ENQUETES ET DONNEES EXHAUSTIVES : UN NOUVEAU DEFI POUR LES MESURES D'AUDIENCES

Lorie Dudoignon[1] & Lila Zydorczak[2]

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

Résumé. Le développement des offres numériques (Haut Débit, TV par ADSL, 3G …) a considérablement modifié la consommation médias des français sur ces dernières années. Les offres médias se sont enrichies. Les contenus sont désormais disponibles en dehors des heures de diffusion (Podcast radio, TV de rattrapage). Et les frontières entre les médias sont devenues floues (consommation de la radio ou de la TV sur Internet, surf sur un écran de télévision …), ce phénomène s'amplifiant chaque jour avec les évolutions technologiques, comme l'émergence des TV connectées.

Le développement du numérique a d'autre part permis de disposer de nouvelles informations offertes par les voies de retour. Celles-ci permettent ainsi de disposer de résultats exhaustifs sur le nombre de téléphones, d'ordinateurs ou de décodeurs TV connectés. Ces voies de retour ont, en contrepartie, des limites : elles mesurent des machines sans tenir compte du ou des individus placés devant.

La mise en cohérence de données issues d'échantillons avec des données exhaustives mais de granularité différente, est une réflexion statistique de base. Mélanger deux sources d'information de natures et de niveaux différents, les croiser mutuellement pour en créer une troisième, plus fine ou plus riche, est une démarche naturelle, que nous qualifierons d'hybride. Pascal Ardilly (2006) écrit ainsi : « Le principe fondamental à retenir est le suivant : lorsqu'on dispose d'une information auxiliaire, il faut chercher à l'utiliser …. ». Le contexte médias n'échappe pas à cette démarche, et Médiamétrie réfléchit en permanence à l'évolution de ses dispositifs de mesure d'audience pour suivre l'évolution des comportements médias. Avec le développement des offres numériques, nous sommes confrontés à la question suivante : comment intégrer ces nouvelles sources d'informations dans la mesure d'audience traditionnellement faite par des enquêtes ?

Cette communication a pour but de faire un état des lieux des données disponibles pour la mesure d'audience de l'Internet, sur un ordinateur « fixe » ou un objet mobile ainsi que pour la mesure d'audience TV et des problématiques d'hybridation ainsi soulevées. Nous exposerons les solutions proposées par Médiamétrie dans ces domaines.

