Tourism statistics: Early adopters of big data?

2017 edition

STATISTICAL WORKING PAPERS

eurostat

Tourism statistics: Early adopters of big data?

2017 edition

Manuscript completed in September 2017

Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use that might be made of the following information.

Luxembourg: Publications Office of the European Union, 2017

The information and views set out in this publication are those of the authors and do not necessarily reflect the official opinion of the European Union. Neither the European Union institutions and bodies nor any person acting on their behalf may be held responsible for the use which may be made of the information contained therein.

PDF	ISBN 978-92-79-71899-1	ISSN 2315-0807	doi: 10.2785/762729	KS-TC-17-004-EN-N
PRINT	ISBN 978-92-79-71898-4		doi:10.2785/202932	KS-TC-17-004-EN-C

Abstract

Data is everywhere; and it is revolutionising the world of official statistics. People and businesses are leaving behind a constant flow of digital footprints, voluntary or unintended. This data deluge can make it difficult to see the wood for the trees; yet big data undeniably has huge potential for many areas of statistics.

The arrival of big data is also changing the working environment for statisticians. They no longer hold a monopoly on producing statistics, but now compete with a wide range of data producers. Ignoring innovation will push statistical authorities out of the information market — a development that could 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 are on the frontline of big data-related innovations of sources and methods. Tourism statistics try to capture physical flows of people — as well as the accompanying monetary flows; 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 in compiling tourism statistics. It discusses 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; user trust; partnerships with 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; the need for new skills, and so on.

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

Keywords: big data, tourism statistics, innovation

Author: Christophe Demunter (European Commission, DG EUROSTAT)

This paper is based on a keynote prepared by EUROSTAT for session 5 of the Sixth UNWTO International Conference on Tourism Statistics – Measuring Sustainable Tourism (Manila, Philippines, 21 – 24 June 2017).

Table of contents

1 Big data & the seven Vs… Or three Vs? Or no Vs at all?

There are probably more articles attempting to define big data than articles actually applying big data sources and techniques. This paper does not define big data, and trying to do so is perhaps not even relevant, as it is an ever-changing concept. To describe or delineate big data, authors have made reference to the three Vs: volume, velocity and variety (see, for instance, Laney (2001) and Beyer & Laney (2012)).

Briefly, ***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?) — and to the variables observed; ***velocity*** refers to how quickly this data is generated and their resolution in time. Several infographics depicting what allegedly happens in an internet minute or second are circulating on the internet itself. Every second 2.5 million e-mails are sent, over 50 000 Facebook updates are posted, over 60 000 Google searches are entered and USD 15 000 is spent online. By the time you read this, these figures will certainly be outdated. However, the internet is not the only data source. Also every second, mobile network operators store many gigabytes of information linked to the whereabouts and service usage of their subscribers, supermarkets gather a constant flow of cashier data, and so on.

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 (posted on, say, Instagram or Facebook), website logs, videos (e.g. camera surveillance), recordings or geo-coded data.

Besides these three ‘core’ Vs of big data, which reflect a more IT oriented perspective, other key Vs have entered the debate in recent years: veracity, validity, volatility and value.

Veracity and **validity** touch upon the quality, reliability and 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 a doubt some of the more complicated challenges for statisticians. Volume and velocity, on the other hand, are relatively 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, which tend to feature non-probabilistic samples and unknown inclusion probabilities. Coverage of big data sources (e.g. not everyone having a mobile phone) or self-selection error (e.g. not everyone using a mobile phone as much) requires new methodological approaches to ensure quality and secure the trust of data users and data producers. For a study providing an overview of methods to address selectivity bias in big data sources, see European Commission (2017).

Volatility refers to how long the data remains relevant and how long it should be kept, bearing in mind the billions of impulses registered every second or the legal framework on retaining 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, for statisticians as a potentially richer or timelier data source; and secondly, for businesses and policy-makers in an era of data-driven decisions. Those businesses or organisations holding the data are a special case. Data is a valuable, marketable asset and can make these stakeholders reluctant to grant access to the data they hold.

As outlined above, some authors tend to jump on any new hype and produce an array of buzzwords — in this case any number of Vs between 1 and 10… Although they can help describe the phenomenon, it is becoming increasingly clear that big data is not always that big, that smart data is not always that smart, and that new sources are not always that new. The promising data sources discussed in this paper could as well be described in a more statistical perspective as *high-detail exhaust data left by the use of IT systems or captured by sensors* (European Commission, 2017).

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 September 2013 in adopting 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 and an ESS task force. The latter brings together members from national statistical authorities, UN organisations, other European Commission services and scientific advisers. Its *Big data action plan and roadmap* (ESSC (2014)) was adopted by the ESSC in 2014.

The task force works on implementing the action plan. Eurostat has 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 and will run until mid-2018. At its heart is a set of pilots run by national statistical institutes. Their purpose is to explore the potential of selected big data sources to produce official statistics and apply the results to specific statistical domains. These source-oriented pilot projects include web scraping (for company characteristics and job vacancies), smart meters, mobile phone data and AIS data (automated tracking of ships).

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

Lastly, Eurostat has launched a series of in-house big data pilots 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 all these activities is to pave the way for bringing big data sources into the regular production process for official statistics.

Given the borderless nature of these new data sources — available on the world wide web or held by organisations across the globe using somewhat comparable data structure definitions — international cooperation on statistics has never been more crucial. It can range from developing strategies for data access and handling to troubleshooting methodological issues and disseminating trusted statistics that meet the needs of 21st-century users.

