Panorama des dispositifs hybrides de mesure d'audience des médias en France

Aurélie Vanheuverzwyn et Lila Zydorczak

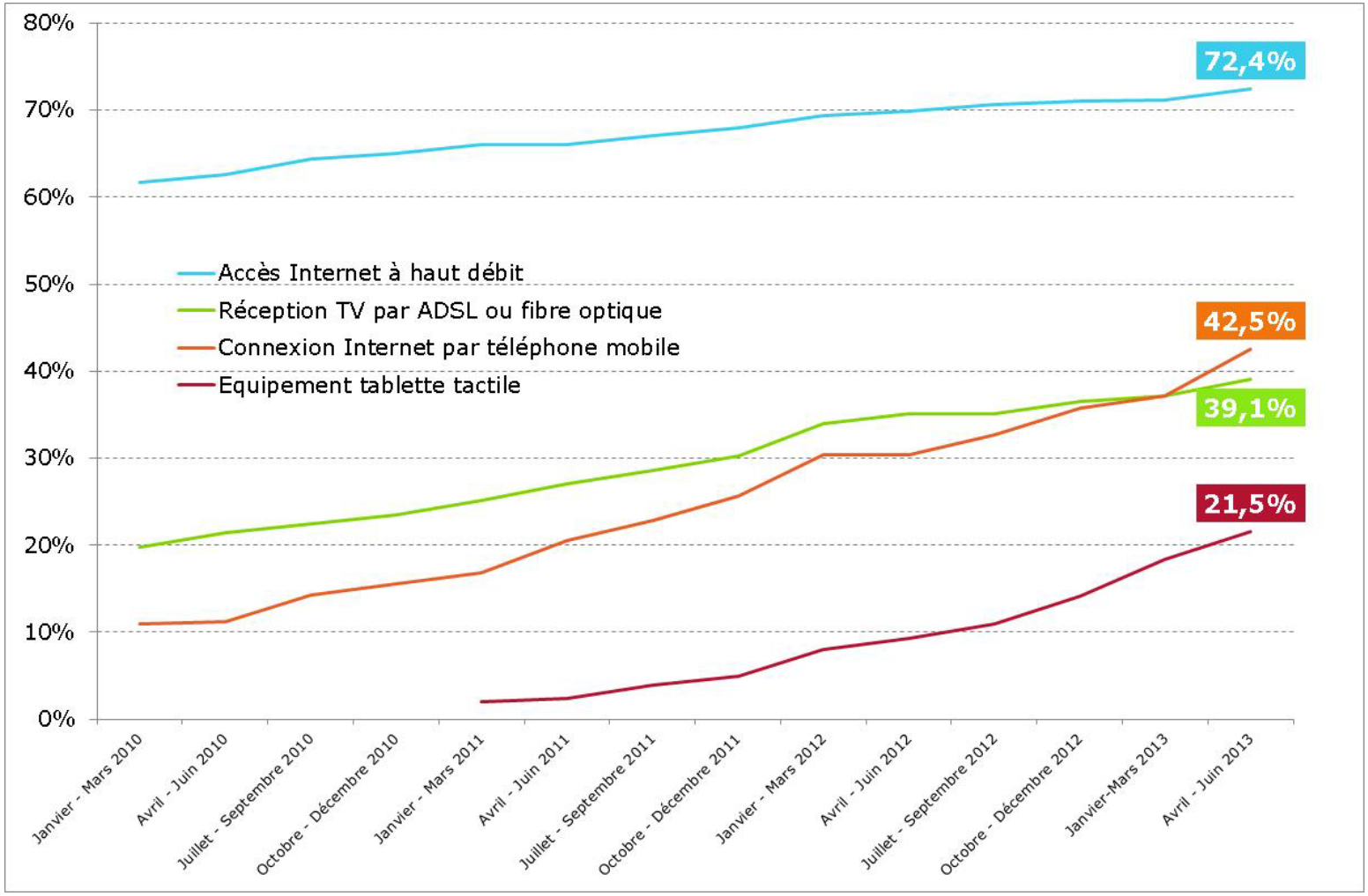
Mediametrie

1

Contexte et problématique

L'avènement du tout numérique

Evolution des équipements des foyers en France métropolitaine depuis 2010



Source : Médiamétrie - Référence des Equipements Multimédias

Tout numérique : une aubaine pour les études médias

Qui dit tout numérique dit voie de retour

Multiplication de bases de données sur l'usage des décodeurs numériques TV, des téléphones mobiles ou de fréquentation des sites Internet

