Big Data and Audience Measurement: A Marriage of Convenience?

Lorie Dudoignon*, Fabienne Le Sager* et Aurélie Vanheuverzwyn*

Abstract – Digital convergence has gradually altered both the data and media worlds. The lines that separated media have become blurred, a phenomenon that is being amplified daily by the spread of new devices and new usages. At the same time, digital convergence has highlighted the power of big data, which is defined in terms of two connected parameters: volume and the frequency of acquisition. Big Data can be as voluminous as exhaustive and its acquisition can be as frequent as to occur in real time. Even though Big Data may be seen as risking a return to the paradigm of census that prevailed until the end of the 19th century – whereas the 20th century belonged to sampling and surveys. Médiamétrie has chosen to consider this digital revolution as a tremendous opportunity for progression in its audience measurement systems.

JEL Classification : C18, C32, C33, C55, C80
Keywords: hybrid methods, Big Data, sample surveys, hidden Markov model

** Médiamétrie (ldudoignon@mediametrie.fr; flesager@mediametrie.fr; avanheuverzwyn@mediametrie.fr)*

Receveid on 10 July 2017, accepted on 10 February 2019
Translation by the authors from their original version: "Big Data et mesure d'audience : un mariage de raison ?"

To cite this article: Dudoignon, L., Le Sager, F. & Vanheuverzwyn, A. (2018). Big Data and Audience Measurement: A Marriage of Convenience? *Economie et Statistique / Economics and Statistics,* 505-506, 133–146. https://doi.org/10.24187/ecostat.2018.505d.1969

During the 20th century, census has gradually declined in favor of sample surveys. The founding act can be considered to be Anders N. Kiaer's paper at the Congress of the International Statistical Institute in 1895 entitled *Observations et expériences concernant des dénombrements représentatifs*. In 1934, Jerzy Neyman published the reference article in sampling theory *"On the two different aspects of representative methods, the method of stratified sampling and the method of purposive selection"*. The growth of telephone equipment then encouraged the use of sample surveys in many fields (public statistics, politics, health, marketing, audience measurement, etc.). The end of the 20th century saw a new paradigm shift with the emergence of Big Data: a return to the census. As a major player in this digital revolution, the media sector has seen its measurement systems multiply and sometimes, inevitably, contradict itself. Médiamétrie, a benchmark institute for media audience measurement in France, has had to change its methods to take advantage of the best of each source.

The first part of the article deals with the relative advantages and limits of survey data and Big Data, with an emphasis on the notion of quality in its various dimensions. This will allow to explain why Médiamétrie has chosen to see survey data and Big Data as complementary rather than in competition with one another. Indeed, we will look at how hybrid approaches: "the mix of two data sources that differ in both nature and level to create a third, richer or more detailed one" have become the natural approach (Médiamétrie, 2010). The second part will illustrate these approaches through two operational implementations in media audience measurement. We shall begin by introducing the hybrid method used to measure internet audiences as part of the French market standard since 2012 – an example of a so-called panel-up approach (Dudoignon *et al.*, 2012). We shall finish by illustrating the so-called log-up approach used to measure the audience of special interest channels (Dudoignon *et al.*, 2014). In both cases, for Big Data to have any meaning or value, we must first understand how it is acquired, which often includes technical aspects performed to "clean" the data and process it in such a way as to create a potentially happy marriage with survey data.

Preamble: Data available in Audience Measurement

Both survey data and Big Data exist for television and especially internet media. In both cases, audience measurement is based on a panel and a semi-automatic system of measurement. In this introduction, our aim is to briefly describe the existing audience measurement systems for television and internet applied by *Médiamétrie* in France.

Internet

Internet audience measurement relies on two types of system: user-centric measurement is dedicated to tracking internet site and app audience behaviour for individuals across all of their devices. These systems are based on panels of individuals whose connections are measured using meter software installed on their computers, mobile phones or tablets that feeds data back to Médiamétrie's servers. The second type of system is called site-centric. This kind of measurement relies on the insertion of tags (Box 1) into the websites and apps of clients subscribing to the measurement and produces a total counting of the number of visits, page views and connection times.

Internet Audience Measurement on Computers

Since the home computer is often a shared device, the panel consists of a cluster sample of all individuals aged 2 years and over within a household. Therefore, the primary unit in the panel is the household and the secondary unit is the individual. Primary sampling units are recruited in accordance with the empirical quota method. Once the meter has been installed on all household computers, a pop-up screen or window will appear each time there is a connection. The secondary sampling units (individual household members) then have to identify themselves by ticking the box that corresponds to them. In September 2018, the panel comprised approximately 6,200 households with internet access *via* a computer, i.e. more than 14,000 individuals.

The scope of measurement is not limited to connections to the internet at home. In fact, for the population in employment, a significant proportion of their connections to the internet occurs in the workplace. Nevertheless, the effort required by individuals to take part in

Box 1 – **Description of Measurement Technologies**

What is a Tag?

In web analytics, a tag is an element that is inserted into each content to be measured, so as to count the number of content views. The content can be a page, an app, a podcast or even audio or video content. A code is inserted into the source code of the content. This generates a log on the third-party measurement system server each and every time a content is viewed. This then makes a total counting of connections to the tagged content possible.

What is Audio Watermarking?

A technology used for television audience measurement, audio watermarking consists of the insertion into the broadcast being measured of a mark (similar to a tattoo) that is inaudible to the human ear. This digital tattoo is inserted by a professional embedder validated by Médiamétrie. The principle is to modify the signal broadcasting the program with some additional information, without affecting sequence audibility. At the other end, the watermark is read by the TV meters connected to the TV sets owned by panelists. The mark inserted by the embedder contains identification information for the channel broadcasting the program, as well as regular markers of the broadcast time. In this way, we can differentiate between the audience watching a live broadcast, the audience watching a pre-recording and the audience watching via a catch-up TV platform.

measurement – also known as the “response burden” – prevents us from insisting that all secondary sampling units on the panel are additionally measured at their place of work (if they have a computer with internet access at work). Such insistence would likely lead to very low response rates. Therefore, the system is supplemented by an independent panel of individuals who have internet access on a work computer. In September 2018, there were 2,000 individuals on this panel, and it is linked to the preceding panel by statistical matching (Fisher, 2004).