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

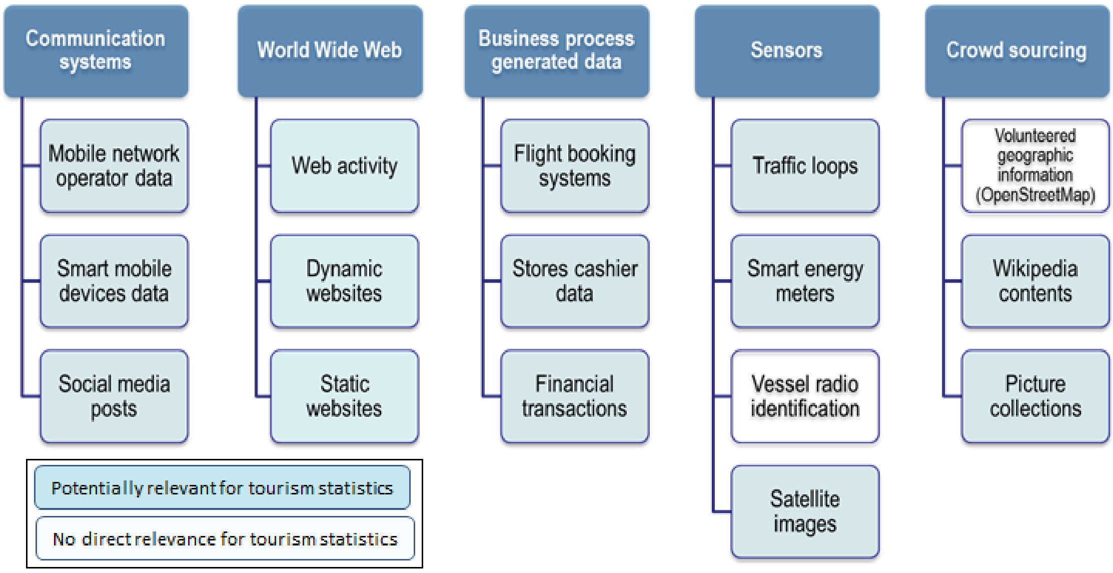
[2] An ESSnet project consists of a network of several ESS organisations aiming to provide results that will be beneficial to the whole ESS.

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

Many big data sources measure human activity or mobility — in other words, flows of people or the transactions they make. With primary tourism statistics measuring physical flows (and the corresponding monetary flows) of people, it comes as no surprise that tourism statistics have been on the frontline of big data-related innovations of statistical sources and methods. In this respect, tourism statisticians can help shape the future of official statistics. They can benefit from experiments and share their experiences; and they can become trailblazers in rethinking statistical systems beyond the traditional methods based on sample surveys targeting households or businesses.

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

Figure 1: Taxonomy of big data sources



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

This chapter briefly introduces the different sources with potential relevance for measuring tourism (highlighted in the classification scheme in Figure 1). While some sources are more promising or more widely used, others are included in the analysis only to provide complete coverage. At first glance the scheme shows that links to tourism are omnipresent. Rather than collecting data, the tourism statisticians of the future will be linking data from various sources into a modernised tourism information system (see also Chapter 4 below).

3.1 Communication systems

This group comprises 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. This group includes social media posts as well.

Until now, mobile positioning data has been a focal source for big data research. Firstly, it exists in all countries (which does not, however, mean that it can be accessed or used in all countries). Secondly, many promising studies or experiments are available. Thirdly, it has potential relevance for many different areas of statistics, making synergies possible.

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 falling 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 help paint a clearer picture of the usual environment, for instance by determining social networks on the basis of call history (who calls whom).

Since the pioneering work of Ahas et al. (2008) in exploring the use of mobile network operator data for statistics (in particular tourism statistics) nearly a decade ago, the landscape has tremendously changed. Up to now, experiments with using mobile network operator data were largely limited to the use of call detail records (CDRs) — 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, the changed behaviour of mobile phone users (use of alternative non-SIM-based messaging services, alternative voice and video call systems) is increasingly affecting the relevance of CDRs. This prompts the need for auxiliary data to assess the selectivity bias of this source, and to correct/calibrate for this bias. The work carried out by the Italian statistical office in assessing the use pattern of mobile phones in a tourism context is a good illustration of this (Dattilo et al. (2016)). It shows that tourists use their mobile phones on 90 % of domestic trips but only 71 % of outbound trips.

On the other hand, MNOs are shifting towards other data sources 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 number of useful signalling events is up to ten times higher as compared with CDRs (De Meersman et al. (2016)). In the case of one Belgian mobile network operator, Proximus (see Seynaeve & Demunter (2016)), the network detects a device's position at least every three hours (unless the device is switched off). For devices with data switched to 'on', this increases to approximately every hour. In practice, however, because of call, message and data usage, 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; one in every three devices is detected within 15 minutes. The mix depends on the actual usage and on the technology (e.g. 4G devices typically give more location points than 2G devices).

Mobile network operator data is a textbook example of how one source can serve multiple statistical domains simultaneously. MNO data can reveal information 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, longer-distance movements can reveal information on trips made outside the usual environment; as such, tourism statistics are interested in the noise that can be observed in the data. In this respect, a key challenge is how to determine the usual environment correctly from the data (and not from the subjective opinion of the survey respondent).

Despite the evident relevance of MNO data for tourism statistics, gaining access to the data for research purposes or for producing statistics unfortunately remains the main barrier to 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 captured by the device itself instead of the mobile network, from its activity sensors or from installed apps. As such, this group can be extended to include other smart mobile devices, for instance tablets, as long as they include sensors and geolocation services.

