), in which respondents were asked to indicate (and comment) likelihood, impact, prevention and mitigation actions for specific big data sources.

number of trips. Figure 3 shows the distribution of SIM cards from one mobile network operator (Proximus, Belgium) (from Seynaeve & Demunter (2016)). For each destination (continent), the distribution of the number of SIMs observed at that destination is given, in terms of the number of times a SIM is observed during the 250 days window. For instance, within the group of SIMs observed at destinations within the EU-28, 36 % are observed only once, 15 % are observed two times, 8 % are observed 3 times, 5 % are observed 4 times, 3 % are observed 5 times, 7 % are observed between 6 and 10 times, 26 % are observed more than 10 times in other EU countries during the 250 days period (see orange line in Figure 3). For other continents, including European countries outside the EU-28, the majority of SIMs observed at these destinations are observed only once during the reference period. Less than 3% of SIMs observed in the remote continent of Oceania are observed on more than 2 trips to Oceania in the reference period (5.5 % twice, 91.8 % only once).

Figure 3: Distribution of SIMs, in terms of the number of times a SIM is observed during a 250 days period, by continent of destination[8]



An important issue when discussing concepts and definitions is the extent to which indicators from new sources are capable of reproducing the existing official data. However, a too restrictive approach can lead to undesirable lack of innovation and suboptimal exploitation of new data. As mentioned earlier (see 4.2), current data is based on current or past methods and sources. In some cases, new data can't reproduce the outcome of the traditional production process, but can instead produce statistics that have a superior relevance for users. For instance, in many European countries, inbound tourism statistics are limited to arrivals and nights spent at tourist accommodation establishments. New sources can give estimates on arrivals and nights spent regardless of the type of accommodation, compensating for the absence of data coming from border surveys or border controls.

5.3 Selectivity bias

Grossing up is relatively straightforward in sample surveys, obtaining the extrapolation factor by inversing the inclusion probability of an individual in the study population. Selectivity can be defined as a general term for self-

[8] Source: Proximus; taken from Seynaeve & Demunter (2016), p7.

selection error resulting from decisions of individuals (i.e. unit specific; e.g. whether to tweet, use a certain mobile provider) or from decisions of the owners of the technological platforms where data is captured (technology specific; e.g. in terms of business concept, technical infrastructure) (European Commission (2017)[9]). As a result, selectivity causes coverage error, measurement error and non-response error, which introduces potential bias in estimates based on big data sources.

To illustrate the selectivity bias, let's take the case of mobile network operator data (based on Seynaeve & Demunter (2016)).

Firstly mobile network operators have information on their market share (and the inverse of the market share would be a good first grossing-up factor to get to population estimates), but the market share can differ by region or by socio-economic group.

Secondly, penetration rates of mobile phone possession and use are not exactly 100%. This issue is similar to the issue of over-coverage or under-coverage of the sampling frame in traditional surveying.

Thirdly, subscribers may or may not make/take phone calls, send/receive message, connect to Wi-Fi networks depending on the time of the day or the place (e.g. while on holidays) or even switch off their device(s). This phenomenon, too, is comparable to the non-response or non-contacts that survey statisticians have to deal with. For the specific case of analysing outbound tourism through network signaling, bias could be introduced by devices being turned off before, or during tourism trips abroad, meaning country/network changes could go unnoticed.

The above sketched problems lead to a selectivity bias that needs to be taken into account when using mobile phone data. While it is generally expected that the use of big data can contribute to a reduction of respondent burden due to surveys, paradoxically the early phases of big data will see the necessity to collect auxiliary information via surveys to enable data scientists to correct for unevenly distributed market shares, for variable use patterns or for non-observation of devices.

Within the European Statistical System, initiatives are being set up to collect this kind of auxiliary information to support big data sources, not only for mobile network operator data but also for e.g. social media. Available data shows that the effects can be very significant. Recent data by the Italian statistical office ISTAT (Dattilo et al. (2016)) evaluates the mobile phone use by Italian residents during tourism trips. Nearly 90% of respondents made calls during trips within Italy but the intensity of use dropped to just over 70% for trips abroad. On the other hand, Wi-Fi internet (not SIM) appears to be relatively higher during trips abroad, possibly avoiding perceived roaming charges.