Internet Audience Measurement on Tablets

The principle of internet audience measurement on tablets is very similar to that for computer measurement. Given that tablets are still hardly used in businesses, the scope for measurement is for the moment restricted to households. The panel consists of a cluster sample of individuals from within the recruited households. The latter must install a measurement app on all tablets used in their home and must change the settings to ensure their connections are sent to Médiamétrie’s servers. As soon as the app is launched, the user can be identified. In September 2018, the panel consisted of 2,000 households, or 5,200 individuals aged 2 and older.

Internet Audience Measurement on Mobile Phones

Unlike computers and tablets, mobile phones are devices that are primarily for personal use. Consequently, the panel is made up of individuals recruited by quota sampling. The minimum age for measurement participants is set at 11 years old, and in accordance with the constraints imposed by France’s Data Protection Act of 6th January 1978, participation by minors is subject to the consent of an adult with parental authority. Like the system for measuring connections *via* tablets, the panelist must install an app on its mobile phone. This app routes the connections to Médiamétrie’s servers. All internet traffic on the phone is attributed to the main user of the phone. Any use of the mobile phone by a secondary user is therefore, by convention, assigned to the main user. In September 2018, the panel consisted of 11,000 individuals aged 11 and older.

Measurement of Secure Connections

Participation in user-centric measurement systems begins with the signature of an agreement between Médiamétrie and its panelists. This agreement details the respective commitments of Médiamétrie and the panelists. In particular, Médiamétrie undertakes to collect panelist user data for purely statistical purposes. Furthermore, Médiamétrie undertakes to never disclose the identity of its panelists to any third party for advertising or commercial purposes. Finally, it undertakes to take all necessary precautions to preserve the security of the data collected and, in particular, to prevent any distortion, corruption or unauthorized third-party access to this data. In return, the panelists undertake to keep their participation in the survey and the means of their participation confidential, in order to avoid any attempt at influence by stakeholders, publishers or operators with an interest in audience measurement. They also undertake to install the measurement software, to log on where appropriate, to inform Médiamétrie of any change in their situation, and to agree to be contacted by Médiamétrie.

Once the agreement has been signed, the panelists authorise Médiamétrie to have full access to their internet usage data, including their HTTPS connections and their IP address. However, for technical reasons, data collected from secure connections is in some cases less detailed than data gathered from HTTP connections. For example, for the measurement of connections *via* tablets, only the domain name will be available in the logs returned to Médiamétrie's servers in the case of an HTTPS connection, whereas the full URL will be collected for an HTTP connection.

Television

Médiamétrie's Médiamat panel is the reference in television audience measurement in metropolitan France. This measurement is based on a panel of individuals consisting of a cluster of some 5,000 households that own at least one television set. All active television sets are included in the measurement scope, i.e. those that are used at least once a month to watch television. Each of these TVs is connected to a TV meter that uses audio watermarking technology (cf. Box 1) to detect the channel being watched on the TV at any time. Individuals in the household must participate in the measurement by stating that they are in front of the TV using a remote control connected to the meter. Médiamétrie's servers continuously collect the data recorded by the TV meters. Although the panelists are instructed to state the presence of all household members in front of the TV screen, only the audience results for individuals aged 4 and older are fed back.

The TV return path (Box 2) is technically possible in two scenarios: ADSL, cable and satellite set-top boxes when they are connected to the internet, and smart TVs. We should note that although most television sets on the market today are smart TVs, in reality it is still quite rare for them to be connected to the internet. In these two scenarios only, return path data are available from the operator distributing the broadcast and they indicate which channel or service the set-top box is turned onto. No measurement is taken for any usage of the television without the set-top box. For example, if the television is connected to several modes of reception – *via* DTTV and an ADSL set-top box – any programs watched *via* DTTV will not be measured.

Quality of Survey Data and Big Data

Although there is no single definition of survey data quality (Dussaix, 2008), this is even more true of data quality in general. We can, however, keep in mind that quality is a real concern for most statistical agencies and that most of these would agree that it is a multidimensional concept that is difficult to assess (Lyberg, 2012). For our discussion, we have chosen to retain the six dimensions of quality used by Statistics Canada and the Australian Bureau of Statistics. They are: relevance, accuracy, time-to-market, accessibility, interpretability and consistency (Brackstone, 1999; Institut de Statistique du Québec, 2006). We should also note that the OECD adds two additional dimensions: credibility and cost-effectiveness in their assessment of the quality of statistical output (OECD, 2011). It is not a question here of discussing the definition of the dimensions of the quality of the surveys but of proposing a comparative analysis of "survey data" vs "Big Data" on each of these dimensions.

