FUSION DES ETUDES DE REFERENCE DE LA MESURE D'AUDIENCE DES MEDIAS RADIO, TV, INTERNET ET PRESSE

Aurélie VANHEUVERZWYN

Médiamétrie, Direction Analyses et Méthodes Scientifiques

Introduction

L'évolution des modes de consommation média et du comportement du public et la fragmentation des audiences ont placé l'évaluation du potentiel de couverture de la marque média, tous supports confondus, au cœur des stratégies médias : les annonceurs et leurs conseils ont besoin de nouveaux outils d'arbitrage dans la sélection des moyens de communication.

Dans ce contexte, il a été demandé à Médiamétrie de mettre au point, dans un cadre économique réaliste, une étude des comportements pluri-média. Une première réponse a été apportée par l'étude Media In Life qui consiste à recueillir, sur une journée, l'ensemble des activités média et multimédia quart d'heure par quart d'heure. La richesse des enseignements de cette étude a vite suscité de nouveaux besoins, en particulier un besoin de profondeur temporelle.

Un nouvel objectif a alors été fixé à Médiamétrie : construire un dispositif permettant de disposer de résultats aussi fins que les enquêtes de référence mono-média. Pour des raisons budgétaires, l'approche de panel « single source » a dans un premier temps été écartée et c'est vers une approche de fusion statistique que nous nous sommes orientés.

L'objectif de ce papier est de présenter la méthodologie retenue, la démarche de validation associée et les difficultés rencontrées lors de la première vague de production. Les premiers résultats de cette étude ont été publiés en décembre 2008.

1. Méthodologie de rapprochement

1.1. Le concept de fusion

La fusion statistique est un cas particulier du traitement de la non réponse. On considère deux fichiers de données : le fichier « donneur » dans lequel sont renseignées un certain nombre de variables que l'on souhaite injecter au fichier « receveur ». Les deux fichiers comportent des variables communes, les variables relais, et des variables absentes de l'autre fichier, les variables spécifiques.

Il existe deux grands types de méthodes pour injecter les variables spécifiques au fichier receveur. La première est une approche par modèle où la valeur de chacune des variables spécifiques est estimée de manière indépendante à partir des valeurs des variables relais à partir d'un modèle estimé sur la base du fichier donneur.
La seconde consiste à détecter, pour chaque individu du fichier receveur, un ou plusieurs « sosies » au sein du fichier donneur. L'individu receveur récupère alors les valeurs observées sur son sosie pour l'ensemble des variables spécifiques.

Si la première approche donne, variable par variable, une estimation sans doute plus juste, elle ne permet pas de maintenir les corrélations observées sur le fichier donneur. C'est pour cette raison que nous avons privilégié la seconde approche et c'est celle-ci qui est détaillée par la suite.

1.2. Les grandes étapes du rapprochement

L'objectif est de déterminer les individus que se ressemblent et de les assembler dans une base unique. Pour ce faire, les grandes étapes sont les suivantes.

1.2.1. Détection des variables relais

Il s'agit dans un premier temps de lister ou construire les variables communes aux deux fichiers. Parmi ces variables, certaines ne seront pas retenues comme variables relais du fait de leur faible pouvoir explicatif sur les variables spécifiques.

Certaines variables relais pourront être considérées comme prioritaires / obligatoires : ce sont les variables de contrôle. Celles-ci seront utilisées pour constituer des strates au sein desquelles seront faites les fusions : les sosies des deux fichiers seront parfaitement ressemblants sur ces critères.

Ensuite, au sein de chaque strate, donneurs et receveurs seront appariés en fonction de leur similitude au regard des variables relais.

1.2.2. Détermination de la fonction de distance

Différentes fonctions de distance peuvent être utilisées. Le choix de l'une plutôt que l'autre dépendra notamment de la nature des variables relais. Les fonctions de distance utilisées sont les suivantes :

Distance de Manhattan : $\sum_{i=1}^{n} \left| x_i - y_i \right|$

Distance euclidienne : $\sqrt{\sum_{i=1}^{n} \left(x_i - y_i \right)^2}$

Distance de Levenshtein : distance intégrant une fonction de coût permettant de pénaliser les différences en tenant compte de la probabilité d'apparition d'une valeur.

Distance de Jaccard : distance permettant de mesurer la proximité entre deux jeux de variables binaires.

1.2.3. Sélection des sosies et imputation des variables

Dans un premier temps, on affecte à chaque receveur son donneur le plus proche. A l'issue de cette première étape, un même donneur peut être affecté plusieurs fois.

Dans une seconde étape, on affecte les donneurs non encore attribués aux receveurs qui leurs sont les plus proches. A l'issue de ces deux étapes, donneurs et receveurs peuvent apparaître plusieurs fois.

On s'attache dans une dernière étape à supprimer les couples inutiles. Pour ce faire, on sélectionne l'ensemble des couples pour lesquels donneur et receveur apparaissent plus d'une fois dans la base. Avant de supprimer un couple, on s'assure que le donneur et le receveur apparaissent tous deux dans d'autres couples dont la proximité est plus grande.

