The issues and opportunities of opening data in the energy sector

Sylvain Moreau,
head of the Service de La Donnée et des Études Statistiques (SDES), Commissariat Général du Développement Durable (CGDD), Ministry of Environmental Transition

For: In R. Lavergne & H. Serveille, editors of *The digital and environmental transitions*, a special issue of *Responsabilité et envrionnement - n° 87 - July 2017*.

Abstract:
Owing to the growth of the Internet, gigantic files on the consumption of energy by households and firms hold a high potential for data processing. Using them for statistics does not differ all that much from the work that public offices of statistics are used to doing with administrative files — apart from the fact that they contain a wealth of time series and geographical information on a scale never previously attained that can be used to design ever more numerous and detailed indicators. Following the application of a new act on energy in France, public statisticians have a key role in opening the data transmitted by energy producers and distributors to the public. Foreseeably, this will have a major impact on systems of statistical information in the field of energy and on the work methods of public offices of statistics.

The phrase “big data” has been increasingly used during the past few years in the information sciences.[1] The volume of data produced by the giants of the Internet, along with certain scientific disciplines (genomics and astronomy in particular), has swollen massively. These gigantic data pools originate in the advances made in techniques for storing and processing large volumes of a variety of data (mainly text and images) that are being produced in an ongoing stream, whence the phrase “three Vs” for referring to the volume, variety and velocity of this data flow. Specific tools and methods (such as the algorithms for making on-line recommendations on websites in the retail trade) are being used to process these big data. It is now possible to have access to very detailed data (in particular at the geographical level) for responding to quite precise demands for information that used to remain unsatisfied.

In short, the phrase “big data” refers to quite different phenomena: on the one hand, the sources of data now available that did not use to exist and, on the other, the methods for managing and analyzing these data, which might be in non-standard formats.

Public offices of statistics are, of course, concerned with these new data sources. Some national institutes of statistics (such as ISTAT and CBS in Italy and the Netherlands respectively) soon worked out a strategy for using big data and have even launched experiments. Several international work groups have been set up — in particular by Eurostat and the United Nations, notably the United Nations Economic Commission for Europe (UNECE) — to adopt a joint position for public offices of statistics on processing data from these new sources, to set priorities and to pool investments. At the French National Institute of Statistics and Economic Studies (Institut National des Statistiques et des Études Économiques, henceforth INSEE), the work group New Sources has dwelled on this topic, which figures in the program being drafted for INSEE in 2025. What position should public offices of statistics adopt?

[1] This article has been translated from French by Noal Mellott (Omaha Beach, France). The translation into English has, with the editor’s approval, completed a few references.

As of 2011, INSEE, like certain public offices of statistics outside France, launched a program (Cash Register Data) for integrating data from big retail chains in the consumer price index. This program has cast light on the questions to be settled before processing such data.

— The first question has legal ramifications: how to maintain long-term access to big data for public offices of statistics? When the program started, no legal framework guaranteed this access. Since then, a relevant article in the Digital Republic Act has settled this question.[2]

— The second question is technical: how to manage and store streams of large volumes of data? The infrastructure must be adapted. In response, INSEE has set up a storage platform for big data.

The particularities of big data

When talking about big data, what first comes to mind is the data used for managerial purposes by private operators. Some of these data sources could be tapped by public offices of statistics, for example: the invoicing data produced by big retail chains (already used by INSEE) or by mobile telephone companies; or the information from smart electricity meters. In some respects, the problems of using such data are similar to those of using administrative data, which public offices of statistics have been processing for a long time now. These problems concern the quality and representativeness of the data, the fit between the data available and the phenomenon to be measured, the duration of the data source, the lack of control over changes in the contents and format of the data, and the very size of data pools. Public offices of statistics are familiar with these problems. INSEE and other ministerial services of statistics have acquired several decades of competence in handling them.

