EUROPEAN
STATISTICAL
SYSTEM

Position paper on access to privately held data which are of public interest

Opening up new data sources for a new generation of official statistics – in light of the growing European Digital Single Market and the revision of the Public Sector Information Directive

November 2017

EXECUTIVE SUMMARY

This paper presents the position of the European Statistical System (ESS) on the issue of access to privately held data which are of public interest.

Successful pilot projects and experiments carried out by statistical offices in partnership with private data holders have demonstrated the huge potential of access to big data for statistical purposes, with significant benefits to be expected for the citizens, businesses and the society at large but also in terms of cost savings for public administrations and burden reduction for private companies. Ensuring an effective and sustainable access to new data sources will indeed allow statistical offices to produce innovative statistical products and services in a more efficient way, to increase their capacity to provide new insight into evidence-based policy-making, to lessen the burden on statistical respondents and businesses, and to boost new opportunities in the growing Digital Single Market. Overall, this will lead to the emergence of a new generation of trusted official statistics.

However, persisting obstacles and the lack of clarity concerning the conditions on access to privately-held data which are of public interest are hindering the capacity to fully exploit the potential of the Digital Single Market and the use of new data sources on a large scale in a sustainable perspective. These obstacles and unclear conditions pertain to issues such as data ownership, privacy and data protection, practical modalities of access, costs issues. They are important disincentives for data holders to share their data in particular in the context of fast evolving developments. The current obstacles need to be addressed at European level and a level playing field must be ensured for all stakeholders.

The EU must take a strong policy initiative on this subject as a matter of priority. The review of the Directive on the reuse of public sector information (PSI Directive) provides a unique opportunity to tackle the issue of data access for official statistics. The ESS invites the European Commission therefore to include in a future proposal amending the PSI Directive, relevant provisions that would establish a general principle of access to privately-held data which are of public interest, while addressing in broad terms key aspects that would ensure that such access is effective and sustainable.

1. INTRODUCTION

The European Statistical System[1] (ESS) welcomes the launch by the European Commission of a public consultation on the review of the Directive on the re-use of public sector information (PSI Directive) and more particularly its decision to further explore the issue of privately held data which are of public interest as part of this review.

The ESS agrees with the European Commission that this review should also provide the opportunity to further the goals of the Digital Single Market Strategy in the field of data economy. Therefore it provides hereafter its position with regard to ensuring access to privately held data for official statistics purposes. Clearer rules for statistical offices to access data of general interest held by private actors are needed; they will help open up new data sources and create a thriving environment for brand new statistical products and services. Through direct access to these new data sources, significant progress in terms of evidence-based policy making as regards to scope, timeliness and accuracy of official statistics will be made possible while lowering the existing burden on respondents. High-quality and timely statistics are vital for the optimal functioning of the Digital Single Market itself and access to privately held data for statistical purposes is also of great importance from that perspective. We are, however, aware that data access has to be balanced by consideration of data assets competitive value for business.

The ESS wishes also to recall its position on the issue of access to and re-use of machine-generated data held by private entities that has been provided in the framework of the consultation launched further to adoption in January 2017 of the Communication on "Building a European Data Economy"[2].

2. NEW DATA SOURCES FOR A NEW GENERATION OF OFFICIAL STATISTICS

A thriving data economy in Europe requires the growth of data ecosystems, which support the reuse of available data for public interest such as generating official statistics and which fosters the reuse of statistical data for digital services. Data are moving extremely fast, everywhere and in large quantities. As data specialists working for the common good of the society, it is vital for statistical offices to remain as close as possible to where the data emerge. These new data sources are of a very large variety: transaction data from mobile telecom operators, sensor data from personal communication devices or from smart electricity consumption meters, road traffic loops, data obtained from the internet such as social media or web-scraped data from job vacancy or real estate agencies' websites, scanner data, electronic reservation systems data, electronic data on credit card transactions, etc. These data are to a large extent generated automatically by the machines and held by private actors.

Statistical surveys and administrative registers are and will remain important sources for the production of official statistics. However, the traditional data gathering methods could be in the future enhanced and enriched by big data

[1] The ESS is the partnership between Eurostat, the national statistical institutes and other national authorities responsible in each Member State for the development, production and dissemination of European statistics. It includes the statistical offices of the EU Member States and of the EEA and EFTA countries
see ESS website at http://ec.europa.eu/eurostat/web/ess/latest-news

[2] Position paper published on the ESS website at http://ec.europa.eu/eurostat/web/ess/about-us/ess-position-papers

analytics. To achieve this, it is essential that data held by private actors can be used by statistical authorities as raw material for innovative value-added services and statistical products, which will also boost the economy by creating new jobs and encouraging investment in data-driven sectors. Increased efficiency and faster delivery by statistical offices of innovative digital products and services will be one of the cornerstones for evidence-based decisions and contribute to lessening the burden on statistical respondents and notably on businesses. Improved access to data would enable statistical offices to provide more granular and timely statistics that would be more useful to enterprises and the citizens alike. It would also mean providing valuable feedback information to the data holders or delivering more tailored statistical services to companies, which might help them in return to better develop their business model.