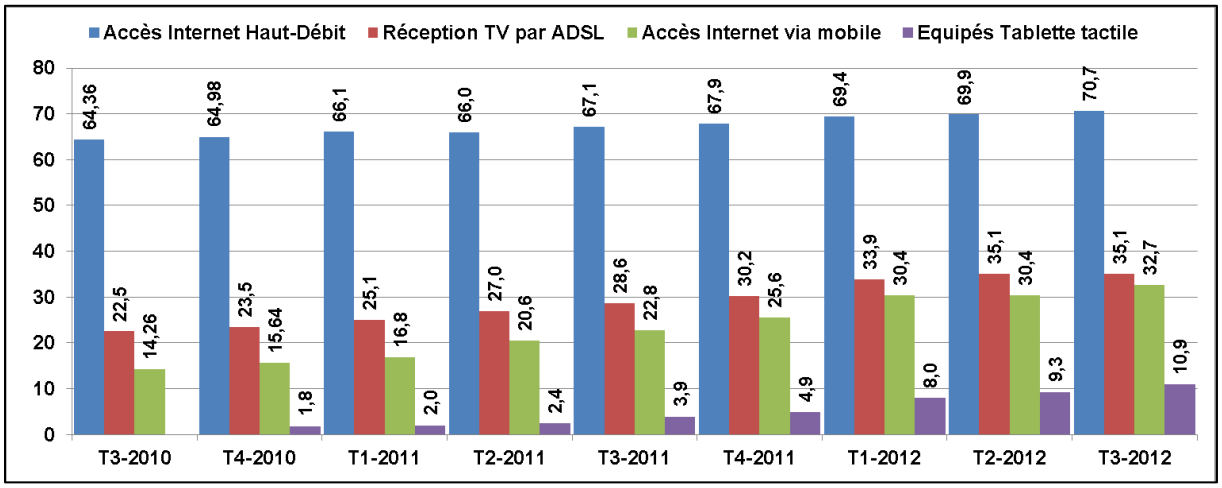
1. Problématique

1.1. Evolution du contexte média

Le développement des offres numériques a considérablement modifié la consommation médias des français ces dernières années. En premier lieu, on a pu observer une démocratisation de l'accès Internet à domicile. On note au troisième trimestre 2012 que plus de 70% des foyers français ont un accès haut débit soit une progression de 6 points en 2 ans. L'offre de chaînes s'est de plus élargie grâce au passage à la TNT et au développement des offres "Triple-Play" : aujourd'hui plus de 35% des foyers ont accès à la TV par l'ADSL contre 22% il y a deux ans. L'accès au média en mobilité avec la téléphonie mobile et/ou les tablettes tactiles se démocratise très rapidement. En deux ans, le taux d'accès Internet via mobile est passé de 14% à 33%, tandis que le taux d'équipement en tablette est déjà à 11%. Les contenus Radio et TV sont maintenant disponibles en dehors des heures de diffusion (Podcast et Catch-Up).

Ce développement des offres numériques conduit à une évolution très rapide des habitudes que Médiamétrie peut suivre grâce à plusieurs études : la Référence des Equipements Multimédias, l'Observatoire des Usages Internet, Global Catch-Up et Global TV notamment.

Graphique 1 : Taux d'équipements numériques des foyers français

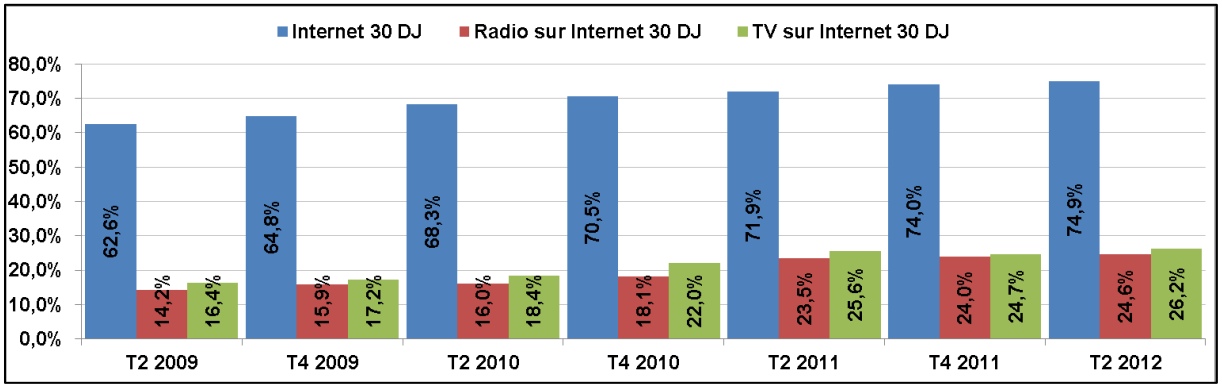


Source : Médiamétrie - Référence des Equipements Multimédias

L'avènement du tout numérique a largement contribué à la délinéarisation des contenus médias. En effet, il favorise l'éclosion de nouvelles formes de consommation des médias via un archivage, qui permet d'accéder sous conditions (récence, acquittement d'un droit, abonnement, etc…) à un grand nombre de contenus et ce, indépendamment de l'horaire de diffusion en "live". Le développement de la Vidéo à la Demande, du Podcast, de la Catch-Up et du time-shifting en témoigne.

Historiquement, le contenu écrit était disponible dans la presse, le contenu audio à la radio et le contenu audiovisuel sur le poste de télévision. L'apparition d'Internet a mis fin à ce cloisonnement puisque tous les médias traditionnels (Presse, Radio et TV) sont désormais en ligne. Ainsi, la frontière entre médias s'amenuise voire disparaît complètement.

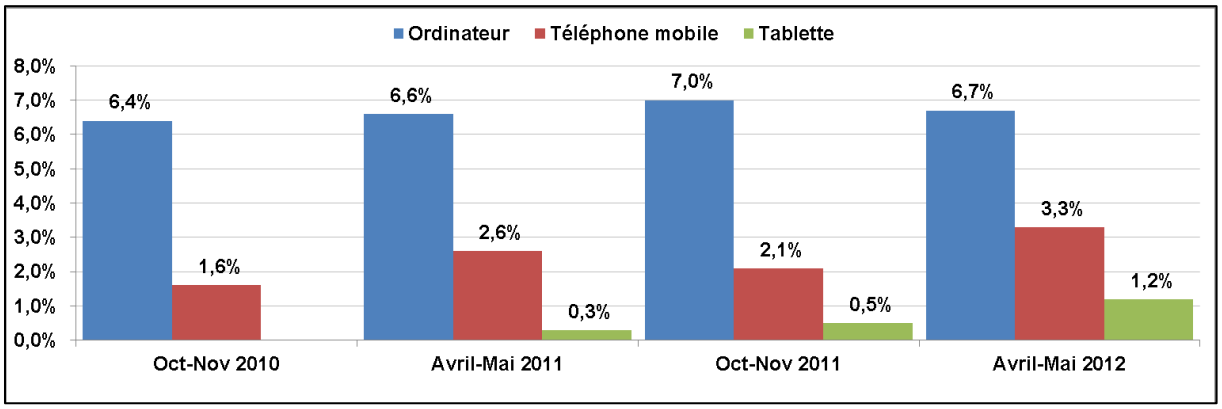
Graphique 2 : Consommation Internet 30 Derniers Jours Radio et TV