When quantifying the selectivity bias and the corresponding under-coverage or over-coverage risks, a correct comparison of 'old' and 'new' sources should put the observed risks in working with big data into perspective with currently observed methodological deficiencies such as dropping response rates (and the ensuing non-response bias) and the significant recall bias from which tourism demand surveys traditionally suffer.

[9] The main objective of this study was to identify existing methods which could be used to address the selectivity in big data sources, in order to be able to make unbiased inference for populations of interest in official statistics (e.g. resident population between 15 and 65 years old).

5.4 Quality, comparability over time

Use of big data sources involves a paradigm shift for official statistics. Statisticians find themselves in the role of data customers instead of data producers, and have to design statistical products from existing data sources. The production of official statistics is driven by high quality standards and principles. To be labelled as official statistics, statistical products produced from big data sources have to meet these quality standards. Bodies such as Eurostat have explored possible accreditation procedures that producers of official statistics can use to for assessing the quality big data sources (see Wirthmann et al. (2014)).

While most aspects of quality are relevant when working with new data sources, the comparability over time is of utmost importance for official statistics. As highlighted in paragraph 5.1, continuity of series is a key strength of official statistics. A shift towards new sources or methods can introduce a significant break in series. A shift towards using big data is not different in this respect. Even if the break in series can be communicated as an actual improvement of the overall quality of the data, it will be perceived as a major inconvenience by long-standing users of the data series.

Again, the example of mobile network operator data is used to illustrate this risk. A recent study (Seynaeve & Demunter (2016)) compared the MNO-based estimate of the number of outbound trips made by residents of Belgium with official statistics on the same variable for a comparable reference period.

For trends analysis, the two sources gave relatively comparable results. For instance, Figure 4 shows the distribution of outbound trips with a destination inside the European Union, by duration of the trip, calculated on the basis of the two sources. In general, the (big) data seems to make sense. Both graphs detect the typical holiday duration of 7 or 14 nights. However, the peak values for a trip duration of exactly 7 or 14 days were more pronounced for the survey based data – possibly due to rounding bias arising when respondents don't remember the exact duration (6 nights? or 7? or 8?) and approximate ("a week" – recorded as 7 days). A more striking observation is that the weight of ultra-short trips of one overnight stay was much higher in the mobile network operator data. Possible explanations include the initial parameter settings for the MNO data (namely a minimum duration of 10 hours and return after 4am) and the memory effect in the traditional tourism surveys (the shorter the duration, the more likely the respondent forgets to report the trip).

A lack of comparability or serious break in series becomes evident when looking at absolute figures – volumes instead of indices or trends – in Figure 5. In the same study, a comparison of the estimated number of outbound trips made by residents of Belgium during a six months period was made. The estimate for trips obtained from MNO data was (more than) twice as high as the official statistics; for nights the deviation was a bit less pronounced but still largely exceeding 50% - see Figure 5. When looking in more detail at the data, relatively larger differences were observed for destination countries close to Belgium (neighboring countries) than for further away destinations (see the referenced paper). Although the proximity of the destination seems to play a role, the differences between the estimates obtained from the two sources tend to be systematic. Differences in scope (official statistics representing only the population aged 15 and over) can explain the discrepancies only partially. The selectivity bias (see also paragraph 5.3) can in this case be caused by the fact that the MNO (Proximus) was an early player in the market and may still have a relatively wealthier customer base (more likely to travel?). More likely explanations are the incomplete fine-tuning of the algorithms used to estimate tourism from

the MNO data and the underestimating of tourism flows due to the recall bias in tourism demand surveys[10].