In some countries scanner data held by retailers or prices scraped from the internet are already used by statistical authorities to calculate the consumer price index for selected items of the sample of goods and services. Smart meters data allow the production of statistics on the electricity consumption of households or businesses very precisely, as well as the development of new related indicators such as energy insecurity indicators. Existing partnerships with telecom operators enable a number of statistical offices to process mobile phone data on an experimental research basis in order to estimate for instance the number of people present in defined areas and analyse their daily, weekly or annual mobility patterns. When combined with administrative registers and supplementary information from other European telecom operators, those data make it also possible to distinguish between resident population and other population such as tourists. Moreover, new data sources will allow the measurement of new emerging phenomena such as the collaborative economy as a whole. The annex to the document provides some concrete examples showing how some of the new data sources have been used with success to develop new statistical products and services as well as the benefits obtained, including quantified impacts in terms of savings or burden reduction achieved. These are only examples but they are the first signs of the emergence of a new generation of official statistics; a generation of statistics of higher quality and of increased relevance.

3. CLEAR RULES ARE NEEDED TO ENSURE EFFECTIVE AND SUSTAINABLE ACCESS

Opening up privately held data of general interest for re-use by statistical authorities will bring important socioeconomic benefits. Yet, statistical offices still face major difficulties in accessing new data sources. The main lesson learned from the experience gained on the ground is the absence of a specific legal basis and the lack of clarity of the conditions under which access to privately-held data can be granted for the fulfilment of public service missions.

For instance, there is a lot of uncertainty that we are facing in relation to data ownership issues such as property rights, protection of the *sui generis* right for databases or trade secrets. The protection of these rights has been very often advanced as an argument to refuse access to data to statistical offices. Although in many cases it has been possible to overcome these obstacles through a constructive dialogue, it is also obvious that risk averse data holders prefer to opt for restrictive interpretations of their own obligations towards their customers and decide not to share the data in their possession. They also fear possible negative impacts in terms of reputation particularly when the data at stake are of a personal character.

The professional independence of statistical offices and their longstanding experience in dealing with personal and confidential data is widely appreciated and perceived as providing extra assurance with regard to the further use of the data. It is deeply rooted in statistical legislation with provisions laying down strict rules on the physical and logical protection of confidential data as well as in the culture and practices of official statisticians. It is fully in line with the General Data Protection Regulation which also includes specific provisions on the processing of personal data for statistical purposes. However, this positive context is not enough to dissuade private data holders from adopting a zero risk approach when they are unclear about what they can do with the data they hold. This is even more the case when they are confronted with uncertainties due to fast evolving developments and disruptive market changes. The lack of clear rules is also the reason to withhold the data when they regard information assets with commercial value.

In parallel, a clear and predictable environment is also necessary for statistical offices to make the required investments and develop sustainable capacities to exploit the new data sources. This environment should balance the value proposition of privately held data with the role of statistical offices acting in public interest. It is therefore important to clarify issues related to the sustainability of such cooperation and possible cost recovery. As a matter of principle, data for the production of official statistics should be provided without any charge to be paid by statistical authorities. Moreover the direct access of statistical authorities could reduce the administrative burden of private entities to deliver data for statistical purposes. At the same time, a contribution towards specific investment costs in relation to data collection could be envisaged in exceptional cases. Possible incentives could also take the form of aggregate customised information derived from the data by the statistical office and provided in return to the data holders consistent with the principle of equal access to all users.

More clarity could also be obtained if adequate procedural safeguards were put in place. These could provide the guarantee of a fair use of the data by statistical authorities and the assurance that access and reuse of privately-held data of general interest remain proportionate and preserve the business models and the competitive position of the data holders. These safeguards could take the form of a structured dialogue with the private operator, whereby the potential use of data for either experiments or routine production of statistics as well as the technical feasibility would be further explored and demonstrated, including the cost-benefits impacts. It could also be envisaged to involve a third party in this assessment. Additional safeguards could consist in defining the scope of the use of the data or to limit it in time. Possible limitations for a further re-use of the data that is not strictly for internal statistical purposes could also be established.