Relevance

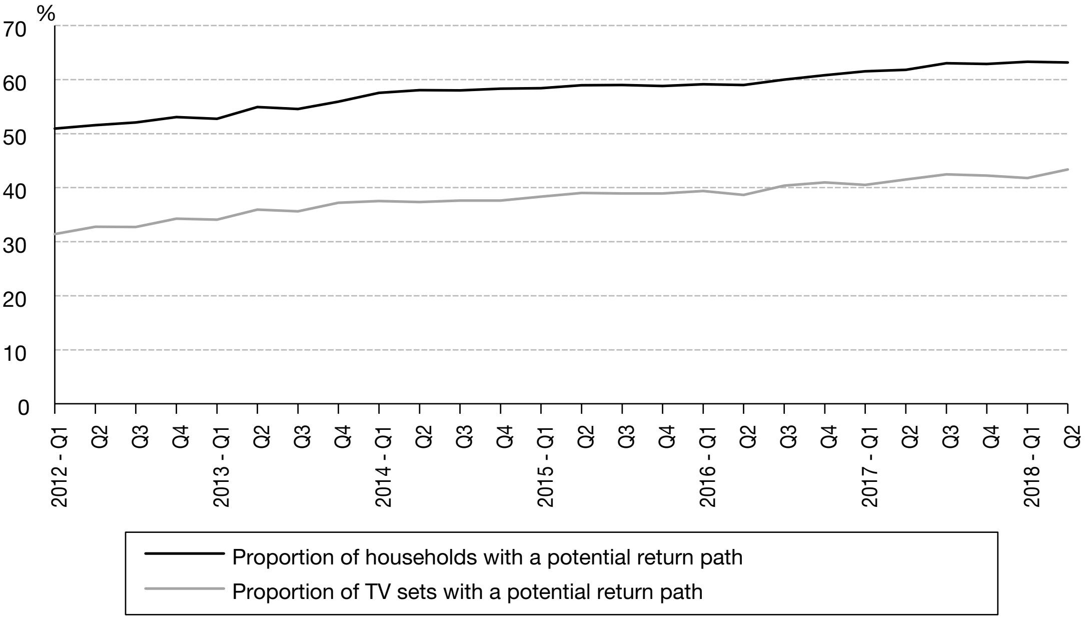
The relevance of a study or a measurement corresponds to its utility and its ability to meet the needs of users or customers. This criterion is obviously the first choice when assessing quality. The relevance of panel audience measurement is not generally called into question insofar as these systems have been designed in close collaboration with their users. In fact, for each media, a committee composed of members representing broadcasters and users, advertisers and media agencies, publishers and operators, and Médiamétrie, has been created on a parity basis. The objective of each committee is to define, orientate and validate measurements and surveys that serve as a reference for each of the media types concerned.

However, panel audience measurement cannot fully meet every need, in particular, when it comes to measuring very confidential or very fragmented usages, given that these would necessarily be poorly represented – or even not represented at all – within a sample. Increasing the sample size is clearly not a pertinent answer because the relevance of a study includes the budgetary constraints of its users. Conversely, Box 2 – **What is the Potential of Return Path Data in Television?**

A TV return path is the possibility for a broadcaster to collect some digital information back from users, about their TV consumption. The return path is technically possible for all set-top boxes that are connected to the internet, and for smart TVs. In concrete terms, this type of data collection is implemented by telecoms operators and satellite operators such as CanalSat (one of the major French suppliers of cable programs).

It is estimated that this return path is currently possible for a little over 60% of French households with at least one television set, but for barely more than 40% of television sets. In fact, the set-top box is very often connected to just one main television set and is not linked up to additional sets. This represents a potential, since not all set-top boxes that can be connected to the internet are necessarily connected.

Figure A
Evolution of the Potential of Return Path in Television



Coverage: Metropolitan France.
Sources: Médiamétrie – Home Devices.

Big Data does not fully meet the needs of users since it can identify machine usage but not individual usage. It is therefore essential to pre-process this kind of data to clean it up and transform it into meaningful information. Below are some real examples of this kind of pre-processing. They provide some valuable information on emerging and niche usages that cannot be measured by samples because of their volume. On the first criterion, relevance, the complementarity between survey data and Big Data for the purposes of audience measurement is clear to see.

Accuracy

In our context, accuracy means correctly describing the media behaviour of French people. Although it is generally acknowledged that results from surveys are flawed because of sampling errors and the problem of non-response there is a tendency to think on the contrary that Big Data is accurate because it covers the entire scope of measurement. It is nothing of the sort. Actually, as we noted above, Big Data brings in information about machines and not about individuals, which is an obvious source of error. Furthermore, if the technologies used to measure are not properly controlled, they can lead to implementation or interpretation errors. This brings us back to the pre-processing phase that should partially clean up these interpretation errors. As far as implementation errors are concerned (for example, wrong implementation of an internet tag), the best way to proceed is to install a monitoring system to detect these flaws as early as possible and to correct them before too large a volume of data becomes affected. It should be noted that this type of monitoring is also necessary for panel measurement since it uses content marking technology (web tag or audio watermarking for television) for the purposes of audience measurement.

Time-to-Market (or Speed of Delivery)

Time-to-market refers to the lag between the analysis reference period and the delivery of results. In the context of media audience measurement, this is a very important criterion. Any excessive delay in delivering results would render these results obsolete and of very limited interest to users. For internet, results are generally made available monthly and must be published in the month following the analysis period. For television, the delays are much shorter. The first audience results for daily programs are published from 9am the following morning. These results are then consolidated eight days later with the inclusion of time-shifting viewing in the seven days following the original broadcast.

For survey data, site-centric data or return path data, when automatic measurement technologies are used, raw data can in theory be acquired almost in real time. The freshness of the results can therefore be ensured as soon as the pre-processing and processing operations of these data are performed in limited time. In both cases, this involves the implementation of very strict, automated and industrial production processes.

Accessibility

Audience measurement results are accessed *via* reporting interfaces that are available to all subscribers. This kind of interface in particular can manage various user permissions, and thus grant access to less or more information depending on their subscription. From the user point of view, accessibility will be considered as satisfactory if the results consultation tool is both ergonomic and efficient in terms of computational and display times. Internally, our teams tasked with producing results and performing additional analysis have ready access to all of the data. Nevertheless, even in-house, this access is limited to anonymous data. Only the management and panel coordination teams have access to personal information that can be used to contact panelists.

Technical difficulties of access to Big Data are increasingly rare these days and are no longer a priority issue for development. By contrast, legal constraints oblige us to limit access to this type of data and even to reduce the quantity of information gathered. Although in the past, digital data could sometimes be collected without the knowledge of the individuals, this kind of practice is no longer possible, in Europe at least. Most stakeholders who are currently gathering this kind of (site-centric or return path) data have had to put in a lot of effort to become compliant with the European General Data Protection Regulation (Box 3).