Source : Médiamétrie - Observatoire des Usages Internet (Base : individus de 11ans et plus)

Près d'un quart de la population des individus âgés de 11 ans et plus consomme aujourd'hui la TV ou la Radio sur Internet.

Graphique 3 : Consommation TV par support



Source : Médiamétrie - Global TV (Base : individus de 15 ans et plus)

L'hypothèse qu'un écran n'est dédié qu'à une seule activité est ainsi dépassée. On a aujourd'hui accès à la télévision sur tous les écrans que ce soit via un ordinateur, un téléphone mobile ou une tablette numérique. Même si l'écran de télévision reste le principal support de consommation, on note une augmentation rapide de la consommation de contenus de télévision sur les tablettes tactiles bien que l'émergence de cet équipement soit relativement récente.

1.2. Le tout numérique : une aubaine pour les études médias

Le numérique n'est pas seulement une technologie qui améliore la qualité de la réception ou de la transmission. Au-delà de l'enrichissement des contenus qu'il permet, il ouvre la possibilité de mesure des usages numériques via la voie de retour.

La voie de retour peut être globalement définie comme la récupération de logs de connexion dans lesquels les usages sont tracés. Elle n'est certes pas disponible sur tous les médias ou toutes les normes mais elle permet de tracer l'intégralité des transactions (saisie d'une adresse réseau, changement de chaîne, lancement d'un programme en VOD...) et de reconstruire ainsi fidèlement le comportement de l'utilisateur.

Cette voie de retour est par exemple disponible à partir des décodeurs numériques et des box ADSL pour la TV, via la 3G pour l'Internet mobile et grâce au taggage des sites Internet

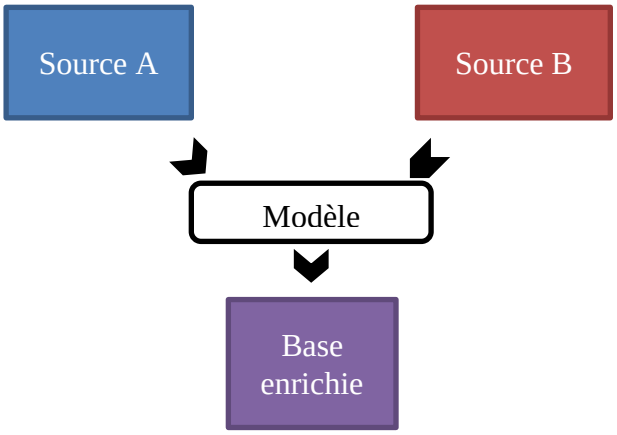
pour l'Internet fixe. Le périmètre de la voie de retour n'est cependant pas le même que celui des enquêtes média. La voie de retour correspond plus à une mesure des machines qu'à une mesure des individus. En particulier, on ne peut pas distinguer dans les logs un usage individuel d'un usage à plusieurs. Cela peut notamment poser problème pour le média TV où l'audience impliquant plusieurs membres d'un foyer est très fréquente.

D'autre part, la couverture de la mesure voie de retour n'est pas la même selon le média. Pour le média télévision, seuls les écrans connectés sont mesurables par voie de retour alors que dans l'enquête média tous les écrans TV sont mesurés. Pour l'Internet fixe, seule la consommation réalisée en France est mesurée dans le cadre de l'enquête média tandis que la voie de retour intègre le surf depuis l'étranger.

1.3. Faire évoluer les dispositifs de mesure d'audience

Médiamétrie a donc considéré ces données issues de la voie de retour comme une aubaine et a naturellement fait évoluer les dispositifs de mesure d'audience pour intégrer ces nouvelles sources d'information en créant des mesures dîtes hybrides.

On entend par mesure hybride le processus qui consiste à « mélanger deux sources d'information de natures et de niveaux différents, les croiser mutuellement pour en créer une troisième, plus fine ou plus riche, est une démarche devenue naturelle » (Livre Blanc des Mesures Hybrides de L'Internet, Médiamétrie-Médiamétrie//NetRatings 2010).