Finally, concerning the practical modalities to access the data, our experience shows that such access can take a large variety of forms, ranging from access to raw data as collected by the data holder to access to data processed on the basis of algorithms specially designed and provided by statistical experts. Using algorithms provided by the statistical offices could not only reduce the cost of making data available but also represent an additional safeguard with regard to privacy. In general, data holders should be entitled to give access to their data as they are and should not be required to provide them in any specific format. It is indeed generally agreed that the burden on data providers needs to be minimised and data shall be used in their original format. The same approach applies regarding the metadata; without the relevant metadata, data will be of limited use for statistical purposes. It is therefore essential for statistical offices to access these metadata in order to be able to ascertain contents, quality and usability of the data while avoiding any extra burden on data holders and eventually leaving the details open for bilateral discussion depending on the data concerned. More generally, statistical offices are committed to transparency in the methods, algorithms, techniques, tools, etc. they use and this overall commitment should represent an additional incentive for data holders to share their data.

4. TACKLING THE ISSUE IN THE CONTEXT OF THE REVISION OF THE PSI DIRECTIVE

Replying specifically to the relevant questions put forward by the European Commission on the issue of access by public sector bodies to data of public interest coming from private sector entities (Part III of the public consultation on the review of the PSI Directive), the ESS position is that access to such data and their use by public authorities for reasons of public interest should definitely be allowed and organised in such a way that effective access is ensured.

Motivations or incentives on the side of private entities for sharing data of public interest with public authorities need to be considered as well. Clarity and legal security on conditions of use of privately held data is no less important for them to engage with public bodies. Private entities are keen to contribute to the common good and assume their corporate social responsibility. They also require clearer conditions for data sharing with public authorities and recognise that it could foster the data economy and eventually enrich their own offer of products and services.

Pragmatism should be the preferred approach concerning the modalities of data access: the most suitable mode is the one that suits both sides the most. This depends on the nature of the data to be accessed and their intended use by statistical offices. It is largely a matter that can be discussed and agreed upon at a later stage, keeping in mind the necessity to accommodate both sides' needs and constraints. Again, what finally counts is that practical modalities ensure effective access and reuse of the data by statistical authorities in each specific case.

As to which specific legal measures must be put in place to enable data access and use by public sector bodies/statistical authorities, the ESS considers that a comprehensive framework setting out common rules at European level would be an ambitious and necessary option. A strong consensus is needed among all stakeholders before arriving at this point and a lot of joint work would need to be done to achieve this ultimate goal. However, the issue of access to data of public interest cannot be left unanswered now as the risk of fragmented approaches across the EU is increasing, making it even more difficult to address it in the future. A more practical avenue would consist at this stage in affirming in EU law a general principle of access to privately-held data which are of public interest and addressing in broad terms the main elements for such access to be effective at operational level.

The ESS would therefore invite the European Commission to tackle these issues in the context of the revision of the PSI Directive. A possible proposal by 2018 to amend the PSI Directive will indeed provide the right place and the right moment to make real progress in this domain even if limited to the establishment of a principle and the clarification of some operational aspects. Reversing the concept of access and reuse of public sector information as set out by the PSI Directive would lead to the recognition of a principle of access by public authorities to privately-held data of general interest (so-called 'reverse PSI approach'). It would only be logical to introduce such a reverse PSI approach in an amended PSI Directive while also possibly extending its scope to data held either by other public sector bodies currently excluded or by private entities fulfilling public sector tasks or by educational and research establishments. The objective should be to create a thriving data ecosystem with a virtuous circle for data sharing among as many private and public sector entities as possible and highlight the mutual benefits for both sides for an increased exchange of data in the digital environment. It would also allow businesses and private organisations to better embrace corporate social responsibility in the field of information.

5. CONCLUSION: THE EU MUST TAKE ACTION NOW

As long as a general principle of access to privately-held data of public interest is not legally recognised and there is a lack of enough clarity regarding the conditions governing such access, statistical offices will remain prevented from exploiting the full potential of the Digital Single Market and the new data sources on a large scale in a sustainable perspective. This situation hinders our capacity to move away from limited experimentation based on ad hoc and voluntary cooperation agreements with data holders to the systematic use of these new data sources leading to a new generation of trusted official statistics. In that vision statistical authorities could act as statistical information hubs providing factual, timely and appropriate information to a wide variety of users in a fast changing world at the service of the society as a whole. The ESS has already supported a number of successful pilots and experiments and invested significant resources targeting the reuse of privately held data for statistical purposes; the momentum should be seized now together with the growth of the European data economy.

The initiatives that have been taken at national level remain limited and fragmented across the EU. The current obstacles cannot be effectively removed in each Member State in isolation, but need to be addressed at European level. There is also evidence showing that the diversity of national legislation and practices may lead to loopholes in some sectors which goes against transparency, fair competition and equal treatment of economic operators. The EU should therefore ensure a level playing field especially for those operators who are active within the single market.