Récupération des logs de connexion pour :
- les décodeurs numériques et la TV par ADSL
- la 3G et l'Internet Mobile
- le surf internet et les cookies

Donc, possibilité d'avoir des données exhaustives…

… mais souvent de granularité différente ou sur un périmètre partiel

Des données « machine » vs « individu »

Usage individuel ou collectif

Couverture de la mesure :
- écrans connectés pour la TV vs tous les écrans, quelque soit le mode de réception
- connexions à Internet quelque soit le lieu, le pays vs panel à domicile en France métropolitaine
- etc…

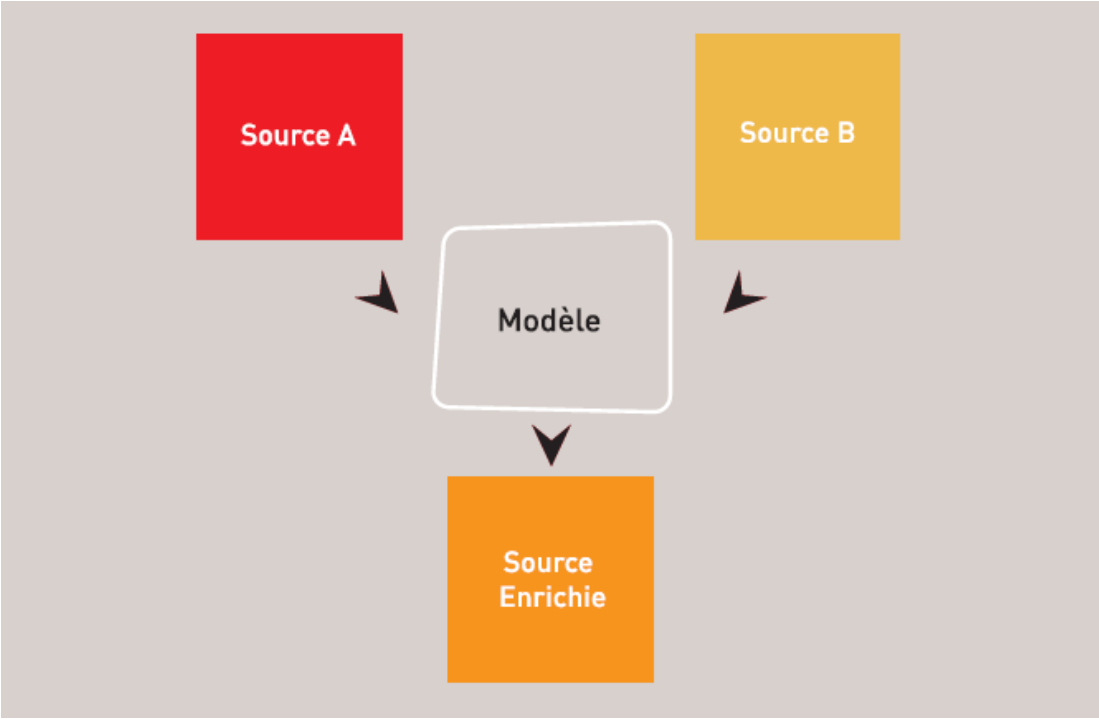
Qu’est-ce qu’une mesure hybride ?

Evolution naturelle des dispositifs

Intégrer les nouvelles sources d’information pour faire évoluer l’existant

Mixer "sondages" et "exhaustivité"

Mélanger deux sources d’information de natures et de niveaux différents, les croiser mutuellement pour en créer une troisième, plus fine ou plus riche



Les approches méthodologiques

L’approche Panel-up

La mesure exhaustive (logs issus de la voie de retour) vient enrichir l’information issue de l’enquête média (généralement un panel).