Cela se traduit plus généralement par le mélange de données d'enquêtes avec des données issues de la voie de retour.

Deux approches sont possibles pour les mesures hybrides selon les besoins. Une première qu'on appellera "panel-up" où la mesure voie de retour vient enrichir l'information issue de l'enquête média (le plus souvent un panel). Dans cette approche la mesure voie de retour va être considérée comme une information auxiliaire que l'on intègre dans le redressement de l'enquête. Bien sûr, il faudra faire très attention aux périmètres de chacune des mesures afin que l'information complémentaire vienne améliorer la qualité de l'enquête.

La seconde approche qu'on appellera "log-up" consiste en une qualification des données voie de retour. On construit un modèle à partir des données de l'enquête qui nous permet d'estimer le profil des consommateurs du média. L'intérêt de cette approche est d'apporter des informations complémentaires à des sites/chaînes à faibles audiences et qui n'ont donc pas beaucoup de visibilité dans l'enquête média.

2. La mesure d'audience hybride Internet Mobile

2.1. Objectif et contexte de la mesure

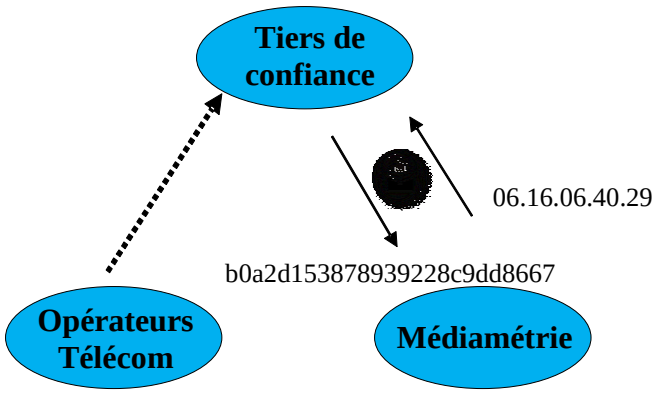
L'objectif de la mesure est de fournir au marché publicitaire l'audience, les usages et le profil des mobinautes en France métropolitaine ; par mobinaute, nous entendons le visiteur de sites ou d'applications via un téléphone mobile. L'étude répond par exemple aux questions suivantes : Quels sont les sites et applications mobiles les plus visités ? Quels sont ceux en affinité avec une cible donnée ? Quel est le profil des visiteurs d'un site donné ? Quels sont les sites qui dupliquent le plus avec un site donné ?

Cette étude née en fin d'année 2010 fût nativement hybride. Elle repose en effet sur deux fondements. En premier lieu, les logs des connexions 3G sont mis à disposition grâce à un partenariat entre les trois principaux opérateurs Orange, SFR, Bouygues Telecom qui fournissent la totalité des logs de connexion anonymisés de l'ensemble des abonnés en France. En second lieu, un panel de 10 000 mobinautes est qualifié sur des critères sociodémographiques et géographiques.

2.2. La nécessité d'un tiers de confiance

Afin de respecter les règles de confidentialité de la CNIL, le recours à un tiers de confiance s'est avéré impératif. Celui-ci permet de faire le lien entre un panéliste Médiamétrie et les logs opérateurs tout en conservant la confidentialité du panel vis-à-vis des opérateurs.