Intelligibility (or Interpretability)

Whether for panel data or Big Data, the intelligibility of the data mainly relates to technology. It is possible to think of the raw data generated by tagging technology for television and internet (what we call logs) as hardly intelligible at all. Only after pre-processing will this data be translated into an interpretable format. The statistician obviously cannot work alone. This type of data necessitates close collaboration between the technical teams who develop the tagging solutions, the I.T. teams who collect and process the data, the statistical teams who devise the analysis, and the customer liaison teams who install the tagging solution into their websites and channels.

Although they may appear complex, media content tagging solutions can, after translation, provide intelligible data that is also easy to enrich with metadata describing the content in detail (e.g. for online video content, the ability to specify if it was a series, which season, which episode, and when the original TV broadcast was, etc.). Automatic measurement solutions that do not use tags are generally much less intelligible. Take, for instance, internet audience measures based on the capture of network traffic for a device. Over 90% of the collected data is irrelevant, since it cannot describe the behaviour of the individual using the device. The data collected actually includes all of the technical information flows, e.g. updates to software and applications, which are totally transparent for the user.

Rendering this kind of data intelligible is a real challenge, since any mistake in filtering the data usually leads to an interpretation error. With tagging solutions, it is possible to only collect

Box 3 – **European General Data Protection Regulation: The Changes Affecting Professionals**

The new European regulation which came into force on 25th May 2018 introduces or strengthens the following principles.

- Strengthening of the rights of persons: Users must be informed of the collection and use of their data. At all times, they must be able to give their consent, or object if necessary. Users have new rights: in particular, the right to restriction of processing; the right to data portability; the right to erase data.
- Responsibility of agents (data controller and processor): The regulation reduces the obligations of prior formalities at the CNIL (the French authority for data protection). On the other hand, the new regulation introduces the principle of demonstrability: the ability to prove compliance with the regulation at all times through detailed documentation of all personal data processing activity. In concrete terms, the data controller undertakes to: keep up-to-date detailed registers of personal data processing activity; to systematically carry out impact assessments before each processing activity that presents a high risk to the rights and freedoms of natural persons; to ensure the compliance of any data processors. The regulation also strengthens the sanctions to be applied against the data controller in the event of a non-compliance: up to 20 million euro or 4% of global turnover.
- Privacy by Design: The company must take into account the notion of respect for private life, beginning at the design phase of a product or application. The data controller must implement all technical and organizational measures that are necessary to comply with the protection of personal data, from the design phase and by default.
- Creation of a Data Protection Officer role (DPO): This new expert will identify and coordinate the actions to be taken within the company or organization that pertain to the protection of personal data: from internal communications to checks on regulatory compliance, as well as being the point of contact with the supervisory authority.

data that is useful, which therefore makes these solutions a lot easier to interpret.

Consistency

Without the hybrid approaches, a stakeholder could end up with several figures representing the performance of an identical content. For example, the average number of viewers for the video content over a period, and the number of set-top boxes tuned into that video content for at least one minute. These two indicators are based on different units and are not comparable, but they may alarm unaware users who see both of them published. Médiamétrie should therefore provide the necessary consistency. Firstly, by clearly explaining the concepts and indicators, as well as how to interpret them. Next, by offering solutions to reconcile these different-natured data so as to produce a consistent measure. Panel data and Big Data consistency is then the very essence of Médiamétrie's hybrid measures.

In addition to the six dimensions described above, another one that must be considered regarding Big Data is confidence (or, to use the OECD term, credibility). Some media stakeholders have installed site-centric or return path systems of measurement. As is the case, for example, of the biggest players on the web – GAFA[1] and the telecoms operators. Such players use these to offer measurement services to publishers who also use their distribution platform. As it is generally very hard to be the judge of one's own case, even if one possesses the utmost discipline and honesty, other market players will always call their credibility into question. In such a context, "proprietary" Big Data often requires certification by a trusted third party to be recognized and shared by the market. This is the role of ACPM[2] in France which certifies the number of newspapers and magazines distributed.

Some Examples of Hybrid Approaches to Media Audience Measurement

Two approaches to hybrid measurement are theoretically possible. The approach chosen depends on the user's expressed need. In the first approach, which we call panel-up, Big Data enriches the information gathered in the media survey, which is usually a panel, as described in the preceding section. In this approach, Big Data will be considered as auxiliary information that is taken into account in order to improve the precision of the survey results. The second approach, which we call log-up, involves the enrichment of Big Data. We construct a model based on the survey data, thereby allowing us to

[1] *Google, Apple, Facebook and Amazon, the four American giants that dominate the digital market.*

[2] *Press and Media Statistics Alliance.*

estimate the consumer profile for this media. We will now illustrate each one of these approaches.

Hybrid Internet Audience Measurement on Computers

Coexistence of Two Complementary Measures

In the context of internet audience measurement on computers, two types of complementary measures have coexisted for a number of years now. As detailed in the first part of this article, user-centric measurement is provided by Médiamétrie//NetRatings. It is based on a panel of 16,000 individuals that can estimate the audience and usage of all websites in France. For their part, site-centric measurement tools can provide comprehensive results for website and app consumption in terms of page views, visits and duration. Subscribers to site-centric measurement systems can only access their own results and may not see their position compared to competitors. We call this proprietary measurement. They must then refer to the Médiamétrie//NetRatings panel to find their position.

Launch of a Hybrid Measure in October 2012

Médiamétrie wanted to release a hybrid measurement system onto the market that could take advantage of both measures while still respecting a number of constraints:

- All websites should be able to benefit from the accuracy gain delivered by site-centric measurement, not just those that have subscribed to that measurement;

- The site-centric data used should be consistent with the panel measurement scope;

- The resulting hybrid data should be compatible with media planning tools which require individual data to input into their calculation engine.

In consideration of the three aforementioned constraints, we decided to go for a panel-up approach. Site-centric results are seen as counts known for the total population. The fundamental theoretical principle is this: “whenever we possess auxiliary information, we must seek to use it” (Ardilly, 2006). The idea, therefore, is to use this information by introducing additional auxiliary variables when weighting the sample (Dudoignon *et al*., 2012). Site-centric data for around 400 entities was then sent to Médiamétrie. By data, we mean all of the connection logs collected by the site-centric measurement tools.