1.2.4. Validation de l'imputation

Lors de l'étape de validation, deux aspects sont contrôlés :
- la qualité de l'appariement d'une part,
- la cohérence avec les bases d'origine d'autre part.

Sur le premier aspect, on analyse la distribution de la distance entre donneur et receveur ainsi que le nombre d'apparitions des donneurs et des receveurs dans la base finale ; le choix de la fusion étant un compromis entre les deux critères.

En ce qui concerne le second point, on compare dans un premier temps les moyennes et distributions marginales. On compare ensuite les rangs et les corrélations à l'aide de tests de Wilcoxon et de Mantel.

2. Résultats de la première vague de l'étude Cross-Médias

2.1. Description des données étudiées

L'étude Cross-Médias est le fruit du rapprochement des études de mesure d'audience des différents médias autour d'une enquête pivot au sein de laquelle sont recueillies les habitudes de consommation de ces différents médias.

L'enquête pivot est issue du panel Radio de Médiamétrie. Elle est composée d'environ 5000 individus sur lesquels on dispose d'informations socio-démographiques et géographiques, des habitudes d'écoute TV par chaîne et par émission, de la fréquence de connexion Internet au global et sur quelques sites et des habitudes de lecture de la presse.

Les études de référence média par média sont :
- le panel Médiamat pour la TV
- le panel Médiamétrie//NetRatings pour Internet
- l'enquête AEPM pour la presse magazine
- l'enquête EPIQ pour la presse quotidienne
- le panel Radio pour la Radio

2.2. Difficultés rencontrées

La première difficulté a été de rapprocher des études différentes non seulement au niveau du mode de recueil (auto-administré, automatique, téléphone ou face-à-face) mais également au niveau de la nature de l'information recueillie. En effet, dans l'enquête pivot sont recueillies des habitudes de

consommation des médias alors que les enquêtes sources mesurent l'audience effective. Afin de pouvoir utiliser cette information dans le rapprochement, nous avons reconstruit des habitudes dans les enquêtes sources à partir de l'audience effective, malgré l'approximation que cela représente. Cette approximation est d'autant plus grande que, pour la télévision et Internet, la mesure d'audience est automatique alors que dans l'enquête pivot, le recueil est déclaratif. Pour la presse, le recueil est déclaratif dans les deux cas, mais on a d'un côté une enquête dédiée au média dont le recueil est fait en face-à-face ou par téléphone et de l'autre une enquête plurimédia autoadministrée. Par conséquent, les variables considérées comme communes pour la fusion ne sont pas strictement identiques, ce qui a sans doute compliqué l'opération.

De plus, les plans de sondage et tailles d'échantillon diffèrent d'une enquête à l'autre, ce qui implique un rapport nombre de donneurs sur nombre de receveurs différent d'une fusion à l'autre, mais aussi, pour une même fusion, d'une strate à l'autre. D'où la complexité de la validation de l'appariement.

Par ailleurs, chaque fusion est réalisée de manière indépendante Par conséquent, à l'issue des fusions, un individu de l'enquête pivot s'est vu affecter par exemple 5 sosies pour la TV, 4 pour Internet, 1 pour la presse quotidienne et 2 pour la presse magazine. On crée donc autant d'individus virtuels que de combinaisons possibles. Ceci a pour effet de démultiplier les enregistrements dans la base finale et on ne connait pas l'effet de ce dépliage sur les corrélations.

Enfin, s'il est possible de valider les résultats des fusions une à une, la pertinence des duplications obtenues entre supports de différents médias est quant à elle plus délicate à évaluer car on ne dispose d'aucune source théorique.

2.3. Bilan

Malgré ces difficultés, les premiers retours des clients après deux mois d'exploitation ont été très positifs. Le livrable a été jugé en adéquation à leurs attentes et les résultats cohérents avec ceux d'autres études. Pour la seconde vague, des ajustements ont été apportés au dispositif de manière à limiter les difficultés : la taille de l'échantillon de l'enquête pivot a été doublée et des questions ont été ajoutées dans les différents questionnaires de manière à accroître sensiblement la batterie de variables de pont potentielles. La comparaison des résultats de ces deux vagues nous permettra par ailleurs d'apprécier la robustesse du dispositif et d'en estimer l'erreur inhérente.

Bibliographie

[1] Ardilly P., *Les techniques de sondage*, Technip, Paris, 1994.
[2] Derquenne C., « La combinaison de données de sources différentes : appariement statistique vs modèles de prédiction » dans Lavallée P., Rivest L.-P., *Méthodes d'enquêtes et sondages*, Dunod, Paris, 2006.
[3] Lagarenne C., Lorgnet J.-P., « Fusion de fichiers, données manquantes : application aux revenus du patrimoine de rapport des ménages français » dans Droesbeke J.-J., Lebart L., *Enquêtes, modèles et applications*, Dunod, Paris, 2001.
[4] Lejeune M., *Traitements des fichiers d'enquêtes*, Presses Universitaires de Grenoble, 2001.
[5] Santini G., *Mathematical models & methods for media research*, G.S. IT Services, 2003.
[6] Tillé Y., *Théorie des sondages*, Dunod, Paris, 2001.