Big Data et mesure d'audience : un mariage de raison ?

Big Data and Audience Measurement: A Marriage of Convenience?

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

Résumé – La convergence numérique a peu à peu modifié l'univers des données mais aussi celui des médias. Les frontières entre médias sont devenues floues et ce phénomène s'amplifie chaque jour avec la diffusion de nouveaux équipements et de nouveaux usages. En parallèle, la convergence numérique a mis en évidence le pouvoir des Big Data – les mégadonnées, ou données massives – dont la définition comporte deux paramètres joints : la quantité et la fréquence d'acquisition. La quantité peut aller jusqu'à l'exhaustivité, la fréquence peut aller jusqu'au temps réel. Si les données massives pourraient être vues comme un risque de retour potentiel vers le paradigme de l'exhaustif dominant jusqu'à la fin du 19e siècle, alors que le 20e siècle a été celui de l'échantillonnage et des sondages, Médiamétrie a choisi de considérer cette révolution digitale comme une formidable opportunité pour faire évoluer ses dispositifs de mesure d'audience.

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 20rh 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.*

Codes JEL / JEL Classification : C18, C32, C33, C55, C80
Mots-clés : hybridation, données massives, enquêtes par sondage, modèle de Markov caché
Keywords: hybrid methods, Big Data, sample surveys, hidden Markov model

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

Reçu le 10 juillet 2017, accepté 10 février 2019

Pour citer cet 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, 113–146. https://doi.org/10.24187/ecostat.2018.505d.1969

Le 20e siècle a été marqué par un recul progressif de l'exhaustif au profit du développement des enquêtes par sondage. On peut considérer que l'acte fondateur en est la communication d'Anders N. Kiaer lors du Congrès de l'Institut International de Statistique en 1895 intitulée *Observations et expériences concernant des dénombrements représentatifs*. En 1934 paraît l'article de référence de la théorie des sondages de Jerzy Neyman *« On the two different aspects of representative methods, the method of stratified sampling and the method of purposive selection »*. La croissance de l'équipement téléphonique favorise ensuite l'utilisation des enquêtes par sondage dans de nombreux domaines (statistique publique, politique, santé, marketing, mesure d'audience, etc.). La fin du 20e siècle connaît un nouveau changement de paradigme avec l'apparition des données massives : le retour vers l'exhaustif. Acteur privilégié de cette révolution numérique, le secteur des médias a vu ses systèmes de mesure se multiplier et parfois, inévitablement, se contredire. Médiamétrie, institut de référence dans la mesure d'audience des médias en France, a dû par conséquent faire évoluer ses méthodes pour tirer parti du meilleur de chaque source.

La première partie de l'article porte sur les avantages et les limites comparées des données d'enquête et des données massives, en insistant sur la notion de qualité dans ses diverses dimensions. Cela permettra d'expliquer pourquoi Médiamétrie a choisi de considérer données d'enquêtes et données massives comme complémentaires et non comme concurrentes. Nous considérons, en effet, que les approches hybrides, qui consistent à « mélanger deux sources d'informations de natures et de niveaux différents pour en créer une troisième plus riche ou plus fine », sont devenues des démarches naturelles (Médiamétrie, 2010). La seconde partie illustre ces approches au travers de deux mises en œuvre opérationnelles dans le domaine de la mesure d'audience des médias. Nous commencerons par présenter la méthode hybride mise en place dans le cadre de la mesure d'audience Internet, référence du marché français depuis 2012, comme exemple des approches dites « *panel-up* » (Dudoignon *et al*., 2012). Nous finirons par un exemple d'approche dite « *log-up* » mis en place pour la mesure d'audience des chaînes thématiques (Dudoignon *et al*., 2014). Nous verrons, pour ces deux cas, que pour donner du sens et une valeur aux données massives, il faut au préalable bien comprendre leur mode d'acquisition, y compris souvent les aspects techniques, pour les « nettoyer », les transformer de sorte que le mariage avec les données d'enquêtes soit possible et surtout heureux.

Préambule : données disponibles dans les mesures d'audience

Les médias pour lesquels nous disposons à la fois de données d'enquêtes et de données massives sont la télévision et surtout Internet. Pour ces deux médias, la mesure d'audience est basée sur un panel et un dispositif de mesure semi-automatique. Nous proposons, dans ce préambule, de décrire brièvement les dispositifs existants pour les mesures d'audience de la télévision et d'Internet opérées par Médiamétrie en France.

Internet

La mesure d'audience d'Internet repose sur deux types de dispositifs. Les dispositifs dits « *user-centric* », centrés sur l'utilisateur, s'attachent à suivre le comportement d'audience des sites et applications Internet des individus sur l'ensemble de leurs appareils. Ils sont basés sur des panels d'individus et leurs connexions sont mesurées à l'aide de logiciels appelés « *meters* » installés sur leurs ordinateurs, téléphones mobiles ou tablettes et qui renvoient l'information sur les serveurs de Médiamétrie. Le second type de dispositif est qualifié de « *site-centric* », centré sur le site. Ce type de mesure repose sur l'insertion de marqueurs (encadré 1) sur les sites et applications des clients souscripteurs et permet un comptage exhaustif du nombre de visites, de pages vues et de la durée de connexion.

La mesure d'audience Internet sur ordinateur

L'ordinateur étant un équipement partagé au sein du foyer, le panel est constitué par grappage de l'ensemble des individus âgés de 2 ans et plus du foyer. Les unités primaires du panel sont donc les foyers et les unités secondaires les individus de 2 ans et plus. Le recrutement des unités primaires est réalisé selon la méthode empirique des quotas. Une fois le *meter* installé sur l'ensemble des ordinateurs du foyer, une fenêtre (ou *pop-up*) apparaît à chaque connexion et les unités secondaires, les individus, doivent s'identifier en cochant la case qui leur correspond. En septembre 2018, le panel est composé d'environ 6 200 foyers disposant d'un accès Internet *via* un ordinateur, soit plus de 14 000 individus.

Le champ de la mesure ne peut se réduire aux seules connexions à domicile. En effet, sur la population des actifs occupés, une part importante des connexions à Internet depuis un ordinateur est réalisée depuis le lieu de travail. Néanmoins, la charge que représente pour les individus la participation au dispositif de mesure, qualifiée dans la littérature de « fardeau de réponse », nous empêche d'imposer à l'ensemble des unités secondaires du panel d'être également mesurées sur leur lieu de travail si elles y disposeraient d'un ordinateur avec accès à Internet, sous peine d'un taux de réponse très faible. Le dispositif est donc complété par un panel indépendant d'individus ayant un accès Internet *via* un ordinateur sur leur lieu de travail. Ce panel est composé en septembre 2018 de près de 2 000 individus et il est rapproché du précédent par fusion statistique (Fisher, 2004).