Consistency between Site-Centric and Panel Data

Site-centric data is not inherently comparable with panel data for the same entity. In particular, they differ in two aspects: geographical coverage and the terminals measured. Indeed, site-centric measurement counts connections across all devices (computers, mobile phones, tablets, games consoles, etc.) and regardless of the country where the connections occur. In order to introduce site-centric results as weighted auxiliary variables in the panel calibration, the two scopes must be exactly comparable. Consequently, we developed a pre-processing step for site-centric data in order to ensure this consistency. Firstly, the site-centric data is filtered on the device being measured, in this case the computer. Connections from abroad are then dismissed. Other more technical filters are also applied which can notably exclude logs that contain connections performed by robots.

The final step consists of aggregating URLs consistently between the two measures. The objective of this last step is to ensure that these auxiliary variables are consistent between panel and population. The only way to ensure this consistency is to tag all of the URLs of the various entities.

Problems Encountered

The problems encountered were first and foremost related to the representativeness of the entities introduced in the panel calibration. Unfortunately, no site-centric results are available for all web content. Some stakeholders are opposed to subscribing to a site-centric system of measurement. Others have proprietary measurement systems that have not been certified by a trusted third party.

Moreover, it was hard to envisage how we could introduce these 400 entities as weighted auxiliary variables in the panel calibration. Therefore, we decided to make a carefully judged selection of entities. The first rule used was to only include entities for which the number of visitors in the panel was greater than 100, and to minimize the correlation between the entities we introduced. The final selection of entities had to respect the following constraints:

- Consistently cover all population targets in terms of gender, age and socio-professional category;

- Be varied in terms of content (news, travel, cars, etc.);

- Be of limited size in order to allow convergence of the calibration algorithm, without discriminating the calibrated weights distribution, as this would limit the gain in precision.

In the end, a little over 150 entities were chosen to be included in the basis for panel calibration. The introduction of these additional auxiliary variables in the weighting process directly impacts on the quality of the calibrated weights. The ratio between the maximum weight and the minimum weight is higher and we observe that calibrated weights accumulate towards the limits, which lead to a loss of accuracy and to greater instability of the results (Roy *et al*., 2001).

Currently, the CALMAR macro program is used for the calibration (Sautory, 1993). Tests are conducted with new algorithms to summarize the auxiliary information – calibration to the principal components (Goga *et al*., 2011) – or to relax the benchmark constraints on some auxiliary variables – ridge regression calibration (Alleaume *et al*., 2013) –, these algorithms allowing either to improve the quality of the calibrated weights or to introduce a larger number of entities.

Extension of the Method to Global Internet Measurement

Since October 2017, the French market standard for internet audience measurement has been the Global Internet measurement, i.e. on the three screens (computers, mobile phones and tablets). The Global Internet measurement is based on the three panels described above, which have a common part. Indeed, some panelists belong to several panels and are measured on several types of devices. In September 2018, the number of panelists measured on several of their devices is about 6,000 individuals.

The three internet panels are combined by statistical matching to produce audience results on three screens, taking into account the duplication between devices. The site-centric measurement described in the previous section allows the identification of the device used by the user to connect, but without distinction between mobile phone and tablet. A hybrid method by calibration similar to that carried out on the computer internet audience measurement is performed on the sample resulting from a first statistical matching between the panels on mobile phones and tablets. A second statistical matching under constraint of weights conservation is then performed with the computer panel to create the hybrid measurement of the Global Internet.

Hybrid Measurement for Television

As indicated above, panel audience measurement does not always allow the most detailed measurement of very fragmented usages. This is true for Médiamat whose 5,000 households are insufficient to offer a daily service to thematic channels that are exclusively received *via* satellite (CanalSat), ADSL, fibre optic or cable.

In response to the need to assess the value of special interest channels, we chose the log-up approach because it can provide these channels with additional information at little cost, which is always an important consideration and especially so for this category of stakeholders whose marketing research budgets are limited. We are only dealing here with TV data for television channels (i.e. broadcast and not video on demand – VOD). Advertising distribution models are very different between broadcast and VOD or digital platforms, at least for the moment, in France.

To clearly understand the solution developed by Médiamétrie for the hybrid measurement of special interest channels, we must understand firstly, the differences between set-top box usage and individual viewing. To begin with, we notice deviations between set-top box usage and the usage of the television that the set-top box is linked up to. For example: the set-top box can send backlogs that do not correspond to human activity, such as automatic reboots. Furthermore, the set-top box may be switched on and the television switched off: this is very often the case overnight.

In addition, deviations were observed between TV usage and watching TV alone, since the TV remains primarily a family media and a significant part of viewing time is spent watching (the same television) together. Around 40% of the time that individuals aged 4 and over spent in front of the television involves multiple simultaneous viewers, and this figure peaked at 60% for certain weekend time slots (Source: Médiamétrie//Médiamat).

We therefore use a two-step method. The first step is to shift from set-top box to television set. We begin by pre-processing the raw logs, so as to clean up any technical log data and to establish the audience tickets. For each channel viewing, we obtain data of the type: start time, finish time, channel identifier. Next, we proceed to truncate the set-top box usage for those times when the television is probably off. To do this, we shorten the longest audience tickets. The parameters of the truncating function can be estimated from the observed audience tickets durations in the Médiamat panel for the same universe (Figure I).

The second step is to individualize the audience tickets obtained in the 1st step at television set level. This second step presents the most difficulties.