The EU must send a strong signal and as a priority decide on a policy initiative in the form of a regulatory intervention even though it is limited in scope and intensity. Tackling the issue in the context of the revision of the PSI Directive through a reverse PSI approach would be a first step in this direction. This policy initiative could be accompanied by soft law measures. For example, guidance could be proposed at European level in support of regulatory intervention to incentivise businesses to share their data or the development of standards and technical solutions for reliable identification, processing and access to data.

ANNEX

Scanner data

Mobile phone data

Internet platform data

Data provided by electricity dealers

Credit and debit cards data

Smart meters

Web scrapped data

France: Scanner data for Consumer Price Index

Objective

Scanner data owned by retailers will be substituted for that collected in shops for industrial food products, household cleaning products and non-durable health and beauty products sold in food stores. They will be used to calculate the CPI in 2020. Since 2010, experimentations with 4 voluntary retailers representing 30 % of the market have been conducted. These experiments have enabled Insee to identify and cope with methodological and IT problems (comparison of the CPI calculated from these scanner data with usual CPI, how to transmit, receive and control the data, how to deal with sales, which aggregation of the data to compute…). Consultation with all the retailers has been organised in June 2016 to explain the project and since last October 2016, the national statistical law has been modified in order to make it possible for Insee and other national statistical authorities to access private data for the exclusive purposes of drawing up statistics. A specific decree and an implementing order dedicated to scanner data were signed in March and April 2017 and Insee has chosen to sign a bilateral agreement with each retail chain. These bilateral agreements make it possible to discuss again with each retail chain, in order to explain once more the objectives of the project and the fact that the costs for them will be null because they (for the great majority of them) already send their data to a consultancy that will transfer the data to Insee.

Results achieved

Accessing these scanner data will allow Insee to produce new statistical products and improve the quality of the standard CPI. For example, more spatially detailed statistics will be published such as regional indices but also more functionally detailed statistics such as indices for specific market segments (for example, organic products). These scanner data will also provide precise knowledge of the structure of household consumption in supermarkets and hypermarkets, enabling accurate calibration of the shopping basket. The precision of the CPI will be better and the quality effect in the CPI (due to the replacement of products) will be more easily evaluated. Around 20 % of the data collection with interviewers will be cut down.

Success factors/lessons learnt

Negotiations/consultations with the retailers are necessary to explain the precise objectives of the project. The new legal national context is facilitating these negotiations/consultations. Experimentations are necessary before launching the production process, to learn how to use these data in practice, to conduct methodological studies and to build the right IT infrastructure.

Website or references on the internet:

https://www.insee.fr/en/statistiques/2912652

https://www.bundesbank.de/Redaktion/EN/Downloads/Bundesbank/Research_Centre/Conferences/2017/2017_05_10_ottawa_group_09_2_paper.html?__blob=publicationFile

https://www.istat.it/it/files/2015/09/6.2-WS-Scanner-data-Rome-1-2-Oct_Leonard-Real-sale-prices-vs-displayed-prices.ppt

Italy: Scanner data for CPI/HICP: the key role of the partnership with the association of large scale retailers

Objective

The reason to look for a partnership within the scanner data project for CPI/HICP compilation lies on the evidence that also the large scale retail trade distribution in Italy shows some relevant aspects of fragmentation in comparison with other European countries. Consequently, Istat looked for a single "door" through which coming into contact with the chains of the modern distribution. This "door" was identified in *GS1 Italy (Indicod - ECR*) representing both GS1 (international body coordinating the dissemination and implementation of EAN codes) and ECR Europe (organization disseminating in Europe techniques, tools and methods of strategic and operational interfacing between manufacturers and retailers and between them and the final consumer). Groups of companies of the distribution of consumer goods operating in Italy are represented in *Indicod - ECR* by the *Association of Modern Distribution* (*ADM*). This is the main association of large scale retailers (over 900 associates and over 32,000 outlets) operating in close cooperation with organizations representing the sector (*Federdistribuzione, ANCC, ANCD*).

In November 2013 Istat illustrated to the representatives of the modern distribution the main features of the scanner data project in the field of estimation of Italian inflation and the board of *Indicod - ECR* accepted to open the collaboration entrusting *ADM* of coordinating this activity. In January 2014 some mutual commitments were agreed for the delivery of the scanner data to Istat for the test phase. A sort of triangle was established: Istat sent a formal request to the most important chains of the retail trade distribution, asking them for scanner data, and these chains (in a first step Coop Italia, Conad, Selex, Esselunga, Auchan, Carrefour, 57% of the turnover of modern distribution) authorized *Nielsen Italy* to send Istat the data; at the same time, Istat reached an agreement also with *Nielsen Italy*.