La mesure d'audience Internet sur tablette

Le principe de la mesure d'audience Internet sur tablette est très similaire à celui de la mesure sur ordinateur. L'usage des tablettes au sein des entreprises étant encore peu développé, le champ de la mesure est aujourd'hui limité au domicile. Le panel d'individus est constitué par grappage au sein des foyers recrutés. Ces derniers doivent installer une application de mesure sur l'ensemble des tablettes du foyer et en modifier le paramétrage de manière à assurer un routage de leurs connexions sur les serveurs de Médiamétrie. Dès lors que l'application est ouverte, elle permet l'identification de l'utilisateur. En septembre 2018, le panel est composé de près de 2 000 foyers, soit 5 200 individus de 2 ans et plus.

La mesure d'audience Internet sur téléphone mobile

Contrairement à l'ordinateur et à la tablette, le téléphone mobile est un équipement à usage majoritairement individuel. Le panel est par conséquent composé d'individus recrutés par la méthode des quotas. L'âge minimum de participation à la mesure est fixé à 11 ans et, conformément aux contraintes imposées par la loi Informatique et Libertés du 6 janvier 1978, la participation des mineurs est acceptée après consentement d'un adulte titulaire de l'exercice de l'autorité parentale. À l'instar du système de mesure des connexions sur tablette, le panéliste doit installer une application sur son téléphone mobile. Cette application permet le routage des connexions sur les serveurs de Médiamétrie. L'ensemble du trafic Internet du téléphone est attribué au panéliste, utilisateur principal du téléphone. L'usage du téléphone mobile par un utilisateur secondaire est donc, par convention, affecté à l'utilisateur principal. En septembre 2018, le panel est composé de près de 11 000 individus de 11 ans et plus.

La mesure des connexions sécurisées

La participation aux dispositifs de mesure *user-centric* se matérialise par la signature d'une convention entre Médiamétrie et ses panélistes. Cette convention liste les engagements respectifs de Médiamétrie et des panélistes. Médiamétrie s'engage notamment à collecter les données d'usage des panélistes à des fins purement statistiques. Elle s'engage par ailleurs à ne jamais divulguer l'identité de ses panélistes à un tiers à des fins publicitaires

Encadré 1 – **Description des technologies de mesures**

Qu'appelle-t-on un marqueur ?

Dans le domaine de l'analyse du Web, un marqueur, ou *tag* en anglais, est un élément introduit dans chacun des contenus à mesurer, afin d'en comptabiliser leur diffusion. Le contenu peut être une page, une application, un podcast ou même un contenu audio ou vidéo. Il s'agit d'une ligne de programme insérée dans le code source du contenu. Il permet de générer un journal de connexions sur le serveur de l'outil de mesure tiers à chaque fois que le contenu est consulté. Il permet donc un comptage exhaustif des connexions sur les contenus marqués.

Qu'est-ce que le watermarking *audio ?*

Technologie utilisée pour la mesure d'audience de la télévision, le *watermarking* audio consiste en l'insertion d'une marque (un tatouage) inaudible à l'oreille humaine dans le signal audio du flux à mesurer. C'est un encodeur, matériel professionnel retenu par Médiamétrie, qui permet d'insérer ce tatouage numérique. Le principe consiste à modifier le signal qui émet le programme en ajoutant de l'information, sans impacter l'audibilité de la séquence. En bout de chaîne, la marque est captée par les audimètres reliés aux téléviseurs des panélistes. La marque insérée par l'encodeur contient l'identification de la chaîne qui diffuse le programme et des repères réguliers sur l'heure de la diffusion. On peut ainsi faire la distinction entre l'audience d'un programme en *live*, c'est-à-dire au moment de sa diffusion, et l'audience d'un programme enregistré au préalable ou sur une plateforme de contenus en *replay*.

ou commerciales. Enfin, elle s'engage à prendre toutes précautions utiles pour préserver la sécurité des données collectées et, notamment, empêcher qu'elles soient déformées, endommagées, ou que des tiers non autorisés y aient accès. Réciproquement, les panélistes s'engagent à préserver la confidentialité concernant leur participation à l'étude ainsi que les modalités de leur participation et ce afin d'éviter toute tentative de corruption de la part des acteurs, éditeurs ou opérateurs, ayant un intérêt dans la mesure d'audience. Ils s'engagent par ailleurs à installer le logiciel de mesure, à s'identifier le cas échéant, à informer Médiamétrie en cas de changement de situation et à accepter d'être contacté par Médiamétrie.

Une fois la convention signée, les panélistes autorisent Médiamétrie à avoir accès à l'ensemble de leurs données d'usage Internet, y compris leurs connexions en HTTPS et leur adresse IP. Néanmoins, pour des raisons techniques, l'information collectée dans le cadre des connexions sécurisées est dans certains cas moins fine que celle recueillie lors de connexions en HTTP. Par exemple, pour la mesure des connexions sur tablette, seul le nom de domaine est disponible dans les logs renvoyés sur les serveurs de Médiamétrie dans le cas d'une connexion HTTPS alors que l'url complète sera collectée pour les connexions en HTTP.

Télévision

Le panel Médiamat de Médiamétrie constitue la mesure de référence de l'audience de la télévision en France métropolitaine. Cette mesure est basée sur un panel d'individus constitué par grappage de près de 5 000 foyers équipés d'au moins un poste de télévision. L'ensemble des postes de télévision actifs font partie du champ de la mesure, c'est-à-dire ceux utilisés au moins une fois par mois pour regarder la télévision. À chacun de ces postes est relié un audimètre qui détecte à tout moment, à l'aide de la technologie du *watermarking* audio (cf. encadré 1), quelle est la chaîne regardée sur le téléviseur. Les individus du foyer doivent participer à la mesure en déclarant leur présence devant le poste à l'aide d'une télécommande reliée à l'audimètre. Les données enregistrées par les audimètres sont collectées en continu par les serveurs de Médiamétrie. Même si les panélistes ont pour consigne de déclarer la présence devant l'écran de l'ensemble des individus du foyer, les résultats d'audience ne sont restitués que sur l'univers des individus âgés de 4 ans et plus.

La voie de retour en TV (encadré 2) est techniquement possible dans deux cas : les décodeurs numériques de l'ADSL, du câble et du satellite lorsqu'ils sont connectés à Internet et les téléviseurs connectés. Il est à noter que même si la plupart des téléviseurs commercialisés aujourd'hui sont connectables, leur connexion effective est encore assez rare. Dans ces deux seuls cas, les logs de connexion sont disponibles auprès de l'opérateur distribuant le flux et permettent de savoir sur quels chaîne ou service est allumé le décodeur. Tout usage du téléviseur fait en dehors du décodeur n'est pas mesuré par l'opérateur : par exemple, si le téléviseur est branché à une antenne TNT et un décodeur ADSL, les programmes regardés *via* l'antenne TNT ne sont pas mesurés par l'opérateur ADSL.