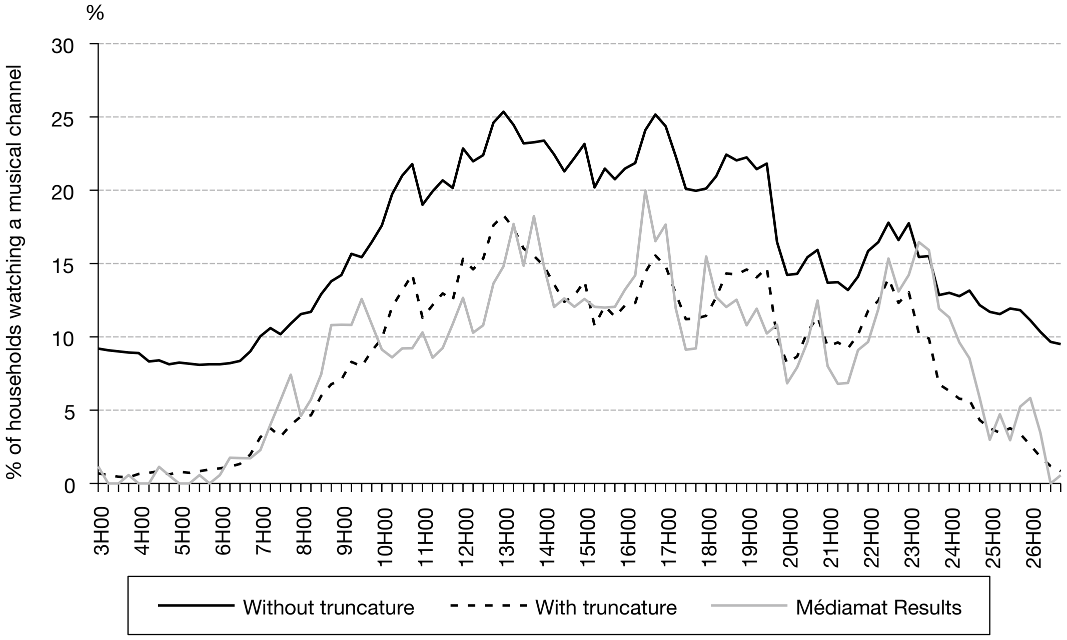
We decided on a modeling approach based on knowledge of the sociodemographic profile of the set-top boxes to be individualized (number of people in household, gender, age, SPG and relationship between individuals). Since the individuals in the household are known, we then only need to determine who is watching the TV when it is turned on. With this approach, we do not therefore use the comprehensiveness of return path data collected by the operators, but only the data from a sample of subscribers who agree to state the nature of their household and who authorize the operator and Médiamétrie to have access to the TV usage data on their set-top box. All of the data is made completely anonymous. Even though the comprehensiveness of the data is not used, the low cost of recruiting a panelist allows us to obtain a large sample size for minimal outlay. This then meets the needs of the thematic channels. The individualization of television set audience tickets without this additional information on household's composition would be hard to envisage.

The individualization of the audience is based on hidden Markov models that can be represented schematically as shown in the Diagram below (Rabiner, 1989; Rabiner *et al.*, 1993).

In our case, the time could be cut into 5-minute steps (but we can choose a longer or shorter time). We then have:

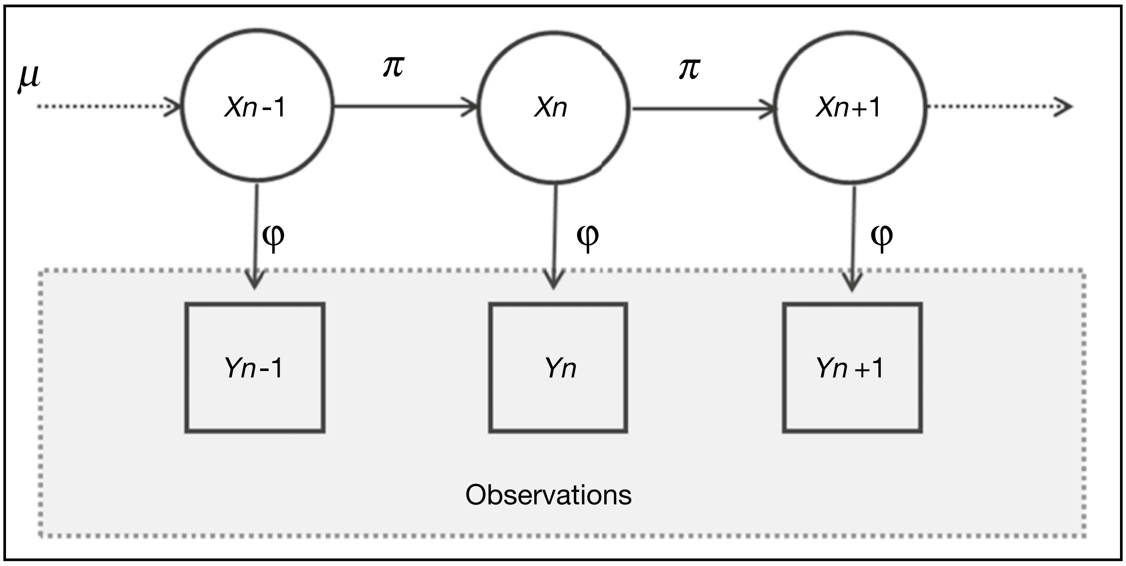
- Observations Y which correspond to the television channels watched, which we group by theme, e.g.: youth, sport, cinema, etc. Y_n is the major theme during the n^{th} time step;

Figure I
Effects of Truncating Function on a Musical Channel



Coverage: Simulation of truncating function on a sample of household subscribed to a French broadcaster.
Sources: Return path data of this broadcaster.

Diagram
Schematic Representation of a Hidden Markov Model



Note: The Markov chain {*Xn*} is not directly observed. Observations {*Yn*} are generated through a memoryless channel, which means that each *Yn* depends only on the state *Xn* at the same moment.

- A hidden phenomenon *X*, which stands for the individuals in front of the television. X_n describes all individuals in the household watching television at time *n*, which enables the correlations between individuals of the same household to be preserved, and therefore the overall levels of watching TV together.