The geo-positioning data and information from activity sensors stored on the device can include very relevant information for analysing mobility, in particular tourists' movements. The geo-positioning data captured by the devices themselves is more precise than the one captured by the mobile network and can be measured at regular times instead of being dependent on communication events between the device and the network. Therefore, smart mobile devices data can be superior in detecting the usual environment of the user of the device and the movements outside this usual environment (i.e. tourism trips). This source can also prove particularly interesting in mixed-mode data collection, where respondents are selected via traditional sampling design methods but part of the data collection is automated from the device (with additional information entries on, for example, the purpose of the trip or purchases made). In the European Union, experiments are ongoing to explore 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 former 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 the mode of transport, purpose of trips, spending, etc. This observation holds true for most of the big data sources discussed in this paper.

3.1.3 SOCIAL MEDIA POSTS

Whether or not they intend to do so, people leave their digital footprint when using social media. Posts can be an information source on people's movements and behaviour.

Although the relevance for measuring tourism flows is obvious, this source faces a number of significant 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 its usefulness to detecting short-term trends rather than volume information or longitudinal trends.

Other challenges include the absence of socio-demographic information (although ongoing profiling exercises are testing whether the socio-demographic status can be detected within the data) and the continuity or sustainability 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 influence Facebook will wield five or ten years from now.

3.2 World Wide Web

The following sections focus on the internet as a data source. (Certain sources mentioned elsewhere in this chapter could also be included under the current 'world wide web' heading.)

3.2.1 WEB ACTIVITY

This group includes the traces left by people while using search engines (e.g. Google Trends data) and visiting websites (e.g. Wikipedia page views).

Web activity can give an indication of which topics interest people at each moment in time. Searching for information on tourism destinations or page views of Wikipedia articles related to destinations can do much to help predict tourism flows. Obviously, interest via search queries or visiting websites does not always lead to a physical visit or a purchase, but a correlation has been found several times (e.g. Sharpe et al. (2016) and Miao et al. (2015)). 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 in producing volume data or absolute numbers, it can be a useful starting point for estimating breakdowns that are otherwise difficult to collect (e.g. tourism activities, by looking at the predominance of searches for, say, 'cruise', 'golf' or 'gastronomy').

'Wikipedia contents' are mentioned under the heading 'Crowd sourcing' in the taxonomy presented in Figure 1. However, a derived source such as *page views* of Wikipedia articles can be a proxy for visits to a destination, measured through the traveller's web activity. Signorelli et al. (2016) have 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 launched a project exploring how Wikipedia page views of (tourist) places of interest can help disaggregate annual or national level tourism statistics into more detailed (and user-relevant) infra-annual or regional series. For this, 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. This would depend on 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 sites such as Wikipedia.

3.2.2 DYNAMIC WEBSITES

Normally, dynamic websites feature structured data and an interface to access and consult a (dynamic) database. Typical examples include tripadvisor.com, booking.com or airbnb.com. In general, the data is obtained via web scraping, where pieces of relevant data are extracted from the web pages returned by dynamic website.

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

Scraping websites has produced some information on the supply (e.g. number of hotels or apartments). By contrast, obtaining information on actual occupancy (nights in use) is more complicated (see, for instance, Schmücker et al. (2016)). One practical problem is that websites can detect and block the bots scraping the information, which means this source may be in general of dubious reliability in providing a longer-term perspective (but is 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 for statistical offices to 'go rogue' on the websites and instead the company operating the website should be informed in advance. On the other hand, informing directly the companies behind the websites may jeopardise the independence and objectivity of statistical offices; unidentified bots can better avoid possible manipulation by the company in an attempt to influence the statistics.

The European project on big data (ESSnet, see Chapter 0) includes scraping job portals to produce job vacancy statistics.

3.2.3 STATIC WEBSITES

Contrary to the dynamic websites discussed above, static websites are composed of a limited amount of webpages which do not change frequently (apart from, for instance, embedded social media feeds). Businesses or organisations normally disseminate content for their stakeholders (customers, fans, etc.) via static websites. Data is obtained by extracting the contents from the html source code and transforming them, clustering them 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 businesses produce a constant flow of data through their regular business processes. The following sections discuss how this data can be relevant for measuring tourism.

3.3.1 FLIGHT BOOKING SYSTEMS

Air travel leaves a trace via airline companies' reservation systems or transaction processors such as Amadeus.

This source is by default incomplete — it only covers air travel, and even then it does not include all airline carriers (low-cost companies, for instance tend to be underrepresented). However, the data can be useful for specific destinations (in particular for islands largely visited by plane) or as auxiliary information for tourism demand surveys, as trip data from sample surveys tends to be rather unreliable for more remote destinations.

3.3.2 STORES CASHIER DATA

Tourists leave a digital trace of their stay from purchases made in local retail stores. Seasonal fluctuations in turnover (or in the 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 understanding tourism's impact and sustainability.

Furthermore, electronic payments in stores could serve as a source for estimating TSA (tourism satellite accounts) tourism ratios for retail industries (the share of cards used 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 to help estimate tourists' countries of origin.

3.3.3 FINANCIAL TRANSACTIONS

Decades before the big data concept emerged, tourism statistics set about using payment card data to measure tourism and travel (in balance of payments terminology). Many tourism statistics systems — and big data sources outlined elsewhere in this chapter — focus on physical flows (accommodation statistics, border counts). This makes payment card data the missing link to monetary information.

The quality of the data has improved over recent years. For instance, it is now possible to distinguish between foreign purchases from e-commerce versus point-of-sale transactions, helping detect economic activity under ISIC([3]) or NACE([4]) using the merchant code. However, despite the huge potential that this source can offer, few experiments have taken place (due to the sensitivity of the data). Burson & Ellis (2014) developed a methodology for using electronic card transaction data to measure and monitor regional tourism in New Zealand.

Despite the applications available from this source, some 'built-in' issues can limit its usefulness. In the non-cashless society that still prevails worldwide, an increase in the volume and value of transactions observed does not necessarily translate into increased tourism figures, because it may be caused only by a substitution effect (as people first begin switching from cash payments to card payments). Auxiliary information on how people use cards would help produce meaningful estimates for absolute values (see also the discussion on mobile network operator data in section 3.1.1).