Figure 4: Comparison of the distribution of outbound trips to EU-28 by duration of the trips[11]

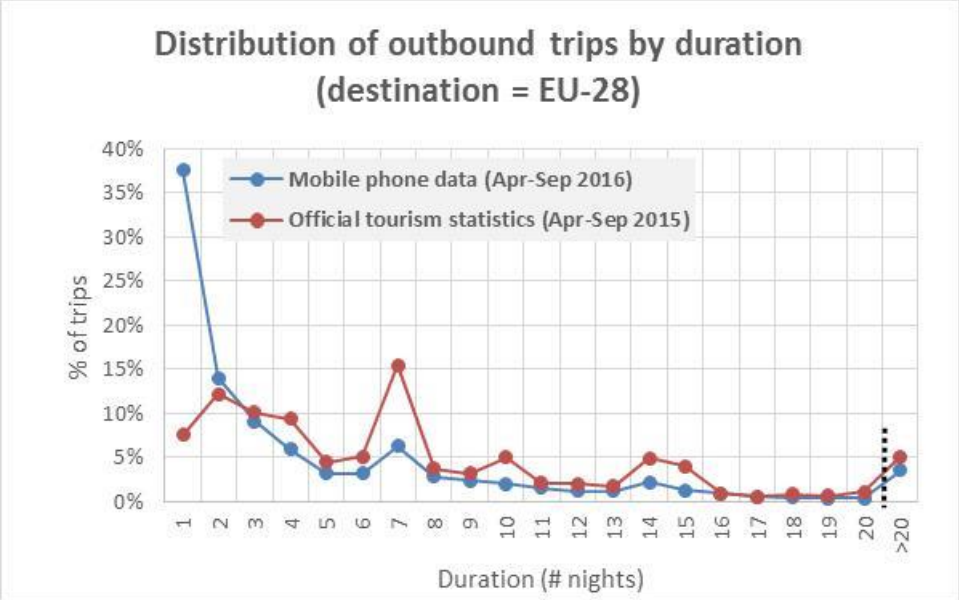
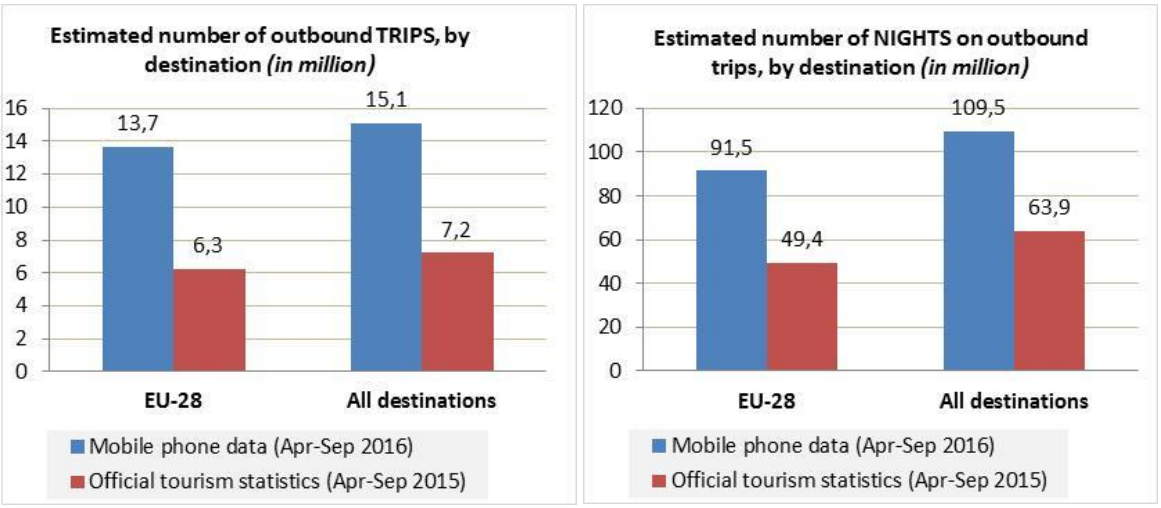


Figure 5: Comparison of estimated number of outbound trips and nights, by destination[12]



From Figure 5, it is clear that a simple shift from one source to the other would introduce an – for users – unacceptable jump in the data. At this stage of the research, the MNO data appears to be a good source for trend analysis, contrary to the analysis of the volume of tourism.

However, a critical assessment of the traditional survey could make the conclusion as well balance in favor of the

[10] It also needs to be pointed out that this particular tourism demand survey has a very low response rate (15%).

[11] Source: Proximus; taken from Seynaeve & Demunter (2016), p13.

