Mesure d'audience croisée (XMM) : modéles et algorithmes

Nicolas Chopin et Guillaume Lecué en partenariat avec Médiamétrie

email:nicolas.chopin@ensae.fr guillaume.lecue@ensae.fr
CREST, ENSAE, IPParis. 5, avenue Henry Le Chatelier, 91120 Palaiseau, France.

Contexte : La WFA (world federation of advertisers) a exprimé le besoin des grands annonceurs 'mondiaux' (Danone, Coca, etc.) d'une mesure de l'éfficacité des campagnes de publicités multi-média / multi-support uniformisée au niveau mondial en 2019. Google et Facebook ont conjointement répondu à cet appel et ont proposé une méthode de mesures basée sur trois papiers de Google [2], [1] et [6]. Ces trois papiers ne s'appuient pas sur l'approche actuellement utilisée chez Médiamétrie, nommée *DAR (Digital Ad Ratings)*, qui utilise les données loguées de Facebook pour le suivi des impressions digitales. Cette dernière approche ne soufre pas des problémes de correspondances "cookies-to-users" et est donc beaucoup plus facile à mettre en oeuvre. Dans l'approche dernièrement proposée par Google et Facebook, le calcul des Reachs se base uniquement sur les données d'impressions de cookies pour en déduire les impressions sur users (les chercheurs de Google avaient anticipé dés 2013 la possibilié que les données loguées ne pourraient plus être utilisées pour la mesure d'audience sur Internet – les problèmes liés à l'utilisation des données personnelles à des fins commerciales devaient commencer à être soulevés).

Suite à la proposition de Google et Facebook, la WFA est revenue auprès des 'mesureurs' nationaux comme Médiamétrie en France, associations locales de chaîne de TV, groupement de sites internet, etc. en leur disant que la WFA aimerait appliquer cette méthode. Des sessions ont alors été organisées par la WFA pour expliquer le modéle de Google/Facebook. Ces sessions ont été animées par des chercheurs de Google.

Du côté Médiamétrie, la demande de la WFA – et des 'grands annonceurs' derrières elle – est tout à fait légitime et est un objectif sur lequel tous les acteurs (éditeur, annonceurs et mesureurs) peuvent travailler conjointement. Cependant l'algorithme proposé par Google peut être challengé et c'est le but de ce stage d'introduire de nouveaux algorithmes pour la mesure d'audience croisée à la fois sur le digital (PC, Tablette et téléphone portable) et la TV.

But du stage est de construire des algorithmes (Gibbs sampler, méthodes statistiques basées sur des modèles et méthodes d'apprentissage basées sur la complétion de matrices) pour répondre au problème de la mesure croisée d'audience digitale+TV pour des populations de la taille du million ou de la dizaine de millions d'utilisateurs. Une fois ces algorithmes construits, on évaluera leurs performances en termes d'estimation du reach et de robustesse. On comparea en particulier leurs performances à celle de l'algorithme de Google.

Le point de départ du stage sera de lire plusieurs documents en lien avec le XMM comme [3], [5] [4] et les documents (rapport et notebooks) réalisés par Daniel Amar et Mehdi Ghrabli durant leurs stage de 2021. Le problème de mesure croisée d'audience digitale + TV est un problème assez récent et donc toute nouvelle idée et initiative personnelle sont les bienvenues.

Rémunération : 700 euros brut par mois,

Lieu du stage : ENSAE (présentiel ou distanciel) + quelques échanges (présentiels ou distanciels) avec le datalab de Médiamétrie.

1 Les données

Il existe plusieurs types de données plus au moins agrégées (individus, foyers, somme total sur le panel, somme totale sur internet) provenant de données de panel (Panel TV+digital 'single source' ou agrégée Panel (Tab, PC, Tel)), des données internet (où on observe la somme totale de cookies de chaque type) et des données de box internet pour lesquelles on observe les données de visionnage publicitaire au niveau du foyer. Le tableau ci-dessous résume les données dont on dispose ou disposera bientôt.