La mesure voie de retour constitue une source d’informations auxiliaires intégrées dans le redressement de l’enquête.

Cette première approche permet d’augmenter la qualité de la mesure panel (correction des biais de recrutement, diminution de la variance statistique).

En revanche, elle ne permet pas de gagner en granularité.

L’approche Log-up

L’enquête média vient qualifier, à l’aide d’un modèle, la donnée exhaustive.

Cette deuxième approche permet de fournir des informations de profil par exemple y compris pour des consommations très faibles (sites, chaînes à faibles audiences).

A noter que les informations issues de cette approche sont entachées des erreurs statistiques inhérentes à toute modélisation.

⇨ Le choix de la méthode dépend de la problématique à laquelle on souhaite répondre

La mesure d'audience de la télévision en France

Population de référence

Foyers équipés d'au moins un téléviseur actif en France Métropolitaine

Foyer : ménage ordinaire au sens de l'Insee (i.e. hors ménages collectifs, résidences universitaires, foyers de travailleurs, maisons de détention)

Téléviseur actif : utilisé au moins une fois par mois pour regarder la télévision

France Métropolitaine : Corse comprise depuis 2009

Au sein de ces foyers, l'ensemble des individus de 4 ans et plus participe au panel.

Champ de la mesure

Consommation de la TV sur un poste de télévision au domicile (résidence principale)

L'élargissement de la mesure aux autres écrans du foyer ou à la résidence secondaire sont des pistes régulièrement évaluées.

Echantillon

5000 foyers, soit environ 11700 individus de 4 ans et plus

Echantillon construit selon la méthode des quotas :

Région, âge et CSP du chef de ménage, activité de la ménagère, nombre de téléviseurs actifs, modes de réception et abonnements TV

Surreprésentation de certaines régions pour les décrochages régionaux de France 3

Sondage par grappe :

On recrute des foyers.

L'audience de l'ensemble des individus de 4 ans et plus du foyer est mesurée.

La mesure d'audience de la télévision en France

Le mode de recueil

Mesure audimétrique (semi-passive)

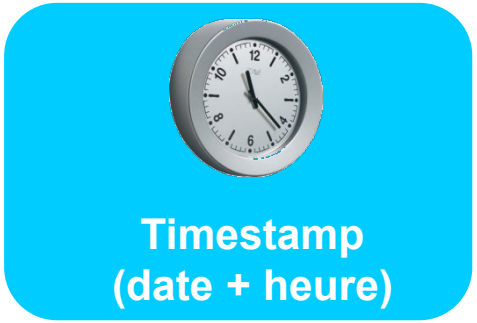
Collecte en continu via des audimètres reliés aux téléviseurs actifs

- Détection automatique de l'allumage du téléviseur et des changements de chaînes
- Les individus déclarent leur présence devant la télévision à l'aide d'une télécommande

Le watermarking audio

Tatouage numérique inséré dans le signal audio et inaudible qui permet de connaître :

- la chaîne regardée
- la date et l'heure de diffusion afin d'identifier si le contenu est consommé en direct ou en différé