The big data produced by more recent sources has, however, particularities that make them different from the data from administrative sources, namely:

— THEIR ACCESSIBILITY. Big data are the property of private operators, who often exploit them for their economic value. The previously mentioned Digital Republic Act has made it somewhat easier for public offices of statistics in France to be authorized to obtain access to these vast pools of data. This access is, however, severely restricted to conducting the statistical surveys required by the law. In this case, access is authorized under certain conditions: *a)* this method of data collection is adapted to the needs of these surveys; *b)* it has, in comparison with other collection methods, advantages in terms of the quality of the data produced or of the cost for the public office of statistics or for respondents in the survey; and *c)* the information extracted from these data bases cannot be used for any other purpose than the specific statistical survey that justified access to the bases.

— THE DIFFICULTY OF HANDLING BIG DATA. Some big data are closely related to the behavior patterns of the households from which they come. In other words, they might be so tightly correlated to what is to be studied that it is hard to quantify the latter. For example, the data might be useful for describing consumption patterns, but of little use for calculating the percentage of households concerned. Access to the data bases of car-pooling websites make it easy to obtain data about the number of rides, the routes or even users' profiles; but this type of data does not yield information about the relative share of car-pooling in transportation as a whole or about the motivations for sharing rides.

[2] Act n°2016-1321 of 7 October 2016 for a "*digital republic*" available at https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000033202746&categorieLien=id.

— THE SENSITIVE NATURE OF THESE DATA. These data are often personal, telling about individual behaviors, and are, *ipso facto*, sensitive. Retail businesses are already designing new services to profile customers by crossing in all sorts of ways a large volume of customer data. Keenly aware of these problems, the National Commission on Informatics and Liberty (Commission Nationale de l'Informatique et des Libertés, henceforth CNIL) has modified its legal theory and work methods, the intent being to safely protect individuals but without curbing innovation.

Why are public offices of statistics interested in big data?

A first reason that public offices of statistics might find it worthwhile to work on big data is to improve and enrich the current production of statistics. Attention has often been drawn to an important issue for public statistics: reducing the time needed to release indicators. Using immediately accessible data obtained through streaming seems offhand to be a natural solution for reducing production time and, thereby, the waiting time for indicators. Smart meters, for example, will be able to produce more frequent estimates of electricity or gas consumption. Through FluxVision, Orange Business Services has promised to convert four million mobile data per minute into statistical indicators of the number of visitors by geographical area and population movements.

This mass of available data is also of interest for producing indicators with a finer grade of information on subcategories or subgroups. For example, the data from cash registers can serve to systematically generate the average price of products or regional price indexes. They allow for more precision without significantly increasing costs (for example, using mobile telephone data to estimate the transportation time per type of transit). These big data provide fuller, even exhaustive, information, which is much more detailed (in particular on a geographical scale) and, for local stakeholders, more relevant. Furthermore, these new data sources could reduce the cost of surveys, in particular of collecting data, and thus make savings. Of course, the situation is much more complicated than just suggested. The reduced cost of surveys has to be compared with the investments necessary for processing massive quantities of data that are, by nature, less easy to handle. For instance, the experiment with using cash register data entailed purchasing catalogs for processing the data and adapting the infrastructure for processing massive streams of data. According to a recently published assessment by Statistics Norway, processing big data did, in effect, reduce the cost of surveys; but it also increased the internal resources devoted to producing the price index. This increase is to be set down not only to the greater complexity of the data used to calculate this index but also to a "quality effect", since more indicators can be produced. As already pointed out, the available data are structurally richer and fuller, and, therefore, have a much higher potential output when processed.

Using big data might deeply modify the system of monitoring for certain statistics. Instead of conducting regular surveys, a few surveys might be carried out to follow up on the findings obtained from processing such highly detailed data. The same information could thus be provided as now but while broadening the spectrum of data available to, in particular, local authorities and stakeholders.

Energy, a field with potential for this work on statistics

Energy is a branch of the economy that has a rich potential for the processing of big data. Part of this potential has already been tapped.

The energy sector has been transformed in recent years. The players in this sector have changed: they are more in number, and are now diversified. Technological advantages now (or will soon) make data available on real-time consumption with an unprecedented level of detail.