[12] Source: Proximus, Statistics Belgium, Eurostat; taken from Seynaeve & Demunter (2016), p14.

newer source. To be continued…

5.5 Independence

Objectivity and independence are among the basic principles of official statistics. Making use of new sources includes the challenge of drawing valid statistics from these data sources and from samples we did not design ourselves (and were not even designed to produce statistics). Official statisticians, who are used to being in "full control" of the entire data production process, suddenly become data users who rely on the market for their ingredients and possibly have to negotiate the recipes to be applied with those who hold the data.

A dominant position of external sources can put the independence of statistical offices at stake, with platforms, MNOs or social media holdings taking over part of the control of the data and its quality[13].

5.6 Skills

The risk of lack of availability of experts consists of, upon receiving data from one of these new big data sources, the statistical office not having the possibility of processing and analysing it properly, due to its staff not having the required skills (Wirthmann et al. (2016)). The use of big data requires skills on model based inference and machine learning, skills in natural language processing, audio signal processing and image processing and a good understanding of distributed computing methodologies.

Furthermore, in an increasing 'data market' and the related cry for skilled data scientists, statistical offices risk losing their staff to other organisations after they have acquired big data related skills.

5.7 Trust

Change always implies regaining trust. This also holds for statistics produced using novel methods or sources and the trust users put in this data. A new aspect, however, is the trust of society in official statistics that make use of big data. The use of their digital footprint risks to be perceived as intolerably invasive by the citizens.

The impact would be a general loss of reputation of the statistical office that might negatively influence the general attitude of persons to collaborate with statistical offices. A negative public opinion might inhibit the use of specific big data sources for official statistics (Wirthmann et al. (2016)).

A suitable communication strategy before going into production and dissemination is crucial. The communication should stress the benefits of big data usage for the citizens, e.g. lower burden on respondents and improved

[13] An interesting example is the recent negotiation between the EU and the European MNOs regarding ending the roaming charges within the EU. To analyse the extent to which EU residents use their phone outside their country of residence, the European Commission made an estimate of the number of days spent abroad by an average European, using – among other sources - labour market statistics (cross-border commuting) and tourism statistics (outbound tourism data obtained via household surveys). If official tourism data were based entirely on MNO data, no objective reference data would have been around to enter the arena with the MNO (and to have a counter-examination of the industry's facts & figures).

statistical data while assuring data security and privacy. Communication campaigns should involve relevant stakeholders with the purpose of raising awareness and informing the public on the purpose of the big data usage for statistics. In this context, respondents consider transparency as key element of the communication strategy (Wirthmann et al. (2016)).

6 Conclusion

The overview of big data sources with relevance for tourism statistics makes clear that the new sources are here to stay. If statistical offices miss the train, others will serve the user needs that official statisticians can not or no longer serve with the same detail and timeliness. The era of national statistical offices' monopoly on statistical information is gone for good. However, in any democratic society, objective and independent data is an essential public good.

NSIs need to invest in skilled staff and in partnerships with those entities holding the data. Long term collaborations need to be set up to guarantee the continuity of data dissemination. The borderless nature of many of these sources necessitates an international collaboration and knowledge sharing – both in terms of governance and methodology.

Besides access and skills, a key question is whether new data can address the policy and research questions the same way as the good old statistics can. Statisticians (and users!) need to show a certain degree of flexibility or even revise the existing concepts and definitions in view of new and richer data sources. When building a new system of tourism statistics, it needs to be kept in mind that many methodological issues inherent to new data sources exist in one way or another also in traditional statistical techniques and as such affect the quality of statistics produced pursuant to those techniques. In the end, both approaches lead to an estimate, not to the true value. Remember, there's no such thing as a ground truth…

Probably the most poorly covered area of tourism statistics is the measurement of sustainable tourism. Big or smart data can be the missing link. Where national level data or annual data is of limited relevance to measure e.g. the impact of tourism on the environment, many of the sources discussed in this paper are likely to produce superior data in terms of geographical and temporal granularity. Previously utopic destination-level data or daily data is now within reach. The combination of different sources, including traditional surveys, can yield a very powerful ecosystem of data, if connected in the right and mutually complementary way.

The abundance of big data sources capable of capturing facets of the tourism phenomenon, makes it evident that the tourism statistician anno 2017 - and the coming decade(s) – is in the frontline of an exciting but challenging data revolution.

