FUSION STATISTIQUE DE DONNEES D'ENQUETES : DERNIERES AVANCEES POUR LES MESURES D'AUDIENCE

Lorie Dudoignon[1]

[1] *Médiamétrie, 70 rue Rivay, 92532 Levallois Cedex et ldudoignon@mediametrie.fr*

Résumé. Sur ces dernières années, on a pu constater une explosion des technologies numériques portable et des consommations associées. Les marques digitales se sont naturellement très vite adaptées à ces évolutions en déclinant leurs contenus sur tous les écrans via des applications sur mobile et tablette. L'audience Internet sur chacun des écrans est mesurée par Médiamétrie auprès de panélistes équipés de « meter » sur au moins un type d'écran. Jusqu'en 2017, Médiamétrie proposait ainsi 3 mesures de référence (une par type d'écran). Ces dispositifs étaient alors complétés par l'étude Internet Global. Cette étude qui, comme son nom l'indique, consiste en une mesure globale des marques digitales sur les différents écrans, était réalisée par fusion statistique des 3 panels vers une étude receveuse ou Hub. Depuis Octobre 2017, l'étude Internet Global est devenue la mesure de référence des audiences Internet. A cette occasion la méthodologie de fusion des 3 panels a été entièrement revue grâce, notamment, au recrutement de panélistes mesurés sur plusieurs types d'écran. L'objectif de cette présentation est de détailler la nouvelle méthodologie mise en place en expliquant les raisons qui ont pu guider nos choix.

Mots-clés. Audience, panel, fusion statistique

Abstract. In recent years, there has been an explosion of portable digital technologies and of associated consumptions. Digital brands have naturally adapted very quickly to these evolutions by displaying their contents on every screen via mobile and tablet applications. The Internet audience on each screen is measured by Médiamétrie with panelists equipped with meter on at least one type of screen. Until 2017, Médiamétrie offered 3 benchmark measurements (one per screen type). To complete, the Global Internet study was providing a global measurement of digital brands on the various screens, by statistical matching of the 3 survey panels on a Hub survey. Since October 2017, the Global Internet Study has become the benchmark for Internet audiences. To do so, the statistical matching methodology of the 3 survey panels was entirely revisited especially with to the recruitment of panelists measured on several screen types. The purpose of this presentation is to explain the new methodology and the reasons for our choices.

Keywords. Audience, survey panel, statistical matching

1. Evolution du marché Internet

L'adoption de nouvelles technologies par les foyers français s'est vivement accélérée sur les dernières années. Ainsi en moins de 10 ans plus de 70% des foyers français possèdent au moins un smartphone alors qu'il a fallu plus de 40 ans au téléviseur pour atteindre un tel niveau de pénétration (mais ça, c'était au siècle dernier).

Fin 2017, 92% des foyers possèdent au moins un écran Internet (ordinateur, tablette ou smartphone).

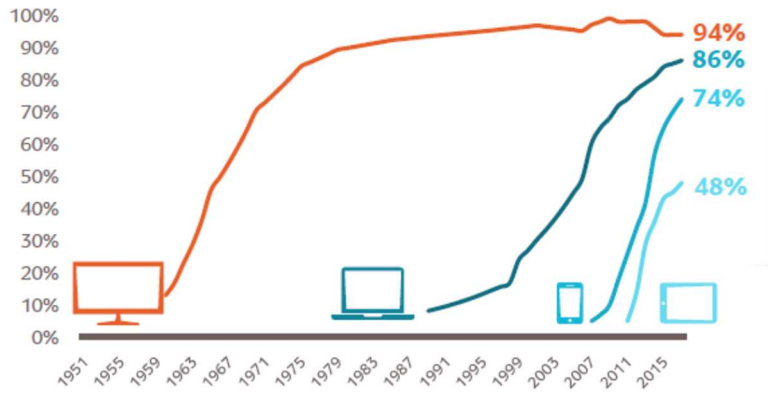

Figure 1 : Evolution de l'équipement des foyers – Source : La référence des équipements multimédia – Home Devices.

La multiplication des écrans au sein des foyers s'est naturellement accompagnée d'une croissance importante des usages Internet.

Fin 2017, on compte en France près de 52 millions d'internautes chaque mois (83% de la population) et 42,2 millions au quotidien. En 10 ans, la population internaute a progressé de 40%.

Figure 2 : Nombres d'internautes par mois – Sources : Médiamétrie/NetRatings – Audience Internet Ordinateur pour 2007 & Internet Global pour 2017.