This way of proceeding was adopted also in the following years enlarging the borders of the partnership within the main framework of cooperation Istat-ADM. As a consequence starting from the data of December 2015, Istat has been receiving scanner data coming from a sample of more than 2,100 outlets belonging to 16 chains of retail trade modern distribution that represent more than 90% of the turnover of hypermarkets and supermarkets in Italy.

Results achieved

All the ongoing activities are monitored within the framework of the partnership with *ADM*. Information received by *Nielsen Italy* concern weekly data (turnover and quantity) on grocery products referred to each of the more than 2,100 outlets. Since 2017 (with data available since December 2015) the coverage of the entire national territory (107 provinces, 2,100 outlets of 16 GDO chains) is completed and the objective is to start releasing in January 2018 the official CPI/HICP estimates using scanner data.

Main benefits brought by the collaboration consist in the possibility to:

- achieve relevant improvements in terms of data quality and efficiency of the statistical production process;

- respond to the growing demand for information on consumer prices with new products (i.e. infra-national indexes, spatial comparison of prices);

- reduce the burden on the Municipal Statistical Offices in charge of traditional data collection;

- realize the desired centrality of the HICP.

Success factors/lessons learnt

Istat in setting the collaboration focused the attention on the mutual benefits that the information on inflation would have obtained from scanner data and on the opportunity for the modern distribution to identify its own role and contribution to the price dynamics of the general index.

Nevertheless, this collaboration still presents some risks because it is not formally defined; thus, in the perspective of entry into production of the use of scanner data for inflation estimation, a next step is needed, that is the signing of bilateral agreements with the chains of modern distribution, containing mutual obligations and strong guarantees for the stability of the data supply to Istat.

The experience carried out on this issue by other NSIs that have already entered into production the regular use of scanner data for CPI/HICP compilation could be strategic in setting the appropriate contents of the agreements.

Website or references on the internet:

http://www.istat.it/it/archivio/168890

The Netherlands: Use of Scanner Data from Supermarkets in Dutch Consumer Price Statistics

Objective

The objective was to replace traditional manual sample-based price observation for the Dutch consumer price statistics by obtaining scanner data from supermarkets. Scanner data includes quantities sold and the corresponding values on a very detailed level. The advantages foreseen were a substantial reduction of data collection effort and costs as well as significantly higher quality due to the improved coverage and level of detail.

The large scale use of scanner data was implemented in the context of a larger redesign of the Dutch Consumer Price Index (CPI). The national data collection regulation allows for collecting scanner data. Many supermarket chains were indeed willing to co-operate and supply scanner data on a regular basis. Statistical methods were developed for integrating the new data source in the CPI production process.

Results achieved

Data from ten supermarket chains are currently utilized. They have market coverage of 90% and account for more than 13% of the CPI-weight. The manual price observation for these supermarket chains has ended completely, saving more than 2000 supermarket visits a year. Supermarkets are very enthusiastic about this, as these visits were felt to be very intrusive. They also appreciate the feedback of price indices for their own chains.

The statistical production process has been simplified. Important quality gains were achieved. Price measurement has become much more accurate, since realised prices are observed instead of prices that are advertised, which ignore for instance the use of loyalty cards. The much more comprehensive volume information allows for improved CPI weights.

As a collateral advantage, specific analyses immediately responding to events have become possible. As an example, the large effect of the recent discovery of Fipronil in eggs on sale could be followed on a weekly basis by using scanner data. This result got considerable public attention, including in the evening news on TV.

In addition to the scanner data for supermarkets, transaction data has been acquired from other branches, such as package holidays, motor fuels, drugstores and DIY-stores. This data is used in the monthly process of the CPI. Statistics Netherlands also invested in bulk web scraping and in techniques to support manual collection of prices on the internet. Through this modernisation of the price collection process, it was possible to reduce the number of visits of price collectors to shops with more than 85% per cent in the time period 2002-2017.

Success factors/lessons learnt

Statistics Netherlands started using scanner data in 2002, and has gradually improved its methods, which can be applied by other NSIs as well. Knowledge and experiences are being shared at international conferences and ESS statistics courses (ESTP courses). In fact, several countries are now using scanner data for official statistics, with similar gains. However, a key success factor is the legal access to scanner data. This is a necessary condition. In the Dutch case this condition is met, but lack of access is holding back other countries.

Another key success factor was the support of stakeholders. As a member of the Dutch Board of Survey Respondents said: "Using or not using scanner data is not a real choice. If you don't, you lose your relevance."