3.4 Sensors

Sensors monitor people's movements, 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 has opened up new possibilities.

Statistics Netherlands (2015) published its first statistics based purely on big data on transport, and more specifically on traffic intensity. These were based on the total counts performed each minute of

([3]) International Standard Industrial Classification (UN).

([4]) Classification of economic activities in the European Union.

vehicles crossing more than 20 000 traffic loops on Dutch motorways. Statistics Netherlands welcomed the fact that results were ‘more quickly available, more up-to-date and more detailed’. The 115 billion observations in this exercise (corresponding to 80 terabytes) amounted to more than seven times the amount of data usually processed by the entire statistical office in a year.

3.4.2 SMART ENERGY METERS

A growing number of electronic devices recording energy consumption are being installed in private homes and on business premises.

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 can draw on the usage pattern to detect whether a given dwelling is likely to be a holiday home (an accommodation category that does not often appear in registers), for instance if energy consumption is concentrated during 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 it was actually in use. On a more macro level, fluctuations in energy consumption measured via smart meters can monitor seasonality with much better temporal (and geographical) granularity.

3.4.3 SATELLITE IMAGES

Although of limited direct relevance for measuring tourism, satellite images can help monitor land use in endangered tourism areas or the urbanisation of natural heritage sites — for instance, construction trends 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 and its subcategory of tourism-related movements. The following sections 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 in section 3.2.1 above. Both web activity and its underlying web contents can be relevant for tourism statistics. Detailed information on the location of sites, attractions or destinations (Wikipedia pages are often geo-located) can help generate inventories of points of interest in a given country, region or destination.

Experiments are ongoing at Eurostat to use information on points of interest retrieved from Wikipedia as a key to improving the geographical granularity of data on accommodation capacity and occupancy. (This also involves comparing Wikipedia with other sources on points of interest, for instance collections made available by producers of traffic and navigation products). Aspects studied include possible bias: for instance, some tourists will not look for any information at all (beach and sun holidays); and repeat visitors are less likely to look up general information on attractions than first-time visitors to a destination. The intensity of Wikipedia usage will also differ depending on the age group or country (according to Eurostat data on ICT usage for 2015, 56 % of 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 between 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 are only one of the many information sources for tourists and because availability of information is only one of the driving factors in their decision-making process (alongside cost, distance and so on).

3.5.2 PICTURE COLLECTIONS

For many tourists, travelling and taking pictures go hand in hand. For a decade now, people have been sharing pictures online rather than in 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), tourists publicly disclosing pictures online are generating data.

A decade ago, Girardin et al. (2008) examined the potential of such digital traces to uncover the presence and movement of people in a city. In their study they highlighted the relevance of this data for tourism (and urban) planning: ‘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 the tourism statistics system

This chapter summarises the insights gained from the previous chapter on big data sources with potential for tourism statistics to sketch out 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 tourism statistics system is not a matter of OR, but AND. While the current *International recommendations for tourism statistics* (UN/UNWTO (2008)) and Eurostat's *Methodological manual for tourism statistics* (European Commission (2014b)) focus mainly on surveys as a data source[5], the new sources outlined in Chapter 3 open up entirely new possibilities for improving and enriching the existing tourism statistics system, making it timelier and closer to user needs. Tourism statisticians' skills will shift from (mainly) designing surveys to (also) putting the building blocks together and solving the multi-piece puzzle.

4.1 Evolution of the tourism statistics system in future years

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

In the coming years we can expect three distinct stages to emerge.

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

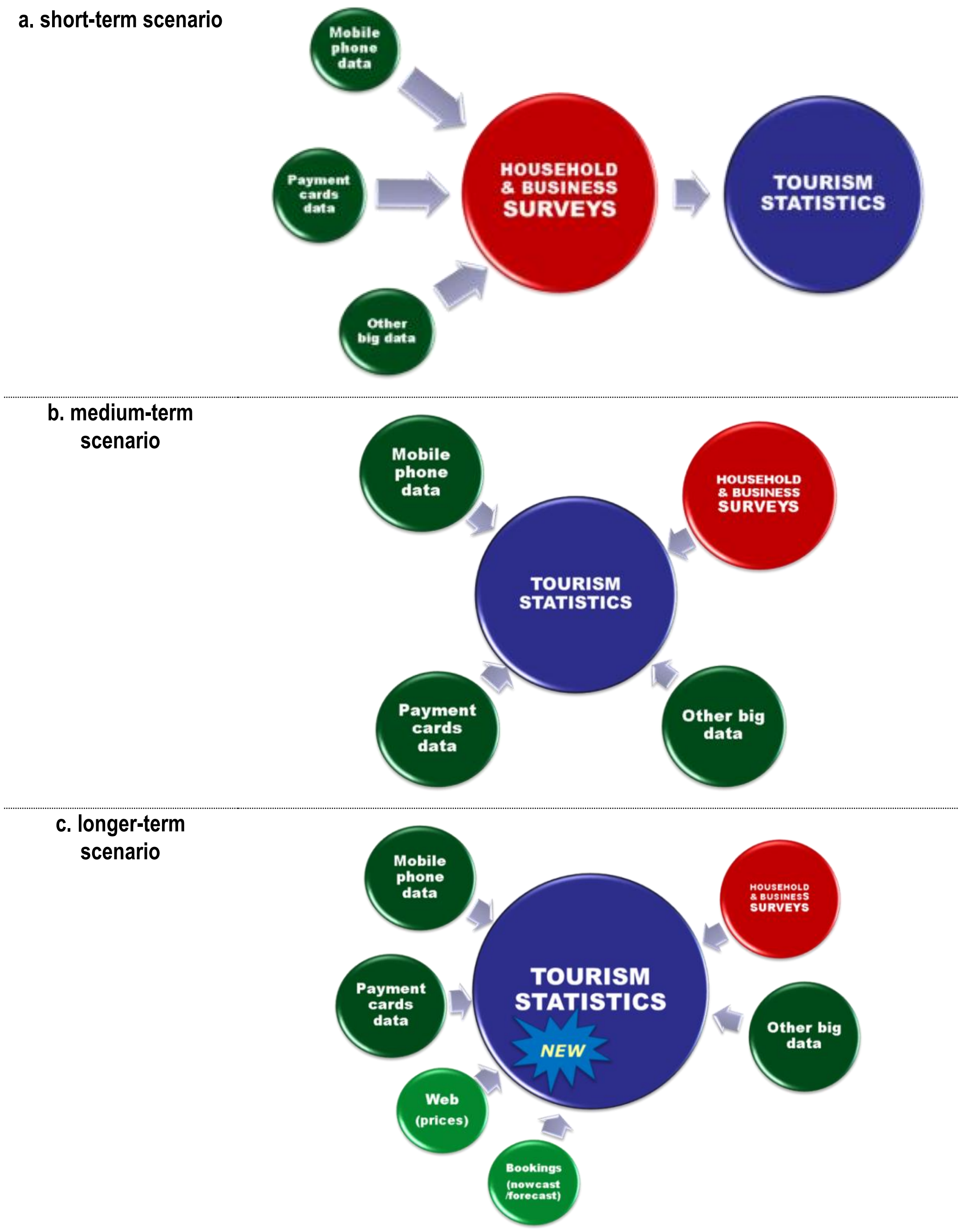
In the medium term, the influence of surveys is likely to decrease in favour of big data. In parallel, new sources will see their impact grow in a more integrated system (see Figure 2b). Traditional household and business surveys will no longer be the main filter, but will rather be one of the many