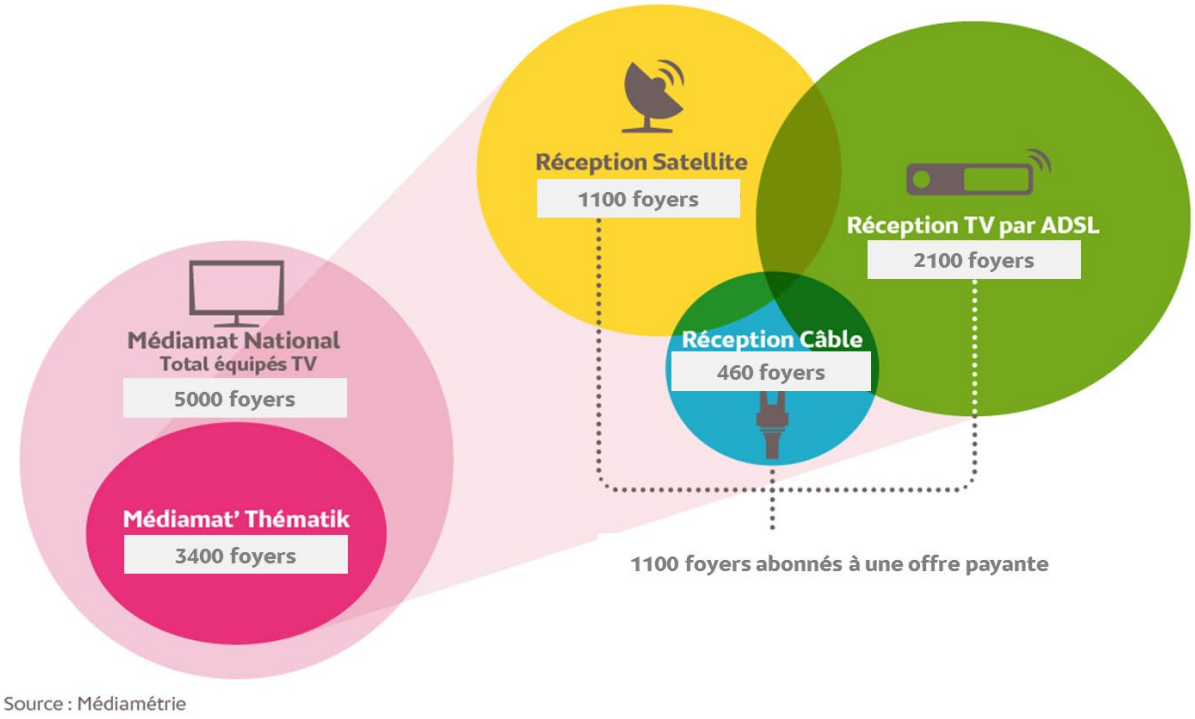


A ce jour, 180 chaînes sont marquées en France.

La mesure d'audience de la télévision en France

La mesure des chaînes thématiques

La réception des chaînes thématiques dépend du type d'offre télévisuelle possédé par le foyer. Les bases de calcul des audiences sont donc plus faibles et les résultats publiés à une fréquence moindre.



Source : Médiamétrie

Besoin exprimé par le marché

Avoir des résultats d'audience individuelle plus fréquents et plus précis grâce à l'utilisation des données voie de retour des opérateurs satellite, ADSL et câble

La voie de retour TV

Sources

Disponible pour les décodeurs numériques ADSL, Câble et Satellite connectés à Internet

Nature de l'information

Logs permettant de savoir l'heure à laquelle la consommation d'une chaîne ou d'un service a débuté

Couverture de la voie de retour TV

55% des foyers équipés TV sont équipés d'un décodeur numérique

36% des postes TV sont reliés à un décodeur numérique

Source : Médiamétrie - Référence des Equipements Multimédias – 2ème trimestre 2013

Il s'agit d'un maximum car tous ces décodeurs ne sont pas nécessairement connectés à Internet.

Les points de divergence avec la mesure d'audience

Différence entre le téléviseur et le décodeur

Le décodeur peut être allumé alors que le téléviseur est éteint

Un poste relié à un décodeur numérique peut être équipé d'autres modes de réception, comme la TNT par exemple

Différence entre le téléviseur et l'individu

En moyenne, plus de 40% du temps passé par les individus de 4 ans et plus devant la télévision l'est à plusieurs (Source : Médiamétrie - Médiamat – Septembre 2012)

Les grandes étapes de la mesure hybride

Principe

L'idée est de « transformer » la donnée d'usage en donnée d'audience individuelle grâce à une modélisation en 2 étapes.

Etape 1 : passage du décodeur au téléviseur

Ecrêtage des usages atypiques qui semblent correspondre au cas où le décodeur est allumé alors que le téléviseur est éteint.