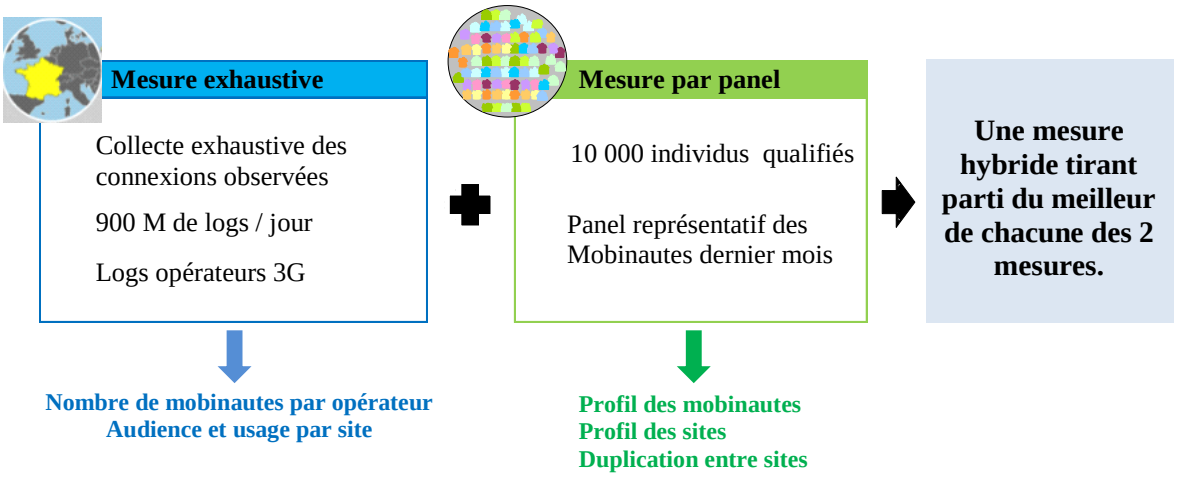
L'étude repose ainsi sur un fonctionnement tripartite. Les opérateurs connaissent le surf de l'ensemble des mobinautes mais ne peuvent identifier les panélistes. Ils sont responsables de l'envoi de la clef de cryptage et du surf associé. Médiamétrie connaît le profil des panélistes et leur numéro de téléphone mais ne dispose pas de leur surf individuel. Le tiers de confiance permet de faire le lien afin de fournir à Médiamétrie le surf des mobinautes panélistes grâce à la clef de cryptage envoyée par les opérateurs.



2.3. Méthodologie de l'étude

Médiamétrie souhaitait tirer le meilleur de chacune des deux mesures. Les logs exhaustifs fournissent le nombre et l'usage des mobinautes au global et par site / application mobile sans erreur d'échantillonnage. L'identité des individus derrière les logs étant tout à fait inconnue, il n'est pas possible de disposer de résultats par cible sociodémographique. Le panel permet quant à lui de dresser le profil de la population des mobinautes ainsi que le profil des sites.

Les niveaux d'audience sont corrigés par redressement afin d'être les plus proches possibles des résultats des logs exhaustifs. Le panel vient aussi pallier la difficulté de faire des calculs complexes sur les logs. En effet, la volumétrie est telle (plus de 900 millions de logs par jour) que les calculs de duplication sont issus du panel. L'article de Vanheuverzwyn et Vouge (2011) présente en détail le dispositif d'étude et les formules de calcul des indicateurs hybrides.



2.4. Avantage et limites

Les avantages de cette méthode sont de fournir, grâce à l'exhaustivité des logs, des audiences et usages par site sans les biais naturels d'une mesure fondée sur un panel (sur la cible ensemble).

De plus, le couplage d'une mesure exhaustive et d'une mesure fondée sur un panel permet de disposer de tous les résultats classiques (classements, profils, duplication, etc...) de pallier les difficultés liées à la volumétrie des logs (notamment la duplication).

Deux principales limites sont néanmoins à considérer. La mesure exhaustive ne couvre que les connexions en 3G (le surf en Wifi n'est pas inclus) et n'est pas disponible pour les BlackBerry. De plus, l'engouement des français pour l'Internet Mobile génère une volumétrie qui augmente très rapidement. Nous sommes dans le champ du Big Data.

3. La mesure d'audience hybride Internet Fixe

3.1. Coexistence de deux mesures complémentaires

Le contexte de la mesure d'audience de l'Internet Fixe (surf depuis un ordinateur qu'il soit fixe ou portable) diffère sensiblement de celui de l'Internet Mobile. En effet, depuis de nombreuses années, deux mesures complémentaires coexistent. Une mesure dite "User Centric" est assurée par Médiamétrie//NetRatings et fondée sur un panel de 25 000 internautes qui permet d'estimer l'audience et l'usage de l'ensemble des sites Internet. Cette mesure repose sur l'installation d'un meter sur l'ensemble des ordinateurs fixes ou portables connectés à Internet qui permet de disposer du surf individuel des panélistes. Un système de mesure dit "Site Centric", mesure propriétaire fondée sur l'insertion d'un marqueur sur chaque page du site, permet de disposer de résultats exhaustifs de consommation du site en termes de pages vues et de visites.