[5] Note that IRTS 2008 mentions explicitly 'other data sources such as credit card records' in the context of measuring tourism spending (paragraph 4.30); EU legislation on tourism statistics 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.

sources feeding the tourism statistics system.

In the longer term, surveys will be gradually (partially) replaced by new sources (see Figure 2c). In spite of the data deluge, big data cannot cover all aspects of tourism information for the time being. New sources will give insights into tourist flows (and spending?), with a revolutionary temporal and geographical granularity, but this information will be complementary to data collected via smaller-scale surveys. Indeed, information on the traveller’s socio-demographic features, the purpose of the trip, the means of transport, the means of accommodation, etc. is difficult to retrieve from big data.

Figure 2: How the tourism statistics system will evolve



a. short-term scenario

b. medium-term scenario

c. longer-term scenario

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). This 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 tourism's impact on the environment, on the economy, on the labour market and on local communities in a meaningful and methodologically sound way.

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

Currently, many pilot projects are ongoing and statistical authorities have begun releasing 'experimental statistics'([6]) based on innovative sources. However, the ultimate aim is to transform the tourism statistics system (or statistics in general) into a data factory using many input sources to serve many output needs simultaneously.

The feasibility of using big data is being explored at length; external sources are now being used as auxiliary information for quality checks or for calibration([7]) (see also Figure 2a). Following this, a next (ongoing) step is for big data to fill current data gaps and produce (experimental) 'flash estimates'. Slowly but steadily, big data will partially(!) replace traditional sources or surveys. Eventually we will need a user-oriented rethinking of the tourism statistics system, taking full account of the opportunities that integrating (big) data sources can offer.

Indeed, current tourism statistics are often a product of sources that were available ten or twenty years ago, rather than of concrete user needs. A 'zero-based budgeting' or 'starting from scratch' approach will be vital if we want to avoid that the statistical system stays with one leg in the 20th century (while emerging competitors do not). This will be the final step, with the evolution becoming a revolution.

The two sections 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 for measuring tourism flows. CDRs and, in particular, signalling information enhancing coverage and completeness, take advantage of the footprint mobile phone users leave behind when they travel. This data can give geographical and temporal detail (destinations, weekends) previously not available to users. The estimates should, however, be adjusted using auxiliary information to compensate for built-in biases (see also

([6]) On 8 June 2017, Eurostat opened of a new section on its website, dedicated to experimental statistics. Marking the occasion, Eurostat's Acting Director-General, Mariana Kotzeva, said 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 fork. Another example, relevant to tourism, is the use of Wikipedia as a new source for producing statistics on the visits to UNESCO World Heritage Sites (see also section 3.2.1 of this paper). The aim here is to measure not only the popularity of the sites, but also the public's 'cultural consumption'.

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

Chapter 5). This information would come from: flight reservation data, to better cover more remote destinations where mobile phones may be under-used while travelling; credit card data; traffic counts; and smart meters.

Specific variables such as purpose of the trip, composition of the travel party and spending 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 from the mobile network operator with follow-up questions presented via an app or pop-up screen to collect the remaining variables of interest.

4.2.2 CASE 2: SPENDING

In the case of spending, payment card data is an obvious source. Point-of-sale (POS) transactions can give information on the products or services purchased (the merchant code produces a link to the economic activity, e.g. accommodation, transport, retail). Data on ATM withdrawals can help 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 non-tourist-related transactions with foreign entities is essential; in this respect, distinguishing between e-commerce and POS transactions seems possible overall. Breakdowns by spending 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 Risks and constraints

The potential benefits of big data were outlined above in chapters 3 and 4. They include: higher overall quality; better timeliness; better geographical granularity; new, previously unavailable indicators; and 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 especially relevant for measuring sustainable tourism, where past (and current) statistical methods fail to address detailed user needs.