Qualité des données d'enquêtes et des données massives

S'il n'existe pas une définition unique de ce qu'est la qualité des données d'enquêtes (Dussaix, 2008), c'est encore plus vrai en ce qui concerne la qualité des données en général. On peut cependant retenir que la qualité est une réelle préoccupation pour la plupart des organismes produisant des statistiques et que la plupart s'accordent à dire qu'il s'agit d'une notion multidimensionnelle difficile à évaluer (Lyberg, 2012). Nous avons choisi pour notre discussion de retenir les six dimensions de la qualité retenues notamment par Statistique Canada et l'Australian Bureau of Statistics que sont la pertinence, l'exactitude, l'actualité, l'accessibilité, l'intelligibilité et la cohérence (Brackstone, 1999 ; Institut de Statistique du Québec, 2006). On notera que l'OCDE ajoute deux dimensions supplémentaires – la crédibilité et la rentabilité – pour évaluer la qualité des productions statistiques (OCDE, 2011). Il ne s'agit pas ici de discuter de la définition des dimensions de la qualité des enquêtes mais de proposer une analyse comparative « données d'enquêtes » vs « données massives » sur chacune de ces dimensions.

La pertinence

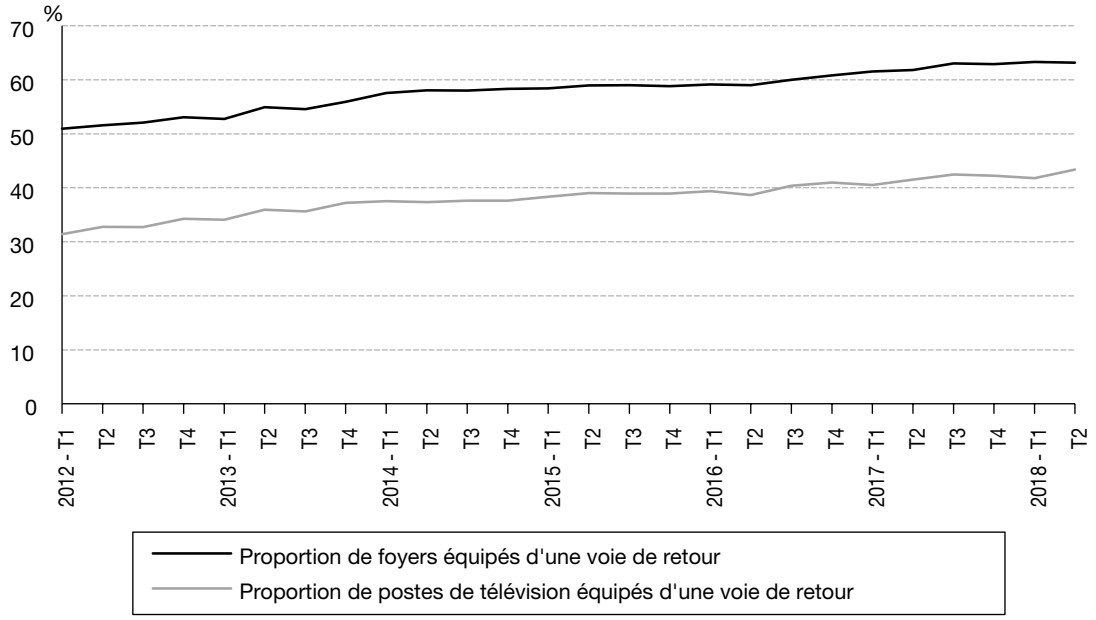
La pertinence renvoie à l'utilité, la capacité à répondre aux besoins des utilisateurs ou clients. C'est donc naturellement que ce critère est, la plupart du temps, le premier retenu pour évaluer la qualité. La pertinence des mesures d'audience

ENCADRÉ 2 – **Quelles sont les données disponibles par voie de retour en TV ?**

On appelle voie de retour en TV la possibilité offerte par certains modes de diffusion de collecter des informations numériques sur la consommation TV des utilisateurs. La voie de retour est techniquement possible pour tous les décodeurs connectés à Internet ainsi que pour les téléviseurs connectés. Concrètement, ce type de collecte est mis en place par des opérateurs Télécom ou des opérateurs satellite comme CanalSat.

On estime que la voie de retour est aujourd'hui possible pour un peu plus de 60 % des foyers français équipés d'au moins un poste de télévision mais pour à peine plus de 40 % des postes de télévision. En effet, le décodeur est bien souvent relié uniquement au téléviseur principal et pas aux postes secondaires. Il s'agit bien d'un potentiel, car tous les décodeurs connectables à Internet ne sont pas nécessairement connectés.

Figure A
Évolution du potentiel de voie de retour en télévision



Champ : France métropolitaine.
Source : Médiamétrie – Home Devices.

par panel n'est globalement pas remise en cause dans la mesure où ces dispositifs sont conçus en étroite collaboration avec leurs utilisateurs. En effet, pour chaque média, un comité composé de membres représentant les diffuseurs et les utilisateurs, les annonceurs et publicitaires, les éditeurs et opérateurs, et Médiamétrie, a été créé sur une base paritaire. Le rôle de chaque comité est de définir, orienter et valider les mesures et études qui, pour chacun des médias concernés, constituent la référence.

Cependant, les mesures d'audience par panel ne peuvent répondre parfaitement à tous les besoins, en particulier, lorsqu'il s'agit de mesurer des usages très confidentiels ou très morcelés qui seront nécessairement faiblement représentés – voire pas du tout – au sein d'un échantillon. L'augmentation de la taille des échantillons n'est évidemment pas une réponse pertinente, car la pertinence d'une étude intègre les contraintes budgétaires de ses utilisateurs. À l'inverse, les données massives ne satisfont pas entièrement les besoins des utilisateurs car elles ne permettent pas d'identifier les usages individuels mais des usages de machines. Un pré-traitement qui vise à nettoyer et transformer ce type de données pour leur donner du sens est indispensable. Nous verrons dans la suite quelques exemples concrets de ce type de pré-traitement. Elles fournissent en revanche des informations précieuses sur les usages émergents ou de niche non mesurables par échantillon en raison de leur volumétrie. Sur ce premier critère qu'est la pertinence, la complémentarité entre les données d'enquêtes et les données massives à des fins de mesure d'audience apparaît donc de manière évidente.

L'exactitude

Dans notre contexte, l'exactitude correspond au fait de décrire correctement le comportement média des français. S'il est communément admis que les résultats issus d'enquêtes sont entachés d'erreurs liées à l'échantillonnage ou au phénomène de non-réponse propre au sondage, on a tendance à penser qu'*a contrario* les données massives, qui peuvent être exhaustives sur leur périmètre de mesure, sont exactes. Or, il n'en est rien. En effet, comme nous l'avons précédemment évoqué, les données massives apportent des informations concernant des machines et non des individus, ce qui constitue une évidente source d'erreur. De plus, les technologies utilisées pour effectuer ces mesures, si elles ne sont pas correctement maîtrisées, peuvent aussi conduire à des erreurs d'implémentation ou d'interprétation. On en revient à la phase de pré-traitement qui doit permettre de nettoyer en partie ces erreurs d'interprétation. En ce qui concerne les erreurs d'implémentation (mauvaise implémentation d'un *tag* Web par exemple), la meilleure façon de procéder est de mettre en place un système de supervision qui détectera au plus tôt ces défauts et de les corriger avant qu'un volume de données trop important ne soit impacté. À noter que ce type de supervision est aussi indispensable pour les mesures par panel qui utilisent dans certains cas des technologies de marquage des contenus dont on souhaite mesurer l'audience, *via* un *tag* pour le Web ou *via* le *watermarking* audio pour la télévision.