References

Ahas, R., Aasa, A., Roose, A., Mark, U. & Silm, S. (2008). *Evaluating passive mobile positioning data for tourism surveys: An Estonian case study*. Tourism Management 29 (2008) 469–486.

Almeida de Oliveira, R. & Abrantes Baracho Porto, R.M. (2016). *Extracting web data to from Tripadvisor as*

support for tourism indicators development in Minas Gerais, Brazil, paper for the 14th Global Forum on Tourism Statistics [link]

Beyer, M. & Laney, D. (2012). *The importance of 'big data': a definition*. [link]

Burson, R. & Ellis, P. (2014). *Using electronic card transaction data to measure and monitor regional tourism in New Zealand*, paper for the 13th Global Forum on Tourism Statistics [link]

Dattilo, B., Radini, R. & Sabato, M. (2016). *How many SIM cards in your luggage? A strategy to make mobile phone data usable in tourism statistics*, paper for the 14th Global Forum on Tourism Statistics [link]

De Meersman, F., Seynaeve, G., Debusschere, M., Lusyne, P., Dewitte, P., Baeyens, Y., Wirthmann, A., Demunter, C., Reis, F. & Reuter, H.I. (2016). *Assessing the Quality of Mobile Phone Data as a Source of Statistics*, paper for the European Conference on Quality in Official Statistics Q2016. [link]

ESSC (2013). *Scheveningen memorandum*. [link]

ESSC (2014). *Big data action plan and roadmap*. [link]

European Commission (2014a), *Feasibility Study on the Use of Mobile Positioning Data for Tourism Statistics*. [link to consolidated report, link to all deliverables]

European Commission (2014b), *Methodological manual for tourism statistics - Version 3.1.* [link]

European Commission (2017). *An overview of methods for treating selectivity in big data sources.* [forthcoming]

Girardin, F., Calabrese, F., Dal Fiorre, F., Biderman, A., Ratti, C., & Blat, J. (2008) *Uncovering the presence and movements of tourists from usergenerated content*, paper for the 9th International Forum on Tourism Statistics. [link]

Groves, R. M. (2011a). *'Designed data' and 'organic data'.* [link]

Hinnosaar, M., Hinnosaar, T., Kummer, M. & Slivko, O. (2015). *Does Wikipedia Matter? The Effect of Wikipedia on Tourist Choices.* [link]

Instituto de Estudios Turisticos (2008). *Memory Effect in the Spanish Domestic and Outbound Tourism Survey (FAMILITUR)*, paper for the 9th International Forum on Tourism Statistics.

Laney, D. (2001). 3-d data management: Controlling data volume, velocity and variety. META Group [now Gartner] Research Note. [link]

Schmücker, D., Sonntag, U. & Wagner, P. (2016). *Assessing the impact of "shared accommodation" for city tourism,* paper for the 14th Global Forum on Tourism Statistics [link]

Seynaeve, G. & Demunter. C. (2016). *When mobile network operators and statistical offices meet - integrating mobile positioning data into the production process of tourism statistics,* paper for the 14th Global Forum on Tourism Statistics [link]

Signorelli, S., Reis, F. & Biffignandi, S. (2016). *What attracts tourists while planning for a journey? An analysis of three cities utilising Wikipedia page views*. [link]

Statistics Netherlands (2015). *A first for Statistics Netherlands: launching statistics based on Big Data*. [link]

United Nations / UN World Tourism Organisation (2008), *International recommendations for tourism statistics*. [link]

Vij, A. & Shankari, K. (2015). *When is big data big enough? Implications of using GPS-based surveys for travel demand analysis*. [link]

Wirthmann, A., Stavropoulos, P. & Petrakos, M. (2014). *Proposal for an accreditation procedure for big data,* paper for the NTTS2015 (New Techniques and Technologies in Statistics) [link]

Wirthmann, A., Karlberg, M., Kovachev, B., Reis, F., Di Consiglio, L. (2016). *Assessment of risks in the use of big data sources for producing official statistics – Results of a stakeholder survey,* paper for the European Conference on Quality in Official Statistics (Q2016). [link]