Although cost efficiency is often mentioned as a major advantage in using big data sources, the cost-cutting opportunities should not be overestimated. Survey fieldwork is a major cost driver for statistics, but handling 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). As 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 coin. In contrast to all the expected benefits, a range of challenges and barriers need to be tackled. This chapter takes a closer look at the more negative side of the big data story. In this respect, it's worth reminding the reader of Gartner’s hype cycle: following a *peak of inflated expectations* and a *trough of disappointment*, a *slope of enlightenment* will follow, resulting in a *plateau of productivity*. Different sources of data are at different stages of this cycle.

When discussing risks and constraints, new sources are typically in a ‘defensive’ position. The results of pilots are compared with existing data — somewhat arrogantly labelled ‘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). A mobile phone penetration rate of only 90 % (with use falling even further when travelling?), is an issue that needs to be assessed and solved — but what about the tourism demand surveys using phone interviews (CATI) on the basis of landline registers, where less than half of the population has a landline nowadays? Or what about the dramatically falling response rates in surveys, sometimes below 50 %, or the significant bias due to the memory effect([8])?

This chapter discusses the risks and constraints of new sources([9]), while putting them into perspective when compared with the shortcomings in the sources and methods used by past generations of statisticians.

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

([9]) 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 on) likelihood, impact, prevention and mitigation actions for specific big data sources.

5.1 Access… and continuity of access

Some of the sources listed in Chapter 3 are ‘open’, but a feature common to many types of big data is that the data is held by private companies.

Those entities holding the data may be reluctant to share it with statistical offices for different reasons: legal uncertainty concerning the obligation to provide such data when it has not happened in the past and data holders are on the other hand wary of personal data protection legislation, internal data monetisation projects or fear of public disapproval (‘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 involved. Alternatively, authorities can take regulatory measures to open privately held data for statistical purposes, giving national statistical offices access to the relevant new data sources.

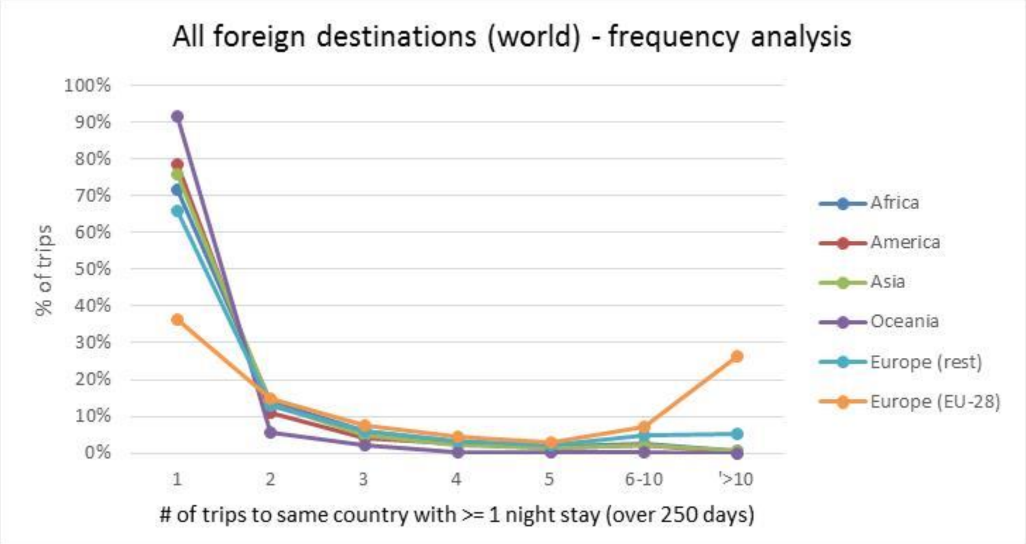
Getting access to the data is not the end of the story. Compared with other data producers, official statistics have a unique selling proposition their robustness and series continuity (‘we have data for 2016, for 2006 and 1996 as well, and we will also have it for 2026’). Relying on external data sources — while statistical authorities have typically controlled the production chain from questionnaire design to 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 of 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

Surveys are designed with the sole purpose of gathering data, but 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 in identifying tourism trips when, for instance, using mobile network operator data. The respondent’s subjective opinion is replaced with parameters defining the distance, frequency and duration of the movements observed in the subscribers’ whereabouts. The choice of how many times a subscriber has to 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 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 a 250-day window. For instance, within the group of SIMs observed in other EU countries, 36 % are observed only once, 15 % are observed twice, 8 % are observed 3 times, 5 % are observed 4 times, 3 % are observed 5 times, 7 % are observed between 6 and 10 times, and 26 % are observed more than 10 times during the 250-day period (see orange line in Figure 3). For other continents, including European countries outside the European Union, most SIMs are observed only once during the reference period. Less than 3 % of SIMs are observed on more than two trips to the remote continent of Oceania (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-day period, by continent of destination([10])



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, an overly restrictive approach can lead to an undesirable lack of innovation and suboptimal exploitation of new data. As mentioned earlier (see section 4.2), current data is based on current or past methods and sources. In some cases, new data cannot reproduce the outcome of the traditional production process, but can instead produce statistics of more relevance to users. For instance, in many European countries, inbound tourism statistics are limited to arrivals and nights spent in tourist accommodation. 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

Selectivity can be defined as the result of coverage problems in frame populations and self-selection, resulting from decisions made by individuals (i.e. unit-specific decisions about, say, whether to tweet or to use a certain mobile provider), or by the owners of the electronic platforms where data is captured (technology-specific decisions about, say, a business concept or technical infrastructure) (European Commission (2017)([11])). As a result, selectivity, when present, introduces bias in estimates made from big data sources.