Etape 2 : individualisation des audiences « poste » obtenues à l'étape 1

Cette seconde étape a pour objectif d'estimer les individus présents devant l'écran. C'est cette étape qui présente le plus de difficultés.

En pratique

Les modélisations mises en place s'appuient sur trois sources d'information :

- un échantillon de foyers abonnés dont on connaît les caractéristiques sociodémographiques de l'ensemble de leurs membres,
- la donnée d'usage (logs de connexion) de ces mêmes foyers,
- le comportement d'audience mesuré au niveau individuel dans le panel Médiamat afin d'estimer les paramètres des modèles.

Travail en collaboration avec un opérateur :

Qualification sociodémographique de 10 000 foyers abonnés

Fourniture de 9 semaines de logs pour ces mêmes 10 000 foyers

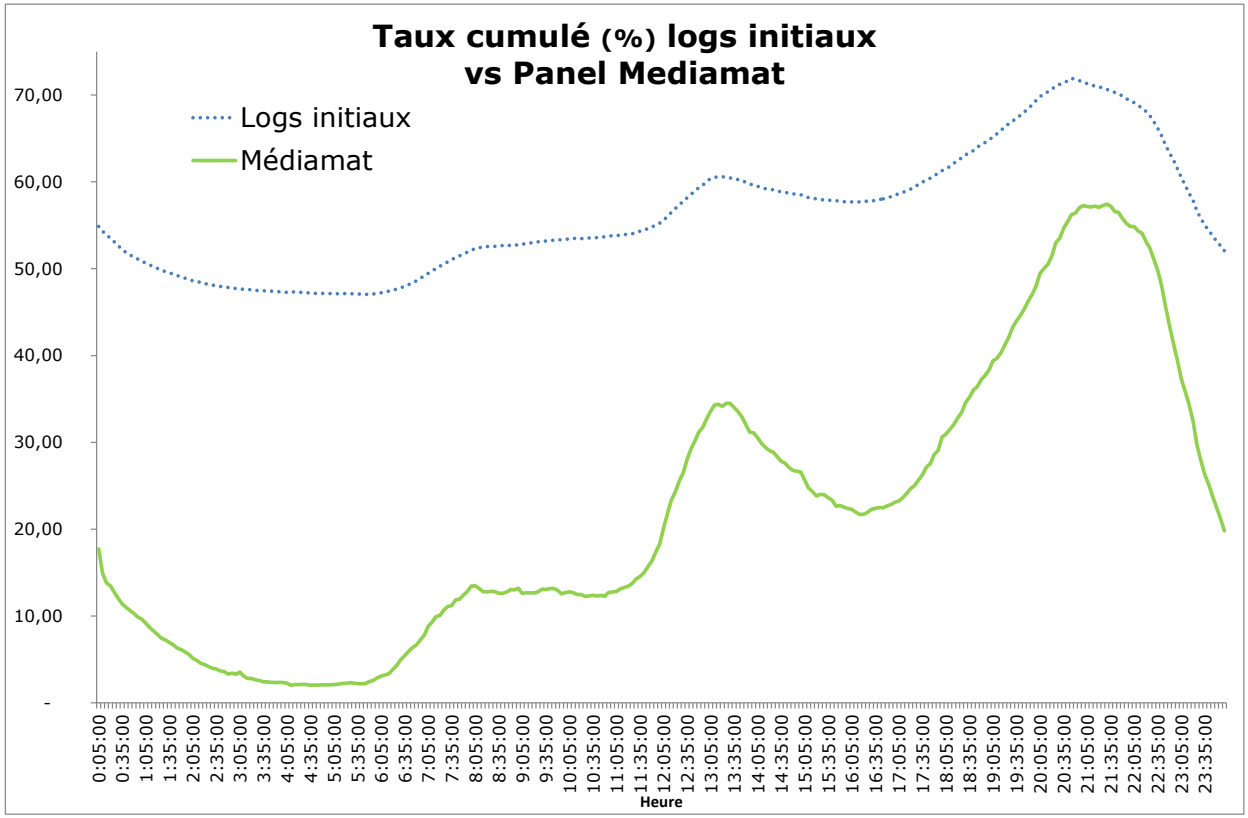
2

Des logs voie de retour aux tickets d’audience

Problématique

Objectif visé

Il s'agit de déterminer les « tickets » correspondant à des cas où le décodeur est allumé et le téléviseur éteint et de les tronquer. En sortie d'écrêtage, la courbe minute à minute doit être très proche de la courbe d'audience issue du panel Médiamat sur la base des postes reliés au décodeur de l'opérateur.