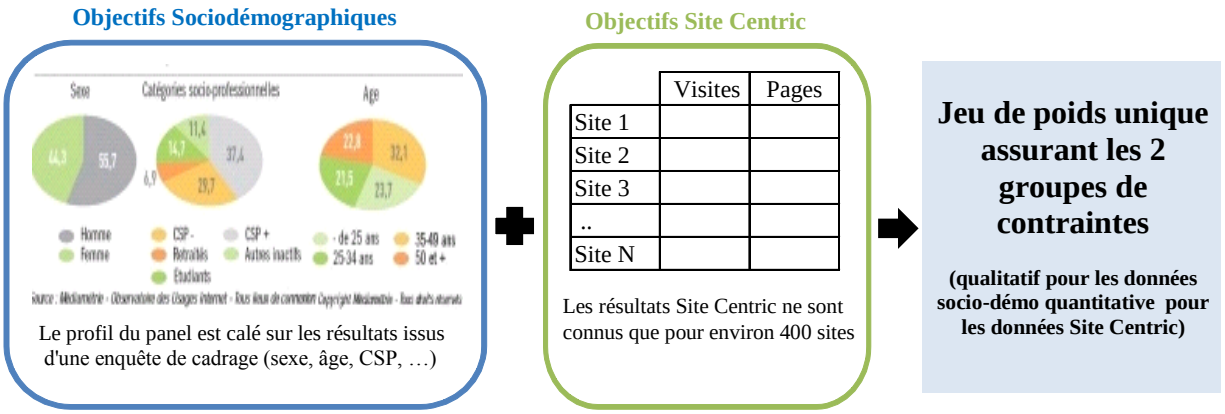
Seuls les souscripteurs de la mesure Site Centric ont accès à leurs résultats. Ils ne peuvent se situer dans leur univers de concurrence avec cette mesure et doivent ainsi se référer au panel Médiamétrie//NetRatings dans cet objectif.

3.2. 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 de la précision apportée par la mesure Site Centric et pas uniquement les sites souscripteurs de cette mesure,
- la donnée Site Centric utilisée doit être cohérente avec le périmètre de mesure du panel,
- la donnée hybride résultante doit être compatible avec les outils de médiaplanning qui ne prennent en entrée que des données individuelles.

Compte tenu des trois contraintes décrites précédemment, nous avons opté pour une approche par redressement.

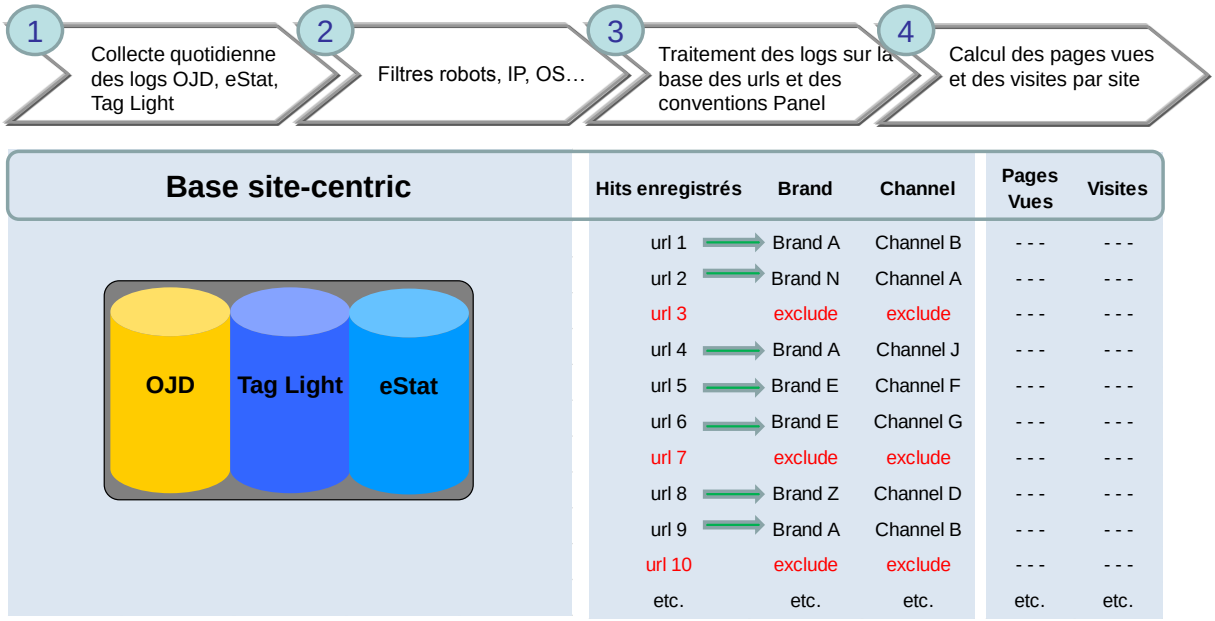


Aux contraintes classiques de redressement (profil des internautes par lieu de connexion), s'ajoutent ainsi des contraintes quantitatives à savoir le nombre de visites par site pour une sélection d'environ 150 sites sur les 400 de départ. Le choix des 150 sites sur les 400 a été fait en tenant notamment compte des critères suivants : faible corrélation entre les entités introduites, sites présentant une audience jugée suffisante, représentativité par cible de la base retenue.