Mobile network operator data is a useful way of illustrating the selectivity bias (see Seynaeve & Demunter (2016)).

Firstly, MNOs have information on their market share (and the inverse of the market share would be a good first grossing-up factor for population estimates). However, the market share can differ by region or socio-economic group.

([10]) Source: Proximus, taken from Seynaeve & Demunter (2016), p. 7.

([11]) The main objective of this study was to identify existing methods that could be used to address the selectivity in big data sources, so as to allow for unbiased inference for populations of interest in official statistics (e.g. residents between 15 and 65 years of age).

Secondly, penetration rates for 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 messages, connect to Wi-Fi networks depending on the time of the day or the place (e.g. while on holiday), or may 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 signalling, bias could be introduced by devices being turned off before or during tourism trips abroad, meaning country/network changes could go unnoticed.

The problems outlined above lead to a selectivity bias that needs to be taken into account when using mobile network operator data. While it is generally expected that the use of big data can help reduce the respondent burden from surveys, paradoxically the early phases of big data will involve collecting auxiliary information via surveys to enable statisticians to make corrections 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 social media and others. Available data shows that the effects can be very significant. As mentioned earlier, recent data from the Italian statistical office ISTAT (Dattilo et al. (2016)) has assessed 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 use (not SIM-related use) appears to be relatively higher during trips abroad, possibly to avoid potential 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. It should take account of current methodological deficiencies such as falling response rates (and the ensuing non-response bias) and the significant recall bias from which tourism demand surveys traditionally suffer.

5.4 Quality, comparability over time

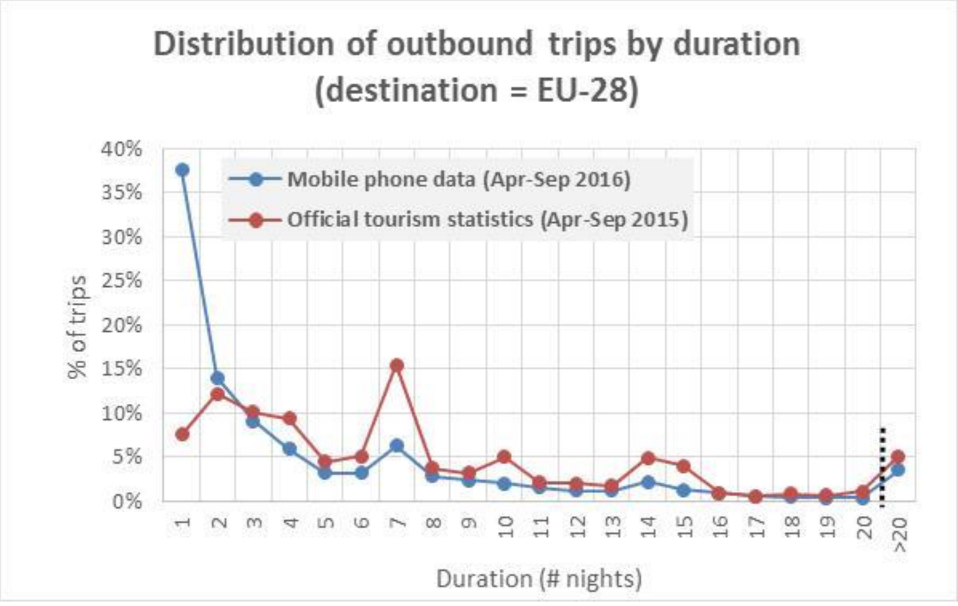
Use of big data sources involves a paradigm shift for official statistics. Statisticians find themselves in the role of data customers instead of data producers, and have to design statistical products from existing data sources. The production of official statistics is driven by high-quality standards 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 assess big data sources for quality (see Wirthmann et al. (2014)).

While most aspects of quality are relevant when working with new data sources, comparability over time is of utmost importance for official statistics. As highlighted in section 5, series continuity is a key strength of official statistics. A shift towards new sources or methods can introduce a significant break in series. A shift towards big data is no different in this respect. Even if the break in series can be communicated as an actual improvement in the overall quality of the data, it will be perceived as a major inconvenience by long-standing users of the data series.

Here again, the example of mobile network operator data illustrates this risk well. 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 trend analysis, the two sources gave relatively comparable results. For instance, Figure 4 shows the distribution of outbound trips with a destination in the European Union, by trip duration, 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 lasting exactly 7 or 14 days were more pronounced for the survey-based data — possibly due to rounding bias arising when respondents do not remember the exact duration (6, 7, 8 nights?) and give an approximation (‘a week’ — recorded as 7 days). A more striking observation is that the predominance of ultra-short trips with (exactly) 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 4 am) and the memory effect in the traditional tourism surveys (the shorter the duration, the more likely the respondent forgets to report the trip).

Figure 4: Comparison of the distribution of outbound trips to the EU-28 by trip duration[12]