c

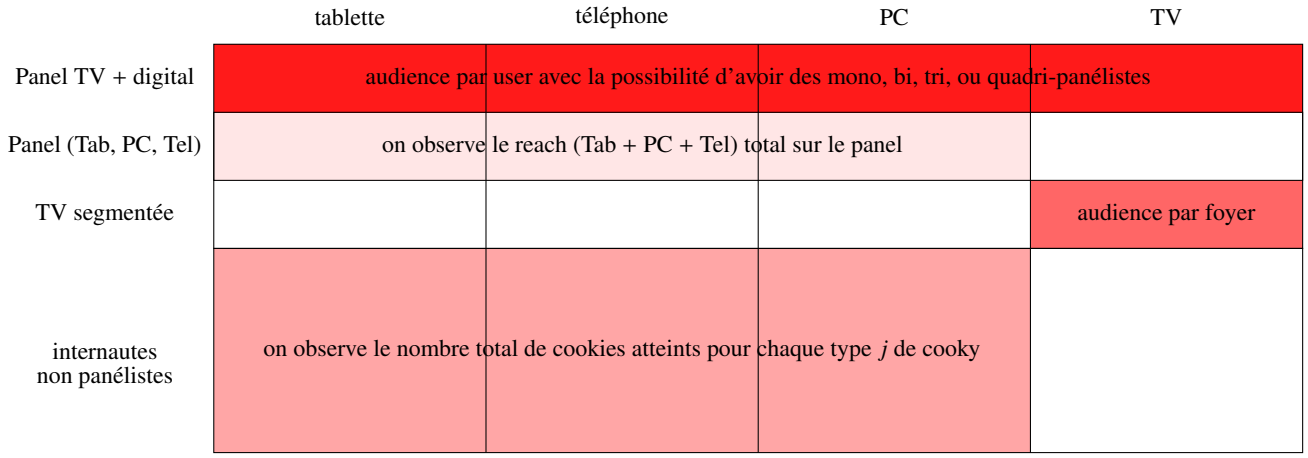

Figure 1: Données multiformes pour la mesure d'audience croisée digitale + TV.

2 Mesure d'audience publicitaire : jargon

On donne ici le jargon utilisé en mesure d'audience publicitaire.

- Reach = couverture : Nombre de foyers ou personnes d'une strate exposés ou moins une fois à une publicité lors d'une campagne publicitaire et sur une période de temps spécifique. La couverture exclue en particulier les duplications. On rapporte souvent le reach par strate; par exemple, le reach des *H 15-24 CSP+* de cette campagne a été de 24% : ce qui signifie que 24% de la population des *H 15-24 CSP+* a été exposé à une annonce publicitaire de cette campagne de pubs au moins une fois. On donne en général le reach total d'une strate socio-démographique. On peut aussi s'interesser à des *slice* de cette strate en ne regardant que les reach des peronnes qui aime le sport par exemple. On peut aussi s'interesser à la distribution du reach : combien de personnes ont été atteintes 1, 2, 3 etc. fois par la campagne (c'est une sortie intéressante car on sait qu'une campagne de pubs peut avoir un effet néfaste sur une marque quand des utilisateurs ont été trop exposés à leurs pubs).

- Frequency = répétition : Nombre moyen d'expositions à une publicité reçues par les individus d'une strate de population ayant été atteint au moins une fois lors d'une campagne. Par exemple, on peut dire que la frequency des *H 15-24 CSP+* a été de 2.2 pour cette campagne; cela signifie qu'en moyenne chaque individu de la strate *H 15-24 CSP+* qui a reçu une pub de cette campagne l'a reçu 2.2 fois. Cela tient en particulier compte de la duplication (qui est un des problèmes lié au manque de lien entre cookies et users) qui est pour 2 sites web par exemple, la proportion d'individus ayant visité les deux sites.

- Impression: On parle d'une impression quand une publicité est apparue dans un encart publicitaire. Le nombre d'impressions pour la strate *H 15-24 CSP+* est donc le reach × frequency (c'est ce qui est aussi appelé le GRP de la strate *H 15-24 CSP+* ou son TRP).

- ad impressions: Nombre de fois où une publicité est affichée.

- R&F: Reach and Frequency; c'est ce qu'on cherche à estimer lors d'une campagne de pubs : pour toutes les strates (comme dans la Figure 2) on souhaite donner le Reach et le Frequency associés à cette strate pour une campagne. En effet, à la suite du calage sur marge (socio+nombre de visite), on effectue une **stratification** des bases de données en fonction des cibles d'audience comme par exemple 'Homme 15/24 CSP+' (voir une liste de 26 strates issues du projet One Next en Figure 2).

strate	Sexe	Age_5	CSP_3
1	Homme	15-24 ans	CSP+
2	Homme	15-24 ans	CSP-
3	Homme	15-24 ans	Inactifs
4	Homme	25-34 ans	CSP+
5	Homme	25-34 ans	CSP-
6	Homme	25-34 ans	Inactifs
7	Homme	35-49 ans	CSP+
8	Homme	35-49 ans	CSPI-Inactifs
9	Homme	50-64 ans	CSP+
10	Homme	50-64 ans	CSP-
11	Homme	50-64 ans	Inactifs
12	Homme	65 ans et +	CSP+
13	Homme	65 ans et +	CSPI-Inactifs
14	Femme	15-24 ans	CSP+
15	Femme	15-24 ans	CSP-
16	Femme	15-24 ans	Inactifs
17	Femme	25-34 ans	CSP+
18	Femme	25-34 ans	CSP-
19	Femme	25-34 ans	Inactifs
20	Femme	35-49 ans	CSP+
21	Femme	35-49 ans	CSPI-Inactifs
22	Femme	50-64 ans	CSP+
23	Femme	50-64 ans	CSP-
24	Femme	50-64 ans	Inactifs
25	Femme	65 ans et +	CSP+
26	Femme	65 ans et +	CSPI-Inactifs

Figure 2: Stratification des bases de données. Ce sont les bases de données stratifiées qui sont fusionnées