L'actualité (ou rapidité de diffusion)

L'actualité correspond au délai de diffusion des résultats depuis la période de référence de l'analyse. Dans le contexte de mesure d'audience des médias ce critère est très important. Un résultat trop tardif sera vite obsolète et d'un intérêt très limité pour les utilisateurs. Pour Internet, les résultats sont généralement mensuels et doivent être publiés le mois suivant la période analysée. Pour la TV, les délais sont beaucoup plus courts. Les premiers résultats d'audience des programmes d'une journée sont publiés dès le lendemain matin à 9 h. Ces résultats sont ensuite consolidés une semaine plus tard par la prise en compte de la consommation de ces programmes en différé dans les sept jours après leur diffusion à l'antenne.

Que ce soit pour les données d'enquêtes, pour les données *site-centric* ou issues des voies de retour, lorsque l'on utilise des technologies de mesure automatiques, l'acquisition des données brutes peut théoriquement se faire quasiment en temps réel. La fraîcheur des résultats peut donc être assurée dès lors que les opérations de pré-traitement et de traitement de ces données sont réalisées dans des temps limités. Dans les deux cas, cela implique la mise en place de processus de production très rigoureux, automatisés et industrialisés.

L'accessibilité

L'accessibilité aux résultats des mesures d'audience est assurée grâce à des interfaces de restitution ouvertes à l'ensemble des souscripteurs. Ce type d'interface permet notamment de gérer des droits pour les différents utilisateurs et donc leur donner accès à plus ou moins d'information selon leur souscription. Du point de vue de l'utilisateur, l'accessibilité sera considérée comme satisfaisante si l'outil de consultation des résultats est à la fois ergonomique et performant en temps de calcul ou d'affichage. En interne, l'ensemble des données est aisément accessible aux équipes chargées de la production de résultats ou de la réalisation d'analyses complémentaires. Ces accès sont néanmoins limités, y compris en interne, à de la donnée anonymisée. Seules les équipes de gestion et d'animation des panels ont accès aux données personnelles permettant de contacter les panélistes.

Les difficultés techniques concernant l'accès aux données massives sont aujourd'hui de moins en moins fréquentes et ne constituent plus un enjeu prioritaire de développement. En revanche, les contraintes juridiques obligent à limiter l'accès à ce type de données voire à réduire la quantité d'information collectée. Si par le passé, des données numériques pouvaient parfois être collectées à l'insu des individus, ce type de pratique n'est désormais plus possible, en tout cas en Europe. La plupart des acteurs collectant actuellement ce type de données (*site-centric* ou voie de retour) ont ainsi dû investir pour se mettre en conformité avec le Règlement Général européen sur la Protection des Données (encadré 3).

L'intelligibilité, ou possibilité d'interprétation

Qu'il s'agisse de données d'enquêtes ou de données massives, l'intelligibilité de la donnée est

principalement liée aux technologies. On peut considérer que les technologies de marquage TV et Internet génèrent des données brutes, qu'on appelle des *logs*, très peu intelligibles. Ce sont les pré-traitements qui vont permettre de traduire ces données dans un format interprétable. Le statisticien ne peut évidemment pas travailler seul. Ce type de données nécessite une étroite collaboration entre les équipes techniques qui développent les solutions de marquage, les équipes informatiques qui collectent et traitent les données, les équipes statistiques qui conçoivent les analyses et les équipes en relation avec les clients qui doivent mettre en place la solution de marquage sur leurs sites ou chaînes.

Les solutions de marquage des contenus médias, même si elles peuvent paraître compliquées, permettent d'avoir de la donnée intelligible après traduction, que l'on peut enrichir aisément de métadonnées décrivant finement le contenu (par exemple, préciser pour un contenu vidéo sur Internet, s'il s'agit d'une série, la saison, l'épisode, la date de diffusion à l'antenne en télévision, etc.). Les solutions de mesure automatique qui n'utilisent pas de marquage sont en général beaucoup moins intelligibles. On pense par exemple aux mesures de l'audience Internet basées sur une capture du trafic réseau d'un appareil pour lesquelles plus de 90 % de l'information collectée n'est pas pertinente car elle ne permet pas de décrire les comportements de l'individu utilisant l'appareil. Elle comprend en effet l'intégralité des flux techniques comme, par exemple, les mises à jour des logiciels ou applications, qui sont complètement transparentes pour l'utilisateur. Rendre ce type de données intelligibles constitue un vrai défi, toute erreur de filtrage de données conduisant le plus souvent à une erreur d'interprétation. Les solutions de marquage permettent quant à elles de ne collecter que de l'information utile et sont dans ce sens nettement plus faciles à interpréter.

La cohérence

Sans les approches hybrides, un même acteur peut avoir à sa disposition plusieurs indicateurs de performance d'un contenu. Par exemple, un nombre moyen de personnes ayant regardé un contenu vidéo et un nombre de décodeurs allumés au moins une minute sur cette même vidéo. Ces deux indicateurs, basés sur des unités différentes, ne sont pas comparables, mais peuvent perturber les utilisateurs non avertis dès lors qu'ils sont tous deux publiés. Le travail de Médiamétrie consiste donc à apporter la cohérence nécessaire. En premier lieu en expliquant clairement les concepts, les indicateurs et comment les interpréter. Ensuite, en proposant des solutions pour rapprocher ces données de

Encadré 3 – **Règlement Général européen sur la Protection des Données : ce qui change pour les professionnels**

Le nouveau règlement européen, entré en vigueur le 25 mai 2018, introduit ou renforce les principes suivants.

• Renforcement des droits des personnes : les utilisateurs doivent être informés du recueil et de l'utilisation de leurs données. Ils doivent pouvoir à tout moment donner leur consentement ou s'opposer, le cas échéant. Les utilisateurs disposent de nouveaux droits : en particulier, le droit à la limitation du traitement, le droit à la portabilité des données et le droit à l'effacement des données.

• Responsabilité des acteurs (responsables de traitement et sous-traitants) : le règlement allège les obligations de formalités préalables auprès de la CNIL. En contrepartie, il introduit le principe de démontrabilité : pouvoir prouver à tout moment la conformité au règlement en documentant de manière détaillée toutes activités de traitement de données à caractère personnel. Concrètement, le responsable de traitement s'engage à : tenir à jour des registres détaillés des activités de traitement de données à caractère personnel, effectuer systématiquement des analyses d'impact avant chaque traitement présentant un risque élevé pour les droits et libertés des personnes physiques, et veiller à la conformité des éventuels sous-traitants. Le règlement renforce aussi les sanctions pour le responsable du traitement en cas de manquement : jusqu'à 20 millions d'euros ou 4 % du chiffre d'affaires mondial.