Furthermore, the act of law on the *"energy transition for green growth"* establishes the principle of opening the data collected by energy networks.[3] In a single year after the passage of this act, the situation has completely changed. This act requires that the parties involved in transporting, transmitting or managing energy communicate to the public administration data related to the total annual consumption of electricity, natural gas, hydrocarbons and heat by branch of the economy and at the neighborhood scale. The administration must release this data once it has dealt with problems of confidentiality (mainly the protection of personal data). The first objective is to help consumers control their consumption and to help local authorities fulfill their new assignments in the field of energy: the drafting of regional plans of sustainable development and of equality between local areas (Schémas Régionaux d'Aménagement, de Développement Durable et d'Égalité des Territoires, SRADDET) and of local climate, air, energy plans (Climat-Air-Énergie Territoriaux, PCAET).

The pivot in this opening of data is the Service of Statistics of the Ministry of the Environmental Transition. It is in charge of rediffusing the data toward local stakeholders. Since the end of 2016, the data are available on the Service's website in an open format that makes them easy to use and process.[4] Within two years, they will be available at the scale of individual buildings.

This opening of data is a lever for new uses and energy services. It will eventually upend the process whereby energy statistics are produced.

Current, statistics follow up on the final consumption of energy in the economy as a whole and by branch (agriculture, industry, transportation, housing, service sector). To break energy consumption down into data at the branch level, information from various sources is collected and processed by the SDES (Service de la Donnée et des Études Statistiques, the service of statistics of the Ministry of the Environment). The SDES completes the data with information from surveys carried out by INSEE or other public services of statistics (in particular in the Ministry of Agriculture). For information about the distribution of consumption by use in residences (by type of building) and in the service sector, the SDES obtains data from the CEREN (Centre d'Études et de Recherches Économiques sur l'Énergie).

The findings as a whole are mainly established at the national level. Work has been done on regionalizing the data. For certain branches of the economy, regional statistics draw from *ad hoc* surveys. For others, the results are estimates or the residuals of calculations. Such is the case for housing, since no statistical source has accurate enough data at the regional level. Experience has taught us that these results might be fragile and questionable.

A strong demand has been addressed to public offices of statistics for reliable local-level data to be made available. A first response to this demand is to open data on consumption, to make it available on a very small geographical scale. In general, the difficulties inherent in big data (*cf*. the beginning of this article) are irrelevant to data of this sort, since they are exhaustive but are not biased via the selection process and are accessible by law.

[3] The so-called TECV Act on the "*energy transition for green growth*". Act n°2015-992 of 17 August 2015, available at:
https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000031044385.

[4] http://www.statistiques.developpement-durable.gouv.fr/accueil.html.

This response does not answer all questions. The uses of energy are still a blind spot. Once the data are available at the scale of buildings however, we could then imagine crossing them with other sources (such as the SIREN data for identifying businesses) so as to obtain a detailed description of energy consumption by branch of the economy. This will not exempt us from the need for other statistical tools (surveys, administrative sources, simulations) for adding value to these data and completing the already available information.

The deployment of smart meters for measuring gas and electricity consumption will eventually open interesting perspectives for following up on energy consumption. Data from the meters could even inform us about the uses of the energy consumed. They could be used to make simulations for households. The nomenclature for industry could be honed. Furthermore, it would be possible to follow up on consumption during the year on an extremely small geographic scale. Finally, “energy poverty” could also be studied.

With respect to opening data, the position adopted by the offices of statistics in government ministries differs somewhat from INSEE’s. The data made available by public offices of statistics are seen as authoritative, as having a proven reliability. In the case of opening data on energy consumption however, the time needed to collect the data from operators and then to make them available is too short to ensure reliability. Only the most egregious errors will have been corrected. The probability of data entry errors is very high. Though undetectable at the national level, they can have major effects at the regional or local levels. This probably requires, at the very least, that the data diffused be certified as “examined” or “unexamined”. If the work on statistics is eventually generalized in this way, thought will have to be devoted to setting up procedures for making the data reliable. For example, the tasks of correction could be decentralized at the regional level (DREAL, Directions Régionales de l’Environnement, de l’Aménagement and du Logement) and then eventually at the local level. This would require a formal procedure for feedback to the producers of the data. This calls for rules and procedures for working together on big data.