- GRPs est le Gross Rating Points = reach × frequency.
- TRPs est le Target Rating Points = c'est le GRPs d'une strate/d'un groupe.
- Publisher = éditeur : sur internet, c'est un site comme le http://lemonde.fr. D'un point de vue publicitaire c'est juste un endroit où on peut mettre des publicité et potentiellement connaître des informations sur le user qui l'a vue. Un *annonceur* est une marque qui achète les espaces publicitaires comme Renaut.
- PPD = Publisher Provided data: Ce sont les données sur les utilisateurs recueillies par les éditeurs. Elles sont en générale recueillies de manière automatique par les éditeurs sur leurs sites (souvent par des solutions de taggage ou par des données logguées donc enrichies sur le user, c'est le cas de MyTF1 où on doit se créer un compte pour pouvoir regarder des vidéos).
- WFA: World Federation of Advertisers
- Media planning: Mélange optimal d'annonces publicitaires sur différent media (pour optimiser le R&F d'une campagne).
- Panel attrition: nom donné au phénomène qui consiste à voir un panel ressembler de moins en moins à une population. L'attrition est le phénomène de "perte" qu'on observe sur un panel (avec les panélistes qui quittent la mesure). L'attrition peut "déformer" la structure du panel par rapport à la population de référence (par exemple si les jeunes participent moins longtemps à la mesure que les plus âgés). Pour compenser l'attrition, on doit recruter de nouveaux panélistes régulièrement.
- CESP : Centre d'étude des Supports de Publicité. Le CESP est l'organisme interprofessionnel des acteurs de la communication concernés par l'étude de l'audience des médias et la mesure de leur efficacité : annonceurs, agences et médias. Le CESP audite toutes les mesures d'audience de référence pour le compte de ses adhérents, quel que soit le média : Internet, TV, presse, communication extérieure, radio, cinéma.
- Census data : Habituellement ce sont les données de recensement. Mais ici c'est aussi les données site-centric.

L'estimation de l'audience d'une campagne de publicités digitale (généralement multi-devices et multi-publishers) consiste à renvoyer un tableau par strate comme dans Table 1 :
où D est le nombre de strates, $U = U_1 \sqcup \cdots \sqcup U_D$ est la population totale stratifiée (ici la strate $d = 1$ est notée U_1 est l'ensemble de tous les individus *H 15-24 CSP+* de U) et $Imp \subset U$ est l'ensemble des individus de la population totale U ayant été impressionés lors de cette campagne. Pour tout individu $k \in U$ on note par $n_k \in \mathbb{N}$ le nombre de fois où l'individu k a été impressionné. En particulier, on a $Imp = \{k \in U : n_k > 0\}$. On peut aussi donner de manière aggrégée le Reach : comme le

strate	Reach (couverture)	Frequency (répétition)	GRP
(d=1) H 15-24 CSP+	$R_1 = \sum_{k \in U_1} I(k \in Imp)$	$f_1 = \frac{1}{R_1} \sum_{k \in U_1} n_k$	$R_1 \times f_1$
$\cdots$	$\cdots$	$\cdots$	$\cdots$
(d=D) F $\geq$ 65 CSP—-inactifs	R_D	f_D	$R_D \times f_D$

Table 1: Rendu de la mesure d'audience d'une campagne de publicités.

nombre d'impressions uniques (= le reach) des hommes. Il peut aussi être rendu une répartition des $(n_k)_{k \in U_1}$: sur les personnes de la strate 1 qui ont été atteintes, 60% l'ont été une fois, 30% l'ont été deux fois et 10% l'ont été au moins trois fois.

Etant donnée une campagne digitale, on cherche à estimer les D couples $(R_1, f_1), \cdots, (R_D, f_D)$. Les difficultés sont :

i) de faire le lien entre "cookies" et "unique users" (car les cookies ont une durée de vie limitée – en fonction du device –, les users changent de devices, il peut y avoir plusieurs users derrière un même cooky). Ces difficultées liées aux cookies ont tendance à surestimer la couverture (reach) et sous-estimer la frequency.

ii) d'estimer la strate démographique d'un cooky car les données des PPD ne sont pas fiables (on a alors recours à un modèle correctif socio-démo qui s'appuie sur les données du panel. On ne traitera cependant pas de ce problème durant le stage et on se placera à l'interieur d'une strate socio-démographique pour laquelle on estimera le reach.

iii) panel attrition: le panel devient de moins en moins représentatif de la population avec le temps; solutions : campagnes de recrutements de nouveaux panélistes et affectation de poids par calage sur marge sur données de recensement.

References

[1] Jim Koehler, Evgeny Skvortsov, Sheng Ma, and Song Liu. Measuring cross-device online audiences. Technical report, Google, Inc., 2016 (To appear).

[2] Jim Koehler, Evgeny Skvortsov, and Wiesner Vos. A method for measuring online audiences. Technical report, Google Inc, 2013 (To appear).

[3] Guillaume Lecué. Wfa and google papers. 2020.

[4] Guillaume Lecué. The xmm problem : models and algorithms. Technical report, CREST - ENSAE, 2020.

[5] Guillaume Lecué and Valentin Patilea. On the virtual id model for cross-media multi-providers R and F measurements. Technical report, 2020.

[6] Evgeny Skvortsov and Jim Koehler. Virtual people: Actionable reach modeling. Technical report, 2019.