• *Privacy by design* : l'entreprise doit prendre en compte la notion de respect de la vie privée dès la conception d'un produit, d'une application. Le responsable de traitement devra mettre en œuvre toutes les mesures techniques et organisationnelles nécessaires au respect de la protection des données personnelles dès la conception et par défaut.

• Création de la fonction de Délégué à la protection des données ou *Data protection officer* (DPO). Ce nouvel expert identifie et coordonne au sein de son entreprise ou organisme les actions à mener en matière de protection des données à caractère personnel : de la communication interne aux contrôles du respect du règlement, tout en étant le point de contact avec l'autorité de contrôle.

natures différentes pour proposer une mesure cohérente. La cohérence des données d'enquêtes et des données massives est donc tout l'enjeu des mesures hybrides mises en place par Médiamétrie.

Un critère de qualité qui n'est pas évoqué dans les six cités ci-dessus mais qui, concernant les données massives, doit être examiné est la confiance (ou crédibilité pour l'OCDE). Certains acteurs médias ont mis en place des systèmes de mesure *site-centric* ou par exploitation de la voie de retour. C'est le cas en particulier des plus gros acteurs du Web, les GAFA[1], ou des opérateurs de télécom. Ces acteurs proposent ainsi des services de mesure à destination notamment des éditeurs qui les utilisent comme plateforme de diffusion. Comme il est en général très difficile d'être à la fois juge et partie, la question de la confiance sera toujours posée par les autres acteurs du marché. Dans ce contexte, les données massives « propriétaires » nécessitent souvent une certification par un tiers de confiance pour être reconnues et partagées par le marché. C'est ce que fait par exemple l'ACPM[2] pour le marché de la Presse, avec sa certification de la diffusion et de la distribution de la Presse et de la fréquentation des supports numériques.

Exemples d'approches hybrides pour la mesure d'audience des médias

Deux approches sont théoriquement possibles pour les mesures hybrides. Le choix de l'une ou l'autre des approches dépend du besoin exprimé par les utilisateurs. Dans une première approche, qu'on appelle *panel-up*, la donnée massive vient enrichir l'information issue de l'enquête, le plus souvent un panel comme exposé dans la partie précédente. Dans cette approche, la donnée massive est considérée comme une information auxiliaire que l'on prend en compte afin d'améliorer la précision des résultats de l'enquête. La seconde approche, qu'on appelle *log-up*, consiste en un enrichissement de la donnée massive. On construit un modèle à partir des données de l'enquête qui nous permet d'estimer le profil des consommateurs du média par exemple. Nous proposons d'illustrer chacune de ces approches.

La mesure d'audience hybride Internet sur ordinateur

Coexistence de deux mesures complémentaires

Dans le contexte de la mesure d'audience Internet sur ordinateur, deux types de mesure complémentaires coexistent depuis de nombreuses années. Comme détaillé dans la première partie, la mesure dite *user-centric* est assurée par Médiamétrie//NetRatings. Elle est basée sur un panel de 16 000 individus qui permet d'estimer l'audience et l'usage de l'ensemble des sites Internet en France. Les outils de mesure *site-centric* offrent quant à eux la possibilité de disposer de résultats exhaustifs de consommation des sites et applications Internet en termes de pages vues, de visites et de durée. Les souscripteurs aux dispositifs de mesure *site-centric* n'ont accès qu'à leurs propres résultats et ne peuvent se situer dans leur univers de concurrence, c'est ce qu'on appelle une mesure propriétaire. Ils doivent ainsi se référer au panel Médiamétrie//NetRatings dans cet objectif.

Lancement d'une mesure hybride en octobre 2012

Médiamétrie a souhaité mettre à disposition du marché une mesure hybride qui puisse tirer profit de ces deux mesures tout en respectant un certain nombre de contraintes :

- tous les sites doivent pouvoir bénéficier du gain de précision apporté par la mesure *site-centric* et pas uniquement les sites souscripteurs de cette mesure ;

- la donnée *site-centric* utilisée doit être cohérente avec le champ de la mesure par panel ;

- la donnée hybride résultante doit être compatible avec les outils de médiaplanning qui ont besoin de données individuelles en entrée de leur moteur de calcul.

Compte tenu des trois contraintes décrites précédemment, nous avons opté pour une approche *panel-up*. Les résultats *site-centric* sont considérés comme des informations auxiliaires dont on connaît le total sur la population. Or le principe fondamental en théorie des sondages est que « lorsqu'on dispose d'une information auxiliaire, il faut chercher à l'utiliser » (Ardilly,

1. Acronyme désignant Google, Apple, Facebook et Amazon, les quatre grandes firmes américaines qui dominent le marché du numérique.
2. Alliance pour les Chiffres de la Presse et des Médias.

2006). L'idée est donc d'utiliser cette information par l'introduction de contraintes de calage supplémentaires dans le redressement de l'échantillon (Dudoignon *et al*., 2012). Les données *site-centric* d'environ 400 entités ont alors été transmises à Médiamétrie. On entend par données l'ensemble des logs de connexions collectés par les outils de mesure *site-centric*.

Mise en cohérence des données site-centric *et panel*

Les données *site-centric* ne sont pas nativement comparables à celles mesurées pour la même entité au sein du panel. Elles diffèrent en particulier sur deux aspects : la couverture géographique et les terminaux pris en compte. En effet, la mesure *site-centric* comptabilise les connexions réalisées depuis tous les terminaux (ordinateurs, mobiles, tablettes, consoles de jeux, etc.) et quel que soit le pays de connexion. Afin d'introduire des résultats *site-centric* en tant que contraintes de calage dans le redressement du panel, les champs doivent être parfaitement comparables. Une étape de pré-traitement des données *site-centric* a par conséquent été mise au point afin d'assurer cette mise en cohérence. Les données *site-centric* sont tout d'abord filtrées sur le terminal objet de la mesure, dans le cas présent, l'ordinateur. Les connexions depuis l'étranger sont ensuite écartées. D'autres filtres, plus techniques, sont également appliqués et permettent d'exclure notamment les logs de connexions réalisées par des robots.

La dernière étape consiste à agréger les URLs de manière homogène entre les deux mesures. L'objectif de cette dernière étape est de garantir l'adéquation de ces variables auxiliaires entre l'échantillon, le panel, et la population. Cette adéquation ne sera toutefois garantie que si l'ensemble des urls des différentes entités sont tagguées.

Les difficultés rencontrées

Les difficultés rencontrées ont concerné tout d'abord la représentativité des entités introduites dans le redressement du panel. En effet, on ne dispose malheureusement pas de résultats *site-centric* pour l'intégralité des contenus du Web. Certains acteurs ne souhaitent pas souscrire à un dispositif de mesure *site-centric*. D'autres acteurs disposent de mesures propriétaires non certifiées par un tiers de confiance. De plus, il était difficilement envisageable d'introduire ces 400 entités comme contraintes de calage du panel. Nous avons donc choisi de faire une sélection raisonnée d'entités en s'imposant de n'intégrer que des entités dont le nombre de visiteurs dans le panel était supérieur à 100 individus et de minimiser la corrélation entre les entités introduites.