Website or references on the internet:

http://www.stat.go.jp/english/info/meetings/og2015/pdf/t6s11p33_pap.pdf

https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.22/2016/Session_2_Netherlands_on_the_use_of_internet_data_for_the_Dutch_CPI.pdf

https://www.cbs.nl/nl-nl/nieuws/2017/35/omzet-eieren-herstelt-na-forse-daling

Poland: Transaction data in consumer price index

Objective

Access to data (scanner data, transaction data other than scanner data) held by enterprises and institutions operating in the field of particular sectors of retail market to use them for the purpose of the consumer price index survey. Work is still in progress and in an experimental phase. Heterogeneous nature of the individual segments of retail market requires different kinds of transactions data and their sources. Transaction data can be used as a source of price information replacing manually collected data by price collectors at outlets or it may be a valuable source of information, enabling the ongoing improvement of the survey in the scope of sample design, weights system, data aggregation and integration methods. In order to minimize the risk of refusal or temporary break in the transmission of transaction data, the relevant provisions in the legal basis for conducting statistical survey in Poland were introduced. As regards scanner data, additional provisions in individual agreements with retail chains are included.

Results achieved

There is no hard data to provide an impact analysis in quantitative terms. Expanding the price survey with transaction data could significantly improve its quality and lead to an increase in precision of estimating price indexes. Moreover, the risk of errors resulting from manual price data collection in selected segments of retail market will be reduced by the automation of the process of obtaining large data sets, their analysis and processing.

Implementation of transaction and scanner data in the index production process is very expensive and time-consuming. These new data sources require new algorithms for, among others, data validation, processing, storage, and extending the IT infrastructure dedicated to retail prices. The costs incurred by the data provider may depend on the required level of detail of the data. Generating dedicated reports is connected rather with staff costs. Processing detailed data may generate material costs (ex. additional software, hardware costs). On the other hand, for some companies visits of price collectors are more burdensome, therefore they are more willing to cooperate with the CSO as regards central transfer of data.

Success factors/lessons learnt

The implementation of transaction data in HICP compilation processes poses a number of challenges (many common concerns shared with other countries). Collaboration with the data owner is a time consuming process as regards establishing organizational and technical rules for cooperation. The scope of the data depends on the willingness of the data provider to cooperate with the NSI (data are provided free of charge). Each data provider as regards retail prices requires individual approach (even within the same sector). It is essential to access the metadata to ascertain content, quality etc.

France: Mobile phone data

Objective
A convention with Eurostat and a research unit from Orange allows Insee to get access to CDR data (reference year 2007) to conduct experimental research work on particular subjects (the evolution of the resident population throughout the day, mobility of people within local areas…). The detailed data are stored within Orange; the metadata are available as well as the technical background for mobile phone data. Different experimentations have been conducted (see the references below) and some are still ongoing. One of the objectives is to learn how to use these CDR data for statistical studies: the algorithm to detect where people live is important as well as the way to characterise the different flows of people from one location to another one. Furthermore, as these raw detailed data are very sensitive, it would be preferable for NSIs to get access to more aggregated data, less confidential. The experimental research work enables Insee to be more familiar with the structure of the data and eventually to propose an aggregated format that would be interesting for statistical usages.

Results achieved
A direct benefit of this project is that Insee now works on a regular basis with a research unit from Orange. Potential usages of these data are huge, especially to produce data and analyses on very small areas (neighbourhoods for example) which are important for our regional units.
For estimation of the number of tourists using mobile phone data, it could be interesting to get data from other European mobile phone operators.

Success factors/lessons learnt
Again, experimental work is necessary to learn how to use the data. One major difficulty which is not easy to overcome is how to extrapolate the results obtained with the data from one operator (Orange) to the whole population. The market share of Orange may differ from one local area to the other one and also between different categories of the population (for example, young people). A common legal framework at the European level could facilitate the development of these new statistical products; as many operators are European groups with subsidiaries in different countries, negotiations with European mobile phone operators could be easier at the European level than at the national level.

Website or references on the internet:

https://www.conference-service.com/NTTS2017/documents/agenda/data/abstracts/abstract_88.html

https://www.conference-service.com/NTTS2017/documents/agenda/data/abstracts/abstract_97.html

France: tourist accommodation offered by individuals via the Internet

Objective

The European regulation on tourism statistics organizes the follow-up of tourism activity in "conventional" collective establishments run by companies. However, in recent years, rental activity through Internet platforms for private tourist accommodation has increased significantly. Therefore, INSEE launched in 2016 an experimentation to follow this segment. A first study presenting the method and results for 2015 and 2016 was published in February 2017. Starting in February 2018, a follow-up will be introduced in the quarterly publication, which until now covers only the professional sector, as experimental indicators.