Cette croissance est particulièrement impressionnante sur le mobile :

Avec plus de 30 millions de mobinautes chaque jour (48%), le smartphone est le 1er écran pour se connecter au quotidien, devant l'ordinateur. Près d'1 individu sur 5 n'utilise même que son mobile pour surfer. Le smartphone pèse 40% du temps passé sur Internet, et 65% chez les jeunes (15-24 ans).

Dans ce contexte, une mesure globale (sur tous les types d'écran) des marques digitales devient incontournable pour le marché.

Cette mesure née en 2013 permettait initialement un rapprochement des mesures Ordinateur et Mobile. Elle a été enrichie en 2015, par des résultats sur les tablettes.

Si au départ, cette étude permettait un complément d'analyse par rapport aux 3 études de référence, elle est depuis Octobre 2017 devenue la mesure de référence.

2. Internet Global – 1ère génération (2013-2017)

La première mesure globale de l'Internet fixe et mobile, reposait sur trois sources de données : le panel d'audience Internet Médiamétrie//NetRatings pour l'ordinateur, le panel d'audience de l'Internet mobile et une étude pivot, le « Hub ».

2.1. Vue d'ensemble sur le dispositif initial

Le Hub est constitué d'environ 8300 personnes qui répondent à un questionnaire sur leurs habitudes de consommation sur les deux supports.

La première étape composée de deux fusions permet de transférer des comportements d'audience des panels Ordinateur et Mobile (bases donneuses) vers les individus du Hub (base receveuse). L'association donneur/receveur est faite en minimisant une fonction de distance basée sur les habitudes d'écoute de chaque écran et en contrôlant le nombre de réplications d'un donneur.

La deuxième étape consiste à redresser la base receveuse ainsi enrichie sur les principaux résultats d'audience par écran, afin de limiter les écarts avec les résultats des deux enquêtes donneuses qui constituent les références sur le marché.

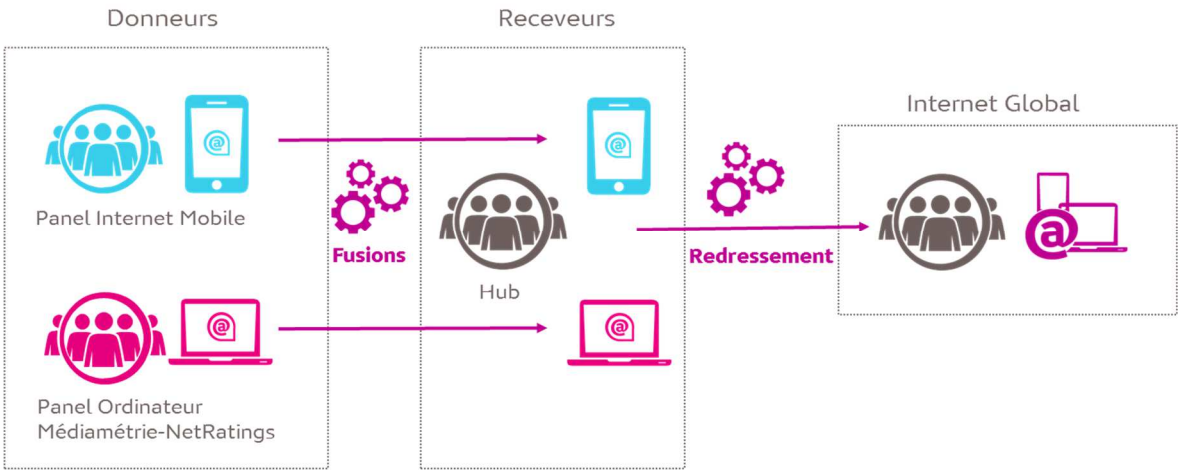

Figure 3 : Internat Global 1ère génération - Schéma.

Une troisième fusion a été ajoutée en 2015, selon le même schéma, pour intégrer le panel Tablette.

2.2. Les axes d'amélioration

Malgré l'étape de redressement, il n'est pas possible avec ce dispositif de garantir sur l'ensemble des marques digitales et l'ensemble des cibles marché, une égalité des résultats par écran entre Internet Global et les panels de référence. Sur le marché des médias, des écarts de résultats entre deux études sont très perturbants pour les utilisateurs-clients. C'est pourquoi, nous avons décidé de faire de l'étude Internet Global la référence qui permet à la fois de disposer de résultats tout écrans Internet et par écran.

Par ailleurs, le socle de cette première étude Internet Global était un Hub dont la fréquence de mise à jour n'était que biannuelle, pour des productions de résultats mensuelles. Nous avons choisi pour la 2ème génération Internet Global un rapprochement direct des 3 panels basé sur les intersections des 3 panels entre eux (i.e. des panélistes bi ou tri-mesurés).