3.3. Mise en cohérence des données Site Centric avec le panel

Les résultats Site Centric ne sont pas nativement comparables à ceux du panel. Ils diffèrent en effet du fait des éléments suivants. La mesure Site Centric tient compte de tous les terminaux : ordinateurs, mobiles, tablettes, consoles, etc. Pour être comparable, les résultats doivent être filtrés sur le terminal ordinateur. La mesure Site Centric tient compte des connexions à l'étranger tandis que le panel est sur un périmètre France métropolitaine. Les connexions depuis l'étranger doivent ainsi être écartées. Les urls marquées dans la mesure Site Centric ne font l'objet d'un contrôle que pour les sites labellisés par l'OJD. Pour les autres, aucune garantie n'est disponible et le périmètre marqué n'est dépendant que de la volonté de l'éditeur. Dans le cadre du panel, les urls marquées doivent respecter des règles et conventions précises.

Afin d'introduire dans le redressement des résultats Site Centric tout à fait comparables à ceux du panel, un recalcul des données Site Centric a été mis au point :



A la fin de ce traitement, on dispose des pages vues et visites pour chacun des sites traités.

3.4. Les difficultés rencontrées

Les difficultés rencontrées ont concerné tout d'abord la représentativité de la base Site Centric. En effet, la base devait respecter les critères suivants : toucher l'ensemble des cibles clés (sexe, âge, catégorie socio-professionnelle), être variée en termes de contenu (actualité, voyage, automobile, etc...), être d'une taille limitée afin de faciliter la convergence de l'algorithme de redressement, … sous contrainte de participation des éditeurs de sites.

De plus l'introduction de plus de 150 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 engendre une moindre précision et une plus grande instabilité des résultats.

Des tests complémentaires sont menés avec de nouvelles techniques de redressement permettant de résumer les contraintes Site Centric (calage sur composantes principales) ou d'introduire une tolérance sur l'atteinte des objectifs (redressement ridge).

4. Perspectives d'application à la mesure d'audience TV

4.1. Contexte de la mesure d'audience TV

Le panel Médiamat de Médiamétrie constitue la mesure de référence. Cette mesure est fondée sur un panel de près de 5000 foyers équipés d'au moins une TV en France métropolitaine. Tous les postes actifs sont mesurés, c'est-à-dire ceux qui servent au moins une fois par mois pour regarder la télévision. Sur chacun de ces postes, un audimètre détecte automatiquement l'allumage de la télévision, les changements de chaîne et la vision d'enregistrement. Les individus du foyer doivent participer à la mesure en déclarant leur présence devant le poste de télévision via une télécommande. Les données stockées par les audimètres sont collectées en continu. En termes de restitution, l'audience TV mesurée est celle réalisée par les individus âgés de 4 ans et plus, au domicile, sur un poste de télévision.

La voie de retour en TV 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 TV connectées (encore assez rares pour l'instant). Dans ces deux seuls cas, les logs de connexion sont éventuellement disponibles auprès de l'opérateur et permettent de savoir sur quelle chaîne ou service est allumé le décodeur. Tout usage fait en dehors du décodeur n'est pas mesuré (si on passe par le tuner TV par exemple).

Le marché souhaite aujourd'hui utiliser la mesure voie de retour pour valoriser des "petites" chaînes.

4.2. Difficultés et méthodes envisagées

Même si techniquement une voie de retour est possible, tous les opérateurs ne collectent pas les données TV. A ce jour, très peu d'opérateurs exploitent la voie de retour TV. Afin de mettre au point une mesure marché, deux points fondamentaux doivent être résolus : la mise à disposition des données par les opérateurs et la certification de ces données.

Par ailleurs, les périmètres mesurés sont différents entre le panel Médiamat et la voie de retour. En effet, d'après la Référence des Equipements Multimédias sur le troisième trimestre 2012, même si plus d'un foyer sur deux (52%) est aujourd'hui équipé d'un décodeur numérique seulement un tiers des postes TV sont reliés à un décodeur numérique. Parmi ce tiers, tous les décodeurs ne sont pas forcément connectés à Internet. La voie de retour ne couvrirait donc au maximum que 33% des foyers en France.