Results achieved

The data are collected thanks to a specific partnership with a dozen major internet platforms. The collection cost is low, since the information is centrally available through the platforms. Each of them transmits aggregated data tables to INSEE, defined uniformly. After defining an estimation methodology for some of the data, each quarter INSEE plays a role of trusted third party by controlling and aggregating these data. INSEE hands out to the platforms global framing information. This enables to answer to the request of an objective measure of the phenomenon extent and also gives each platform the opportunity to estimate its market share with respect to all of these platforms.

The data compiled through this experimentation indicate that this segment of the offer represents 16% of tourist attendance in 2016, with a very strong dynamic (+ 30% over one year). If these figures are taken into account, then tourist numbers in France should be revised significantly upwards: the increase would have been 2.6% in 2016 across the entire scope, whereas it decreased by 1.3% when considering only the segment supplied by companies.

Success factors/lessons learnt

These data were collected on a voluntary basis, without following the traditional legislative framework of statistical data collection. The platforms agree to deliver aggregated data but on the other hand, they are reluctant to share their database, or even a sample. The level of data disaggregation is still rather low: we are currently getting decomposition between the Ile-de-France (Paris area) and the Province and hope to get a more detailed decomposition according to the 13 French metropolitan regions, but not beyond.

Setting up this partnership has required many exchanges at the federation level of the platforms (UNPLV), and many direct contacts with each of them. It remains fragile because of its non-binding nature. However, we remain confident about the possibility of convincing them to follow up with this experimentation. As these actors are international, this example could probably be extended to the level of other member states.

Website or references on the internet:

https://www.insee.fr/fr/statistiques/2589218 and https://www.insee.fr/en/statistiques/2856980

Poland: Electricity dealers (enterprises selling electricity)

Objective

Actualisation of the social Survey Frame. Personal and address data are processed in the Social Survey Frame. The population set of addresses with permanent and temporary residents is used to draw samples for social surveys. It is particularly important that address data are up to date. The Central Statistical Office of Poland has decided to obtain address data and energy consumption data from enterprises selling electricity. The exact scope of data is specified with the Programme of Statistical Surveys of Polish Official Statistics developed by the Council of Ministers. Data sets received from providers contained gaps in address data, especially regarding identifiers of administrative divisions of Poland. It caused problems with linking them with Social Survey Frame.

Currently, analytical work is in progress. Work on defining energy consumption levels is carried out in order to identify vacant and probably unoccupied apartments. That will allow regular update of the Social Survey Frame.

The elaborated methodology is intended to be verified by information obtained from interviewers about inhabited addresses during realisation of surveys. Finally, in the Social Survey Frame unoccupied apartments will be marked.

Results achieved

Marking inhabited addresses in the Social Survey Frame will allow to exclude them from the frame used to draw samples for social surveys. Such solution will decrease costs of interviewers transport and increase the percentage of responses for conducted social surveys.

Success factors/lessons learnt

The close cooperation with enterprises selling electricity is necessary. It helps with creating standards for preparing databases and giving support for data transfer. This project could be replicated in other countries.

Portugal: Use of credit and debit cards' data in the National Accounts

Objective

In the national accounts it is necessary to estimate final consumption expenditure by detailed product categories on an annual basis. Household Budget Survey provides this detailed information for resident households, but only every five years. Detailed data from credit and debit cards movements, broken down by cards' issuing country and by code of the corporation where the expenditure takes place enables estimating final consumption expenditures of households on a detailed basis.

Results achieved

Credit and debit cards' data allowed for estimating final consumption expenditures of households by detailed product, both for resident and non-resident households. It is also used as reference for final consumption expenditure in the compilation of supply and use tables product by product. As this information is available very quickly after the reference period, it is also a very important piece of information for rapid estimates, notably for quarterly consumption expenditure and GDP. This information can also be used on other domains, like tourism statistics (base statistics or satellite account).

The main benefits were increased quality of the national accounts data and greater detail of the information. Until now, no reduction of costs occurred directly with the use of this data. However, this can occur in the future if it is decided to stop the direct collection of data of a specific survey. This information may also be used for calibrating the results of some surveys, allowing for the simplification of the questionnaires.

Success factors/lessons learnt

The key factor for success in accessing this data was the use of somewhat aggregated data. No individual information is received, but instead specific outputs defined based on the needs evaluated by Statistics Portugal.

In the discussions between Statistics Portugal and the institution owner of this data, it was realised that accessing individual data could be a major issue. Therefore it was considered the option of tailored outputs.

The institution providing the data is assuming the (direct and indirect) costs for providing these specific outputs on a monthly basis.