La base des entités finalement choisies devait respecter les contraintes suivantes :

- toucher de manière homogène l'ensemble des cibles de population en termes de sexe, d'âge et de catégorie socio-professionnelle ;

- être variée en termes de catégories de contenu (actualité, voyage, automobile, etc.) ;

- être d'une taille limitée afin de permettre la convergence de l'algorithme de redressement et de ne pas pénaliser la distribution des poids de redressement ce qui aurait pour conséquence de limiter le gain de précision.

Finalement, un peu plus de 150 entités ont été retenues pour intégrer la base de redressement du panel. L'introduction de ces contraintes quantitatives dans le redressement a un impact direct sur la qualité des poids de redressement. Le rapport de poids est plus important, on constate une accumulation de poids vers les bornes ce qui conduit à une perte de précision et à une plus grande instabilité des résultats (Roy *et al*., 2001).

Le redressement est à ce jour réalisé à l'aide de la macro CALMAR (Sautory, 1993). Des tests sont menés avec de nouveaux algorithmes de redressement permettant de résumer les contraintes *site-centric*, calage sur composantes principales, (Goga *et al*., 2011) ou d'introduire une tolérance sur l'atteinte des objectifs, redressement *ridge* (Alleaume *et al*., 2013), ces approches permettant soit d'améliorer la qualité des poids de redressement, soit d'introduire un plus grand nombre d'entités.

Extension de la méthode à la mesure Internet Global

La mesure de référence du média Internet est, depuis octobre 2017, la mesure de l'Internet Global, i.e. sur les trois écrans. La mesure Internet Global repose sur les trois panels décrits précédemment, ces derniers ayant une partie commune. En effet, certains panélistes appartiennent à plusieurs panels et sont mesurés sur plusieurs types de terminaux. En septembre 2018, le nombre de panélistes mesurés sur

plusieurs de leurs écrans est de 6 000 individus. Les trois panels Internet sont rapprochés par fusions statistiques pour produire les résultats d'audience sur les trois écrans, en tenant compte des duplications entre écrans.

La mesure *site-centric* décrite dans la partie précédente permet l'identification du terminal utilisé par l'internaute pour se connecter, mais sans distinction possible du téléphone mobile et de la tablette. À l'instar de l'hybridation réalisée pour la mesure d'audience Internet sur ordinateur, une hybridation est réalisée sur la mesure d'audience Internet en mobilité, issue d'une première fusion statistique entre les panels sur téléphones mobiles et tablettes. Une seconde fusion sous contrainte de conservation des poids est ensuite effectuée avec le panel ordinateur pour créer la mesure hybride de l'Internet Global.

Une mesure hybride pour la télévision

Comme évoqué précédemment, la mesure d'audience par panel ne permet pas toujours de mesurer finement des usages très morcelés. C'est le cas de Médiamat dont les 5 000 foyers sont insuffisants pour proposer des résultats quotidiens aux chaînes thématiques reçues exclusivement par le satellite (*via* CanalSat), l'ADSL, la fibre optique ou le câble.

Pour répondre au besoin de valorisation des chaînes thématiques, c'est l'approche *log-up* qui a été retenue car elle permet d'apporter des informations complémentaires à ces chaînes et à des coûts faibles, critère toujours important mais particulièrement pour cette catégorie d'acteurs dont les budgets en études marketing sont limités. Nous ne traitons ici que de données concernant le téléviseur pour des chaînes de télévision (c'est-à-dire des flux *live* et non de la vidéo à la demande – VOD). Les modèles de distribution des espaces publicitaires sont en effet très différents entre le *live* et les services de VOD ou les plateformes digitales, en tout cas pour le moment, en France.

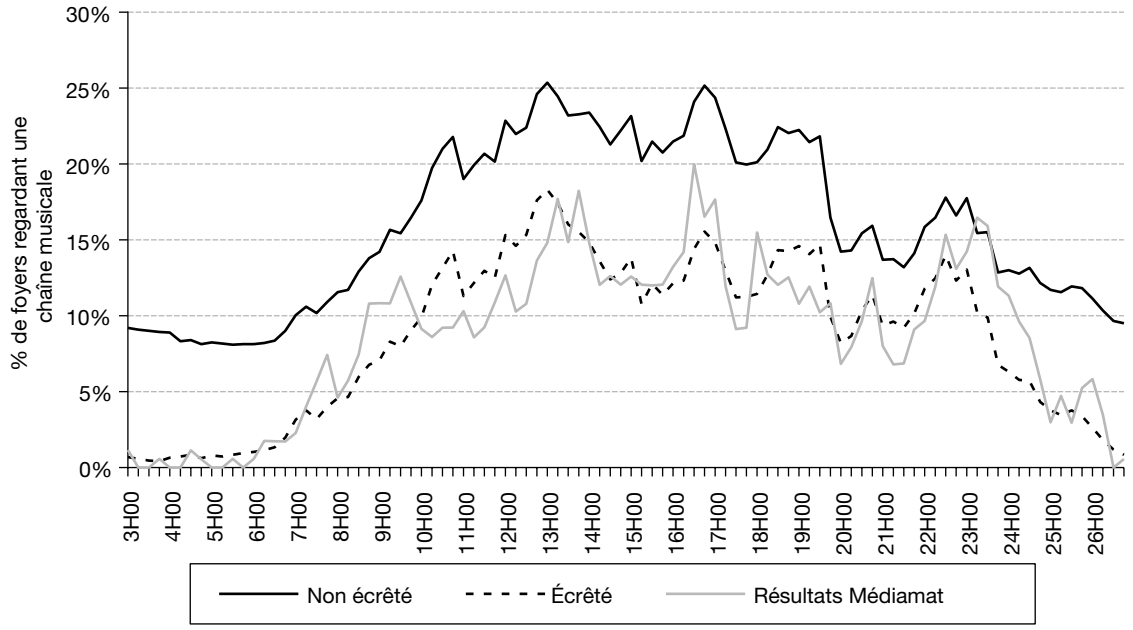
Pour bien comprendre la solution mise au point par Médiamétrie pour la mesure hybride des chaînes thématiques, il faut avant tout comprendre quelles sont les différences entre les usages des décodeurs et les audiences individuelles. On observe pour commencer des écarts entre l'usage d'un décodeur et l'usage du poste auquel ce décodeur est relié. Quelques exemples : le décodeur peut remonter des *logs* ne correspondant pas à une activité humaine, comme des reboots automatiques. Par ailleurs, le décodeur peut être allumé et la télévision éteinte : ce cas est très fréquent en particulier la nuit. Mais on observe surtout des écarts entre les usages d'un téléviseur et les audiences individuelles car le média TV reste avant tout un média familial avec une part importante d'audiences conjointes (i.e. lorsque plusieurs individus regardent simultanément le même poste). Les audiences conjointes représentent ainsi environ 40 % du temps passé devant la télévision par individu de 4 ans et plus, avec des pics pouvant aller à 60 % sur certaines tranches horaires du week-end (source : Médiamétrie//Médiamat).