We chose hidden Markov models because their characteristic properties perfectly describe the phenomenon to be modeled, namely:

- A short memory process: to know who is watching the TV at time $n + 1$, we only have to look at who was watching it at time *n*. We do not need to know the full history of who was in front of the TV;

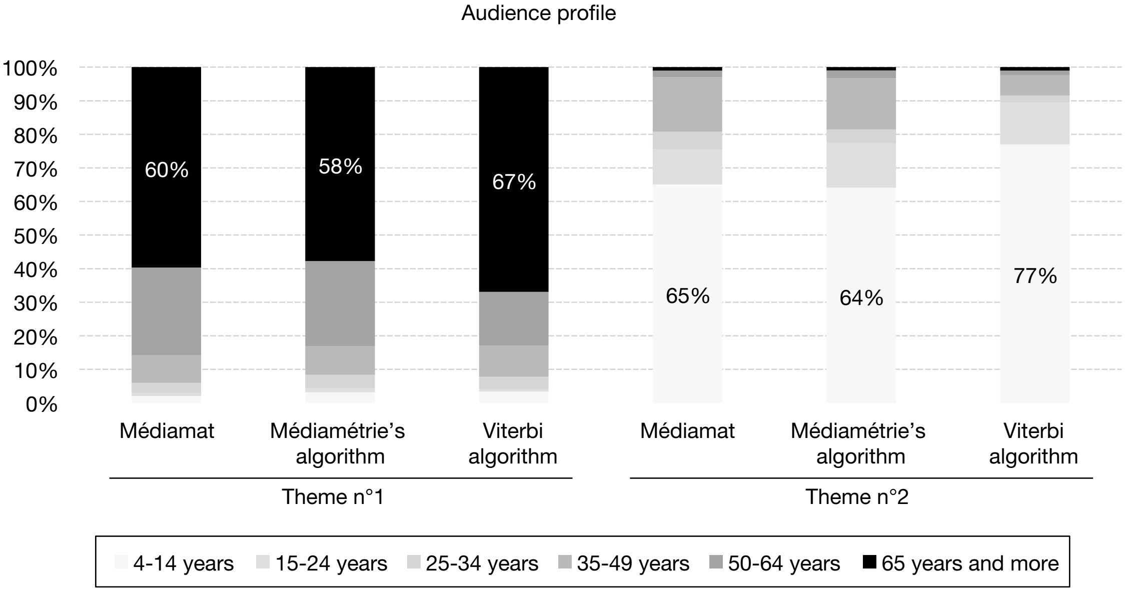
- Observations through a memoryless channel: the TV channel being watched at time *n* only depends on the individuals who are in front of the television at the same time.

The possible states for *X* depend on the size and composition of the household. For a single person household, modeling is pointless (the one individual in that household is watching the TV). For a two-person household, for example a couple, there are three possible states: the reference person alone, the partner alone, or the couple. For a three-person household, for example a couple with one child, there are seven possible states: the reference person alone, the partner alone, the child alone, the reference person with the child, the partner with the child, the couple or the couple and the child.

It can be easily demonstrated that for a household of size *k*, the number of possible states is $2^k - 1$. We have deployed a household typology that describes all household compositions to consider: one person in the household, two persons in the household (couple), two persons in the household (single parent and a child), three persons in the household (couple and child), three persons in the household (single parent and two children), three persons in the household (three adults), etc. For each household type, there is a corresponding sub-model characterized by a set of parameters $M = (\mu,\pi,\varphi)$ where μ is the initial state, π the transition matrix and φ the probabilities of observation. All parameters can be simply estimated using Médiamat panel data, which here serves as a sample for learning.

Once the model parameters are known, we only have to estimate how many people are in front of each television set. Most often, people want to estimate the most likely sequence {*Xn*} by using the Viterbi algorithm (dynamic programming), which allows to do it without calculate all the possibilities. But considering the most likely solution leads to caricatured behaviour estimates (only children in front of youth channels, etc.) and does not reproduce behavioural diversity. We prefer then to use an algorithm with a random component.

Figure II
Comparison of Algorithms – Example of Results on Two Themes with Very Marked Profiles



Reading note: 60% of theme 1 audience is 65 years old and over, in Médiamat panel. With Viterbi algorithm, this is increasing to 67%, that means an over-estimation of older people. While, with Médiamétrie's algorithm, the result is closer to the panel reality with 58%.
Coverage: Audience profile on two themes.
Sources: Tests of individualization on Médiamat panel.

The Médiamat panel is also used as a test sample for the choice of algorithm. Using the panel data, the presences are estimated with the individualization algorithm, then we compare the obtained results with those from Médiamat. The comparisons are not made on a unitary basis (household by household) because the published results are averages and so this could lead to compensations. Instead, the main audience indicators by theme and by channel are compared and we choose the algorithm that minimises the deviations. Figure II gives an illustration of the comparisons that have been made to build the algorithm.

* *
*

The emergence of Big Data – the new Oil – and the development of capacities to store and process this data have raised the prospect of the end of audience measurement in favour of more accurate, more reliable and less expensive measurement systems (Vanheuverzwyn, 2016).

In the first part of this article, we demonstrated that issues surrounding quality were of equal concern for Big Data and survey data. The two examples shown of hybrid approaches clearly show that quality also lies in the processing and modeling that can be applied. Some perfectly good data could lead to incoherent or irrelevant results, especially if we lose sight of the users' needs.

Rather than marking the end, we are observing today an evolution, or even a revolution, in audience measurement towards hybrid measures. There is no question that we must leverage the advantages of different observation systems in order to create others that are more complex and richer. With this outlook, new application fields will open up in research and development. Starting with the theory and practice of surveys. In fact, the utilization of Big Data could be considered as a response to the increasing prevalence of non-response in surveys. The question of the trade-off between bias and variance, estimation bias and calibrated weights variance, has been raised and is worth pursuing. It could lead to the development of more effective calibration algorithms capable of taking many more weighted auxiliary variables into account. It could also result in the development of new hybrid methods based on statistical matching or imputation techniques. Research in machine learning also offers interesting prospects for enriching Big Data and it cannot be ignored in the context of audience measurement.