Estonia: Use of smart electricity meters to produce electricity consumption statistics

Objective

Data from smart electricity meters are used to complement and replace data collection on consumption of electricity by households and by businesses. As part of the ESSnet Big Data activities, Statistics Estonia did a study on the feasibility of using smart meter data for producing information on electricity consumption.

The roll-out plans of smart meters in Estonia expect full coverage in Estonia by 2017. Electricity consumption is one of the most important parts of energy data to create energy balance sheets and statistical data. Access to this data provides an opportunity to reduce the reporting burden on businesses. The Estonian transmission system operator Elering AS manages the Estonian electricity system in real time. For this purpose, they built a data hub. Data for statistical purposes is acquired from the data hub.

Output is final energy consumption by economic activity, by region and monthly, quarterly and annual aggregation for businesses and final energy consumption by household characteristics as they are contained in household registers (size of dwelling, number of rooms and persons, etc.) by region and monthly, quarterly and annual aggregation. Another goal is to identify vacant dwellings and to verify real places of residence of households and related persons. From survey data, it is estimated that 20% of residents do not live in places where they are registered.

Results achieved

The production of electricity consumption statistics of households and businesses by various characteristics on a monthly basis and spatially disaggregated is feasible. Key factor for success is linking the smart meters to dwellings, households and businesses from the respective registers. Empty dwellings could be identified and the results of the survey could be replicated. However, the granularity in terms of time and spatial disaggregation can be considerably improved.

The main advantages of using the smart meter data consist in the possibility to:

- link them with other data sources and gain new knowledge,
- validate or improve current survey based statistics and
- improve considerably the production speed, temporal granularity and the spatial disaggregation of regional statistics.

Success factors/lessons learnt

The project was conducted and delivered successful results because access to the data was possible. It has been based on current legislation.

The data has to come with adequate information on the data source and the data. Metadata is crucial for understanding and correctly analysing the data.

Additional information is necessary to identify the metering point recording the final consumption. Some metering points only transfers electricity. However, this issue could also lead to new knowledge on the structure of the electricity network and the trading of electricity.

Matching of smart meter data with register data is crucial. Issues of identifying the actual consumers instead of the contract owner were recognised.

Address information should be standardized or at least harmonised.

Modelling is necessary, e.g. to assign consumption from dwelling level to single apartments, machine learning for identifying unoccupied households.

Constant relationships with data providers are required and IT systems have to be designed to receive regular data flow.

Analysing the potential of a new data source triggers improving existing data sources and enhances their quality. In addition, new applications are detected in the process, e.g. correlate patterns of electricity consumption with economic activity of businesses, use patterns for forecasting of major economic indicators, e.g. national accounting, infer from consumption pattern to type of energy usage, e.g. heating.

Portugal: Web scraping as a source of HICP

Objective
Statistics Portugal launched in October 2015 the first structured web scraping project to collect prices for the Consumer Price Index (CPI), looking for quality benefits such as cost and time reductions, and improved coverage and frequency. It was defined the following specific objectives: (1) selection of technological option for extraction, storage, processing and analysis of the data, (2) development of a web scraping prototype for CPI price collection, and (3) evaluate the data collected by the prototype from a real website, comparing it with traditional data collection of the CPI.

Results achieved
A technical solution was developed and tested for data extraction, storage, processing and analysis for an existent commercial website, comparing prices traditionally collected. Furthermore, it was possible to adopt the approach in real, substituting the traditional price collected from the outlets of the targeted company with the same data found by the web scraper prototype.

The choice fell on a well-known multinational group that designs and sells ready-to-assemble furniture, where the prices of products belonging to several classes of COICOP (Classification of Individual Consumption According to Purpose) are monthly collected manually. Statistics Portugal previously informed the company about the objectives and principles of the project, having obtained full cooperation, and maintaining a good institutional relationship.

Web scrapers tests started working in the summer of 2016. During six months, data from products were daily collected. It was detected differences in the availability of products on the website, differences in prices, as well as website crashes. The website has shown to be stable, not suffering breaks during the data collection periods. Prices collected were comparable. Following this first approach, and with the same infrastructure, web scrapers were also developed for collecting prices on the mobile phone operator's websites and on a large retailer of home appliances and electronics.

Success factors/lessons learnt
Price collection through web scraping has potential of replicability to other statistical fields as well as to other countries, as a substitute or complement approach to the traditional data collection. It should be explored and extensively studied.

The specific solution was internally developed, requiring low programming skills. The acquisition of knowledge and IT skills were generally based on open training platforms, which is an opportunity to adopt flexible ways with lower costs and time-to-result.

It was success factor involving the data provider from the very beginning of the initiative.

Multidisciplinary teams, which were composed by members with non-IT skills, have provided opportunities to increase synergies.