Nous avons ainsi retenu une méthode en deux grandes étapes. La première consiste en un passage du niveau décodeur au niveau poste de télévision. On commence par un pré-traitement des *logs* bruts afin de nettoyer la donnée des *logs* techniques et de constituer des tickets d'audience. Pour chaque consommation d'une chaîne, on obtient une donnée du type : heure de début, heure de fin, identifiant de la chaîne. On poursuit ensuite par une étape d'écrêtage qui vise à supprimer les usages du décodeur lorsque le téléviseur est probablement éteint. Pour cela, on raccourcit les tickets les plus longs. Les paramètres de la fonction d'écrêtage peuvent être estimés à partir des observations de durées des tickets dans le panel Médiamat sur le même univers/opérateur (figure I).

La seconde étape consiste à individualiser les tickets d'audience au niveau poste obtenus à l'étape précédente. C'est cette seconde étape qui présente le plus de difficultés.

L'approche que nous avons retenue est une modélisation basée sur la connaissance du profil socio-démographique des décodeurs à individualiser (nombre de personnes au foyer, sexe, âge, CSP et lien de parenté des individus). Les individus du foyer étant connus, il nous reste à déterminer à chaque instant qui regarde la télévision quand celle-ci est allumée. Avec cette approche nous n'utilisons donc pas l'exhaustivité des données voie de retour collectées par les opérateurs mais seulement celles d'un échantillon d'abonnés qui acceptent de renseigner les caractéristiques de leur foyer et autorisent l'opérateur et Médiamétrie à avoir accès aux données d'usage TV de leur décodeur. L'ensemble des données est totalement anonymisé. Même si l'exhaustivité des données n'est pas utilisée, le coût marginal de recrutement

Figure I
Effets de l'écrêtage sur une chaîne musicale



Champ : simulation de la fonction d'écrêtage sur un échantillon de foyers abonnés à un opérateur français.
Source : données voie de retour dudit opérateur.

d'un panéliste nous permet de constituer un échantillon de taille importante à des coûts très réduits. On répond ainsi au besoin des chaînes thématiques. Une individualisation des audiences sans ces informations complémentaires n'est que difficilement envisageable.

L'individualisation de l'audience se base sur des modèles de Markov cachés qui peuvent être représentés schématiquement comme sur le schéma (Rabiner, 1989 ; Rabiner *et al*., 1993).

Pour le cas, qui nous intéresse, le temps peut être découpé en pas de 5 minutes (on peut choisir un découpage plus ou moins fin). On a alors :

- des observations Y qui correspondent aux chaînes regardées à la télévision que l'on regroupe en thématiques du type jeunesse, sport, cinéma, etc. Y_n est la thématique majoritaire sur le $n^{\text{ème}}$ pas de temps ;

- un phénomène caché X, qui correspond aux individus devant la télévision. X_n donne l'ensemble des individus du foyer présent devant le poste à l'instant n ce qui permet de conserver les corrélations entre individus d'un même foyer et donc globalement les niveaux d'audience conjointe.

Nous avons choisi des modèles de Markov cachés car leurs propriétés caractéristiques décrivent parfaitement le phénomène à modéliser :

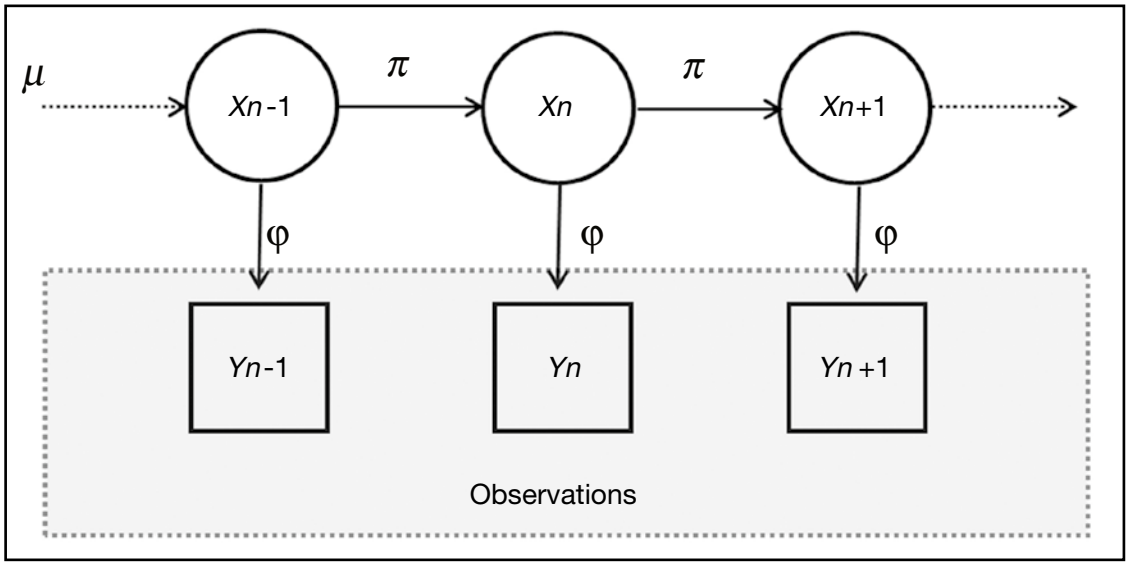
- un processus avec une mémoire courte : pour savoir qui regardera la TV à l'instant $n + 1$, il suffit de savoir qui la regarde à l'instant n. Il n'est pas nécessaire de connaître tout le passé des présences devant le téléviseur ;

- des observations à travers un canal sans mémoire : la chaîne regardée à l'instant n ne dépend que des individus présents devant la télévision au même instant.

Les états possibles pour le processus X dépendent de la taille du foyer mais aussi de sa composition. Pour un foyer d'une personne, la modélisation est inutile (c'est l'individu du foyer qui regarde la télévision). Pour un foyer de deux personnes, par exemple un couple, 3 états sont possibles : la personne de référence seule, le conjoint seul, le couple. Pour un foyer de 3 personnes, par exemple un couple avec un enfant, 7 états sont possibles : la personne de référence seule, le conjoint seul, l'enfant seul, la personne de référence avec l'enfant, le conjoint avec l'enfant, le couple, le couple avec l'enfant. On peut assez facilement démontrer que, pour un foyer de taille k, le nombre d'états possibles est de $2^k - 1$.