De plus, le décodeur peut être allumé et la TV éteinte et inversement. Ces cas de figure génèrent ainsi des logs dont les durées ne sont pas réalistes lorsque le décodeur est resté allumé par oubli ; des audiences manquantes lorsqu'on ne passe pas par le décodeur.

Enfin, la voie de retour est une mesure "machine" qui ne peut en l'état être rapproché de la mesure individuelle de Médiamat. Derrière un log, on ne peut pas savoir combien d'individus étaient présents derrière le poste. C'est un point clé car le média TV reste familial. Par exemple au mois de septembre 2012, plus de 40% du temps consacré à la télévision est passé à plusieurs devant le même écran.

Par rapport au besoin, exprimé par les clients, de plus de fiabilité dans la mesure des petites chaînes et à la différence de périmètre actuel entre la voie de retour et le panel Médiamat, seules des solutions de type "log-up" sont envisagées à ce jour. Dans ce cadre, Médiamat est utilisé pour qualifier les résultats issus des logs. Dans tous les cas, la modélisation devra se faire en 2 étapes.

Une première étape consiste à transformer les données décodeurs en données TV. Pour cela, il faut supprimer les cas d'usage où le décodeur reste allumé alors que la TV est éteinte. Cela peut se traduire techniquement par un écrêtage des audiences trop longues ou du milieu de la nuit … en essayant de se rapprocher de la distribution des sessions d'écoute TV mesurée dans Médiamat. Comme les données voie de retour ne couvrent aujourd'hui qu'un tiers des postes TV, il faut restreindre les analyses aux chaînes thématiques. C'est-à-dire aux chaînes qui sont disponibles quasiment exclusivement via un décodeur numérique. Et donc pour lesquelles on n'a très peu, voire pas du tout d'audience possible sur les postes non mesurables par voie de retour.

La deuxième étape consiste à transformer la donnée poste en donnée individu. L'idéal est d'arriver à individualiser la donnée et pas seulement de fournir un profil moyen. Si on a un échantillon de décodeurs avec des informations sur l'abonné (nombre de personnes au foyer, sexe et âge de chacun des membres), on peut proposer une probabilisation des sessions d'écoute pour chaque individu. Si au contraire, aucune information sur l'abonné n'est disponible, des modèles full comportementaux doivent être envisagés (estimer qui regarde la TV uniquement à partir de la consommation TV). Ce qui semble a priori plus compliqué surtout pour un média familial. De plus, si l'on considère la volumétrie des logs, il semble préférable de travailler sur un échantillon de décodeurs qualifiés plutôt que sur l'exhaustivité des logs sans information sociodémographique sur l'abonné.

Des travaux de R&D sont prévus en 2013 sur ces sujets.

5. Bilan et perspectives

Les méthodes d'hybridation consistent à combiner deux sources données de nature et granularité différentes pour en créer une troisième, plus riche. Au sein des études médias, l'hybridation se traduit aujourd'hui le plus souvent comme la combinaison de données d'enquêtes et de données issues de la voie de retour.

Ces projets d'hybridation doivent faire face à des difficultés diverses. Il faut respecter la confidentialité des informations personnelles (la participation d'un tiers de confiance peut s'avérer obligatoire), garantir la cohérence des données voies de retour, rapprocher les périmètres des deux mesures (mesure foyer vs mesure individu, mesure tous terminaux vs mesure ordinateur, etc...) et veiller à la dispersion des poids de redressement lorsque les résultats voie de retour sont utilisés comme contraintes dans le redressement.

Aujourd'hui, plusieurs méthodologies hybrides ont vu le jour. Notamment la mesure Internet Mobile qui s'appuie conjointement sur un panel et les logs fournis par les opérateurs avec une approche "log-up" et la mesure Interne Fixe qui a basculé en mesure hybride en octobre 2012 avec une approche "panel-up".

Bibliographie

[1] Ardilly, P. (2006). Les techniques de sondage, *Technip*.
[2] Beaumont, J.-F. and Bocci, C. (2008), Another look at ridge calibration, *Metron-International Journal of Statistics*, LXVI, 5–20.
[3] Médiamétrie et Médiamétrie//NetRatings, (2010). Les mesures hybrides – Synergies et rapprochement entre les mesures de l'Internet, *Le Livre Blanc*.
[4] Sautory, O. (1993), La macro calmar : redressement d'un échantillon par calage sur marges. *Documents INSEE*.

[5] Vanheuverzwyn, A. et Vouge, E. (2011). Méthode d'hybridation de données appliquée à la mesure d'audience de l'Internet mobile en France, *article présenté aux 43èmes Journées de Statistique de la SFdS, Tunis, Tunisie*.