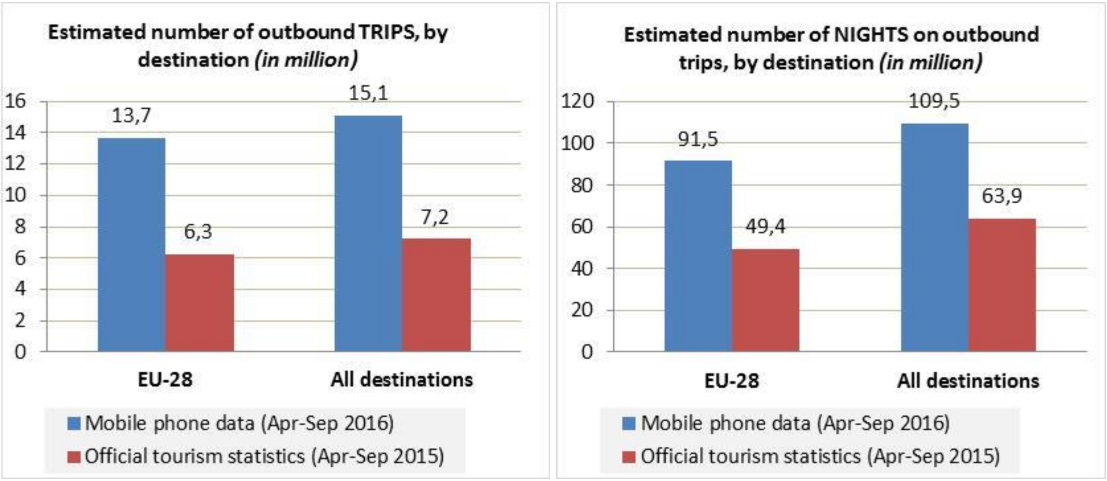


A lack of comparability or a serious break in series becomes evident when looking at absolute figures — volumes instead of indices or trends — in Figure 5. The same study compared the estimated number of outbound trips made by residents of Belgium during a six-month period. The estimate for trips obtained 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 exceeded 50 %. When looking in more detail at the data, relatively larger differences were observed for destination countries close to Belgium (neighbouring countries) than for destinations further away (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 section 5.3) may in this case be caused by the fact that the MNO (Proximus) was an early market player and may still have a relatively wealthier customer base (more likely to travel?). More probable 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[13].

[12] Source: Proximus, taken from Seynaeve & Demunter (2016), p. 13.

[13] Moreover, this particular tourism demand survey had a very low response rate (15 %).

Figure 5: Comparison of the estimated number of outbound trips and nights by destination[14]



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

However, a critical assessment of the traditional survey could equally tip the balance in favour of the 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 may 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 partial control of the data and its quality[15].

[14] Source: Proximus, Statistics Belgium, Eurostat, taken from Seynaeve & Demunter (2016), p. 14.

[15] An interesting example is the recent negotiation between the EU and the European MNOs on ending roaming charges within the EU. To analyse the extent to which EU residents use their phone outside their country of residence, the European Commission estimated the number of days spent abroad by the 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 available to enter the arena with the MNO data (and provide a counterweight to the industry’s facts and figures).

5.6 Skills

A lack of expert availability happens when, on receiving data from a new big data source, the statistical office cannot process and analyse it properly, because its staff do not have the requisite skills (Wirthmann et al. (2016)). The use of big data calls for skills in model-based inference and machine learning, 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 high demand 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 that society places in the official statistics that make use of big data. The use of their digital footprint risks being seen as intolerably invasive by the public.

The impact would be a general loss of reputation for the statistical office that might deter people from working with it. 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. Communication outreach should stress the benefits of big data usage for people, including a lower burden on respondents and improved statistical data, in a context of data security and privacy. Communication campaigns should involve relevant stakeholders with the purpose of raising awareness and informing the public about the purpose of big data usage for statistics. In this context, respondents consider transparency a key element in a communication strategy (Wirthmann et al. (2016)).

6 Conclusion

This overview of big data sources with relevance for tourism statistics makes it clear that the new sources are here to stay. If statistical offices miss the boat, others will serve user needs that official statisticians cannot or can no longer serve with the same detail and timeliness. The era of national statistical offices' monopoly on statistical information has 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 partnerships need to be set up to guarantee the continuity of data dissemination. The borderless nature of many of these sources necessitates international collaboration and knowledge-sharing — on both governance and methodology.

Besides access and skills, a key question is whether new data can address the policy and research questions in the same way as traditional statistics can. Statisticians (and users!) need to show a certain degree of flexibility or even revise existing concepts and definitions to accommodate new and richer data sources. When building a new tourism statistics system it is worth remembering that many methodological issues inherent to new data sources also exist, in one way or another, in traditional statistical techniques — and as such they affect the quality of statistics produced using those techniques. In the end, both approaches lead to an estimate, not to the true value. Remember, there is 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 provide the missing link here. Where national level data or annual data is of limited relevance to measure, say, 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. The kind of destination-level data or daily data we could once only dream of is now within reach. A combination of different sources, including traditional surveys, can yield a very powerful ecosystem of data, if connected in the right way and mutually beneficial.

The abundance of big data sources capable of capturing facets of the tourism phenomenon makes it abundantly clear that the tourism statistician in 2017 — and in the coming decade(s) — will be on 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, pp. 469-486.

Almeida de Oliveira, R. & Abrantes Baracho Porto, R.M. (2016). *Extracting web data 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 Turísticos (2008). *Memory Effect in the Spanish Domestic and Outbound Tourism Survey (FAMILITUR)*, paper for the 9th International Forum on Tourism Statistics.

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

Miao, R., & Ma, Y. (2015). *The Dynamic Impact of Web Search Volume on Product Sales — An Empirical Study Based on Box Office Revenues*. WHICEB 2015 Proceedings. 14. [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]

Sharpe, J. D., Hopkins, R. S., Cook, R. L., & Striley, C. W. (2016). *Evaluating Google, Twitter, and Wikipedia as tools for influenza surveillance using Bayesian change point analysis: a comparative analysis*. JMIR public health and surveillance, 2(2). [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]

KS-TC-17-004-EN-C

Tourism statistics: Early adopters of big data?

This paper, originally prepared for the 6th UNWTO International Conference on Tourism Statistics, gives an overview of the different sources of big data and their potential relevance in compiling tourism statistics. It discusses 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; user trust; partnerships with 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; the need for new skills, and so on.

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

For more information
http://ec.europa.eu/eurostat/

Publications Office

ISBN 978-92-79-71899-1