Nous avons mis en place une typologie des foyers qui décrit toutes les compositions à prendre en compte : une personne au foyer, deux personnes au foyer (un couple), deux personnes au foyer (un parent isolé et son

Schéma I
Représentation schématique d'un modèle de Markov caché

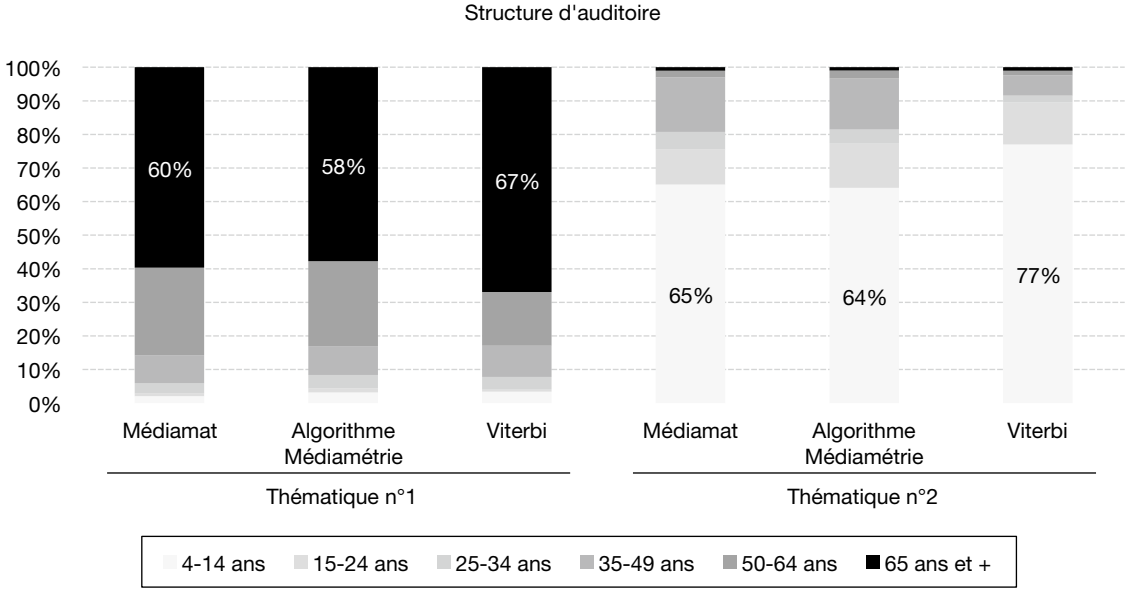


Lecture : la chaîne de Markov {*Xn*} n'est pas directement observée. Les observations {*Yn*} sont générées à travers un canal sans mémoire, c'est-à-dire que chaque *Yn* ne dépend que de l'état *Xn* au même instant.

enfant), trois personnes au foyer (un couple et un enfant), trois personnes au foyer (un parent isolé et deux enfants), trois personnes au foyer (trois adultes), etc. À chacun de ces types de foyer correspond un sous-modèle caractérisé par un jeu de paramètres $M = (\mu,\pi,\varphi)$ où μ est la loi initiale, π la matrice de transition et φ les probabilités d'observation. Tous les paramètres peuvent être estimés simplement à partir des données du panel Médiamat qui en ce sens nous sert d'échantillon d'apprentissage.

Dès lors que les paramètres du modèle sont connus, il suffit d'estimer les présences devant

Figure II
Comparaison d'algorithmes – exemple de résultats sur deux thématiques aux profils très marqués



Lecture : l'auditoire de la thématique n° 1 est composé à 60 % d'individus de 65 ans et plus dans le panel Médiamat. Une estimation des présences avec l'algorithme de Viterbi conduirait à une sur-estimation des plus âgés (67 %). L'algorithme proposé par Médiamétrie donne des résultats plus proches de la réalité du panel avec 58 %.
Champ : structure d'auditoire sur 2 thématiques.
Source : simulation de l'estimation des présences sur le panel Médiamat.

chaque téléviseur. En général, on cherche à estimer la suite $\{Xn\}$ la plus probable et on utilise pour cela l'algorithme de Viterbi (programmation dynamique) qui permet de trouver la solution sans avoir à parcourir l'ensemble des possibilités. Cette approche ne convient pas pour notre problématique car la solution la plus probable conduit à des comportements estimés caricaturaux (uniquement des enfants devant des chaînes jeunesse, etc.) et ne permet pas de reproduire la diversité des comportements. On préfère donc utiliser un algorithme avec une composante aléatoire.

Le panel Médiamat nous sert alors aussi d'échantillon test pour le choix de l'algorithme. On estime les présences en appliquant l'algorithme d'individualisation aux données du panel Médiamat puis on compare les résultats ainsi obtenus avec ceux de Médiamat. Les comparaisons ne sont pas faites de manière unitaire (foyer par foyer) car les résultats publiés sont des moyennes et des compensations peuvent avoir lieu. On compare donc les principaux indicateurs d'audience par thématique et par chaîne et on choisit l'algorithme qui minimise ces écarts. La figure II donne une illustration des comparaisons réalisées pour construire l'algorithme.

* *
*

L'émergence des données massives, l'or noir du 21e siècle, et le développement des possibilités de stockage et de traitement de ces données a laissé entrevoir la perspective d'une fin des mesures d'audience au profit de dispositifs de mesure plus précis, plus fiables et moins coûteux (Vanheuverzwyn, 2016).

Nous avons pu montrer dans la première partie que les enjeux de qualité concernaient tout autant les données massives que les données d'enquêtes. Au travers des deux exemples d'approches hybrides, il apparaît clairement que la qualité réside également dans les traitements, les modélisations qui peuvent être appliqués. Une donnée irréprochable pourrait conduire à des résultats incohérents ou non pertinents en particulier si on perd de vue le besoin des utilisateurs.

Plus qu'une fin, c'est une évolution, voire une révolution, des mesures d'audience vers les mesures hybrides que nous observons aujourd'hui. La nécessité de tirer parti des avantages des différents dispositifs d'observation pour en créer d'autres, plus complexes mais plus riches, est indéniable. Cette perspective ouvre des champs d'application nouveaux en matière de recherche et développement. En premier lieu en théorie et pratiques des sondages. En effet, l'exploitation des données massives peut être considérée comme une réponse à l'augmentation croissante de la non-réponse aux enquêtes. Si la question du compromis entre biais et variance, biais des estimateurs et variance des poids de redressement, a été abordée, elle mérite d'être creusée. Elle pourrait conduire au développement d'algorithmes de redressement plus performants, permettant de tenir compte d'un plus grand nombre de contraintes de calage, ou à la mise au point de nouveaux modèles d'hybridation basés sur des techniques d'imputation ou d'appariement statistique. La recherche en *machine learning* offre également des perspectives d'enrichissement des données massives intéressantes pour des problématiques de ciblage mais aussi potentiellement pour la mesure d'audience.

Mais les réponses que nous pourrons apporter aux besoins d'observation du comportement des individus devront, comme elles l'ont toujours été, s'inscrire dans le cadre du respect de la vie privée et des contraintes juridiques liées au traitement des données à caractère personnel. Et il ne s'agit pas tant là d'une question juridique que d'une question éthique (Tassi, 2014). L'entrée en vigueur du Règlement Général européen sur la Protection des Données et l'ensemble des débats publics qui ont eu lieu en amont, ont permis de mettre en lumière les dérives en matière de mesure des usages sur Internet. Les enquêtes, pour lesquelles le consentement de l'individu est inhérent, retrouvent dès lors un rôle central. □

BIBLIOGRAPHIE

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