3. Internet Global – 2ème génération (Octobre 2017)

Les 3 panels Ordinateur, Mobile et Tablette ne sont plus vus comme 3 dispositifs de mesure indépendants mais comme un seul panel de mesure Internet, avec une participation des panélistes « à la carte ». Les panélistes peuvent, en effet, accepter d'être mesurés sur un, deux ou trois types d'écran.

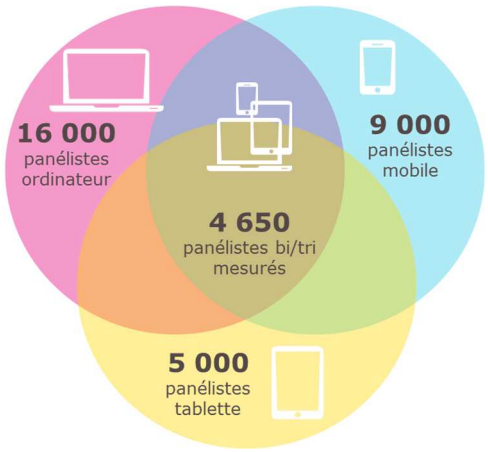

Figure 4 : Internet Global 2ème génération – Panel « à la carte » - Objectifs 2018.

Les rapprochements sont opérés deux à deux sur la base des communs.

Dans un premier temps, on va ainsi constituer deux bases : une base mobilité, obtenue par fusion des données Mobile et Tablette et une base Ordinateur.

En effet, sur ces deux univers Mobilité et Ordinateur, nous disposons de données exhaustives (données site-centric) que nous utilisons pour une hybridation des résultats panel selon l'approche Panel-UP développée par Dudoignon et Zydorczak (2012). Dans cette approche la donnée site-centric va être considérée comme une information auxiliaire que l'on intègre dans le redressement de l'enquête.

Aussi, lors de l'étape suivante qui consiste à rapprocher les données Mobilité et Ordinateur, on souhaite conserver les résultats de cette hybridation et donc les poids associés à chaque « observation ». C'est pourquoi nous nous sommes orientés vers des fusions avec contraintes.

On considère 2 sources A et B (de tailles respectives n_A et n_B) que l'on souhaite fusionner :

- $D_{ij} = d(a_i, b_j)$ est la distance entre les observations i de A et j de B
- $w(a_i)$ et $w(b_j)$ sont les poids des observations i de A et j et B
- $w(a_i, b_j)$ = poids du couple (a_i, b_j) à l'issue de la fusion statistique.

On va donc chercher à minimiser $Z = \sum_{i=1}^{n_A} \sum_{j=1}^{n_B} D_{ij} \times w(a_i, b_j)$ sous les contraintes :

- $\sum_{i=1}^{n_A} w(a_i, b_j) = w(b_j)$
- $\sum_{j=1}^{n_B} w(a_i, b_j) = w(a_i)$

Si dans le domaine de la Statistique d'enquête, les solutions pour ce type de problème sont peu documentées, on trouve une bibliographie importante sur la question du « Problème de transport » qui est en définitive exactement la même chose. C'est pourquoi nous parlerons de fusion par transport de poids.

Dans un problème de fusion statistique, il y a deux éléments clefs qui sont l'algorithme de jumelage et la fonction de distance. Pour chacune des fusions deux à deux, il nous faut trouver une fonction de distance adéquate. Nous souhaitons, pour cette fonction de distance, utiliser les données disponibles sur les panélistes réellement mesurés sur plusieurs écrans et nous nous sommes donc orientés vers des distances procustéennes.

Nous reviendrons dans la présentation, plus en détail, sur la méthodologie retenue. Nous présenterons aussi quelques résultats de tests qui ont pu être réalisés pour valider l'approche ou choisir entre différents packages R.

Bibliographie

Dudoignon, L. & Zydorczak, L. (2012). *Enquête et données exhaustives : un nouveau défi pour les mesures d'audience*, 7ème Colloque Francophone sur les Sondages, Rennes.
Fisher, N. (2004), *Fusion statistique de fichiers de données*, Thèse de doctorat, Montpellier.
Mansi, S.G. (2011), *A Study on Transportation Problem, Transshipment Problem, Assignment Problem and Supply Chain*, Thèse de doctorat, Rajkot.
Santini, G. (2001). Méthode de fusion procustéenne, *Traitements des fichiers d'enquêtes*, Michel Lejeune.