However, the responses that we put forward to address the needs of observing individual behaviour must be, as they have always been, part of a framework that respects privacy and the legal restrictions associated with the processing of personal data. This is not so much a legal question as an ethical one (Tassi, 2014). The entry into force of the European General Regulations on Data Protection and the public debates that took place upstream, made it possible to highlight the drifts in the measurement of internet usages. Surveys, for which the consent of the individual is inherent, therefore regain a central role. □

BIBLIOGRAPHY

Alleaume, F. & Dudoignon, L. (2013). Calage sur information auxiliaire incertaine : proposition d'algorithme de redressement ridge. *Actes des 45e Journées de Statistique de la SFdS,* Toulouse, 2013.
http://papersjds13.sfds.asso.fr/submission_189.pdf

Ardilly, P. (2006). *Les Techniques de sondage*. Paris : Éditions Technip.
http://www.editionstechnip.com/en/catalogue-detail/113/techniques-de-sondage-les.html

Brackstone, G. (1999). La gestion de la qualité des données dans un bureau de statistique. *Techniques d'enquête,* 25(2), 159–171.
https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1999002/article/4877-fra.pdf?st=FSaA6d3F

Brackstone, G. (2006). Le rôle des méthodologistes dans la gestion de la qualité des données. In : Lavallée, P. & Rivest, L.-P., *Méthodes d'enquêtes et sondages.* Paris : Dunod.
https://www.dunod.com/sciences-techniques/methodes-d-enquetes-et-sondages-pratiques-europeenne-et-nord-americaine

Deville, J.-C. & Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, 87(418), 376–382.
https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1992.10475217#.XGbmljNKiiM

Dudoignon, L. & Logeart, J. (2014). Mesure hybride de l'audience TV. *Actes des 46e Journées de Statistique de la SFdS*, Rennes, 2014.
http://papersjds14.sfds.asso.fr/submission_128.pdf

Dudoignon, L. & Zydorczak, L. (2012). Enquête et données exhaustives : un nouveau défi pour les mesures d'audience. *Actes en ligne du 7e Colloque Francophone sur les Sondages*, Rennes, 2012.
http://sondages2012.ensai.fr/wp-content/uploads/2011/01/Dudoignon-Zydorczak-Mesures-Hybrides-Médiamétrie-2012-Article.pdf

Dussaix, A-M. (2008). La qualité dans les enquêtes. *MODULAD,* 39, 137–171.
https://www.rocq.inria.fr/axis/modulad/archives/numero-39/Tutoriel-Dussaix/Dussaix-39.pdf

EUROSTAT (2007). *Handbook on Data Quality Assessment Methods and Tools*.
https://unstats.un.org/unsd/dnss/docs-nqaf/Eurostat-HANDBOOK ON DATA QUALITY ASSESSMENT METHODS AND TOOLS I.pdf

Fischer, N. (2004). Fusion statistique de fichiers de données. *Thèse de doctorat.* Paris : Conservatoire National des Arts et Métiers.
https://cedric.cnam.fr/fichiers/RC899.pdf

Goga, C., Shehzad, M.-A. & Vanheuverzwyn, A. (2011). Régression en composantes principales versus ridge régression en sondages. Application aux données Médiamétrie. *Actes des 43e Journées de Statistique de la SFdS*, Tunis, 2011.
https://www.researchgate.net/publication/292133976_Regression_en_composantes_principales_versus_ridge_regression_en_sondages_Application_aux_donnees_Mediametrie

Institut de la Statistique du Québec (2006). *Le cadre intégré de la gestion de la qualité de l'Institut de la statistique du Québec*.
http://www.stat.gouv.qc.ca/institut/CadreGestion_qual.pdf

Kiaer, A. N. (1896). Observations et expériences concernant des dénombrements représentatifs. *Bulletin de l'Institut International de Statistique*, 9(2).
https://gallica.bnf.fr/ark:/12148/bpt6k61560p?rk=42918;4

Lyberg, L. (2012). La qualité des enquêtes. *Techniques d'enquête,* 38(2), 115–142.
https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2012002/article/11751-fra.pdf?st=NfC31Ekj

Médiamétrie & Médiamétrie//NetRatings (2010). Les mesures hybrides – Synergies et rapprochement entre les mesures de l'Internet. *Le Livre Blanc*. https://www.mediametrie.fr/fr/les-publications-scientifiques-de-2008-2016

Neyman, J. (1934). On the Two Different Aspects of Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection. *Journal of the Royal Statistical Society*, 97(4), 558–625. https://www.jstor.org/stable/2342192

OCDE (2011). *Quality dimensions, core values for OECD statistics and procedures for planning and evaluating statistical activities*. http://www.oecd.org/sdd/21687665.pdf

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 257–286. https://ieeexplore.ieee.org/document/18626

Rabiner, L. R. & Juand, B.-H. (1993). *Fundamentals of Speech Recognition*. Upper Saddle River, NJ, USA: Prentice Hall. https://dl.acm.org/citation.cfm?id=153687

Roy, G. & Vanheuverzwyn, A. (2001). Redressement par la macro CALMAR : applications et pistes d'amélioration. In: Lejeune, M. (Ed.), *Traitement des fichiers d'enquêtes*. Grenoble : Presses Universitaires de Grenoble. https://www.pug.fr/produit/314/9782706110295/traitements-des-fichiers-d-enquetes

Sautory, O. (1993). La macro CALMAR : redressement d'un échantillon par calage sur marges. Insee, *Méthodes*. https://www.insee.fr/fr/information/2021902

Tassi, P. (2014). La data est-elle éthique-compatible et quelques questions posées par les données. *8e Colloque Francophone sur les Sondages*, Dijon, 2014. https://www.mediametrie.fr/fr/les-publications-scientifiques-de-2008-2016

Vanheuverzwyn, A. (2016). Mesure d'audience et données massives : mythes et réalités. *9e Colloque Francophone sur les Sondages*, Gatineau, 2016. http://paperssondages16.sfds.asso.fr/submission_104.pdf