Modèles de durée - généralités

Un modèle souvent utilisé pour les durées de vie : la loi Gamma.

On considère la variable aléatoire T (durée d'un ticket).

T suit une loi $\Gamma(\theta,\nu)$:
$\theta>0$ est un paramètre de forme
$\nu>0$ un paramètre d'échelle.

Ce double paramétrage de la loi lui permet de s'ajuster assez facilement à des données variées.

La densité de T est de la forme : $f(t) = \frac{\theta^{\nu}}{\Gamma(\nu)} t^{\nu-1} e^{-\theta t}$ où $\Gamma(\nu) = \int_0^{\infty} t^{\nu-1} e^{-t} dt$

L'espérance mathématique (ou durée moyenne d'un ticket) est donnée par : $E(T) = \frac{\nu}{\theta}$

La variance est donnée par : $V(T) = \frac{\nu}{\theta^2}$

Application aux données TV

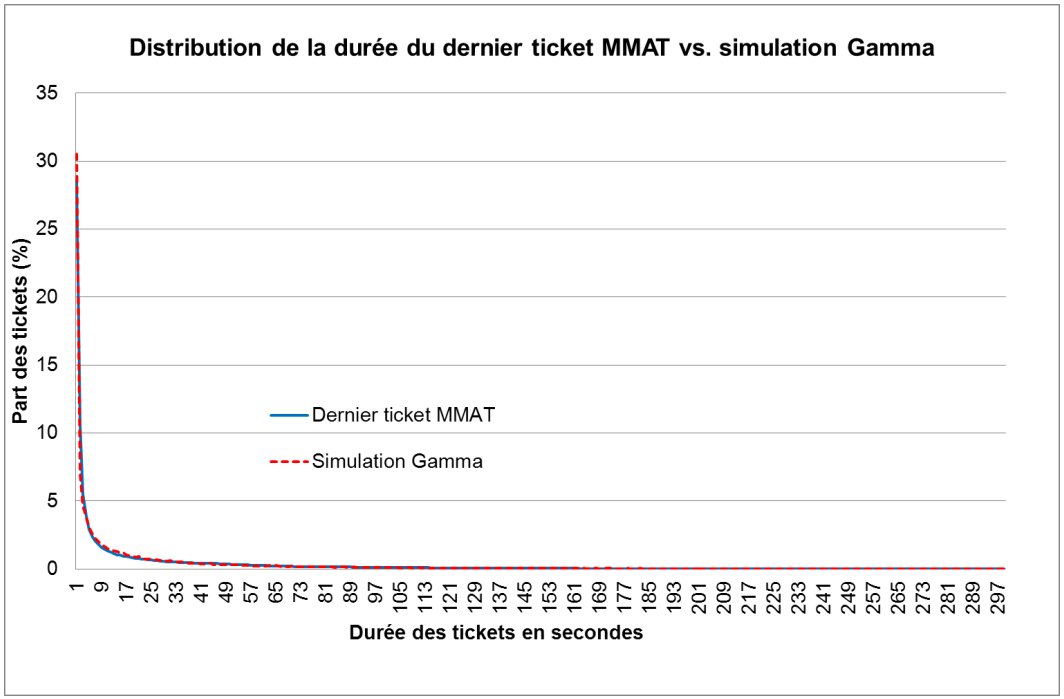
Application à la problématique d'écrêtage

On définit une session TV comme une succession de tickets TV sur le même poste. Seuls les derniers tickets d'une session sont considérés.

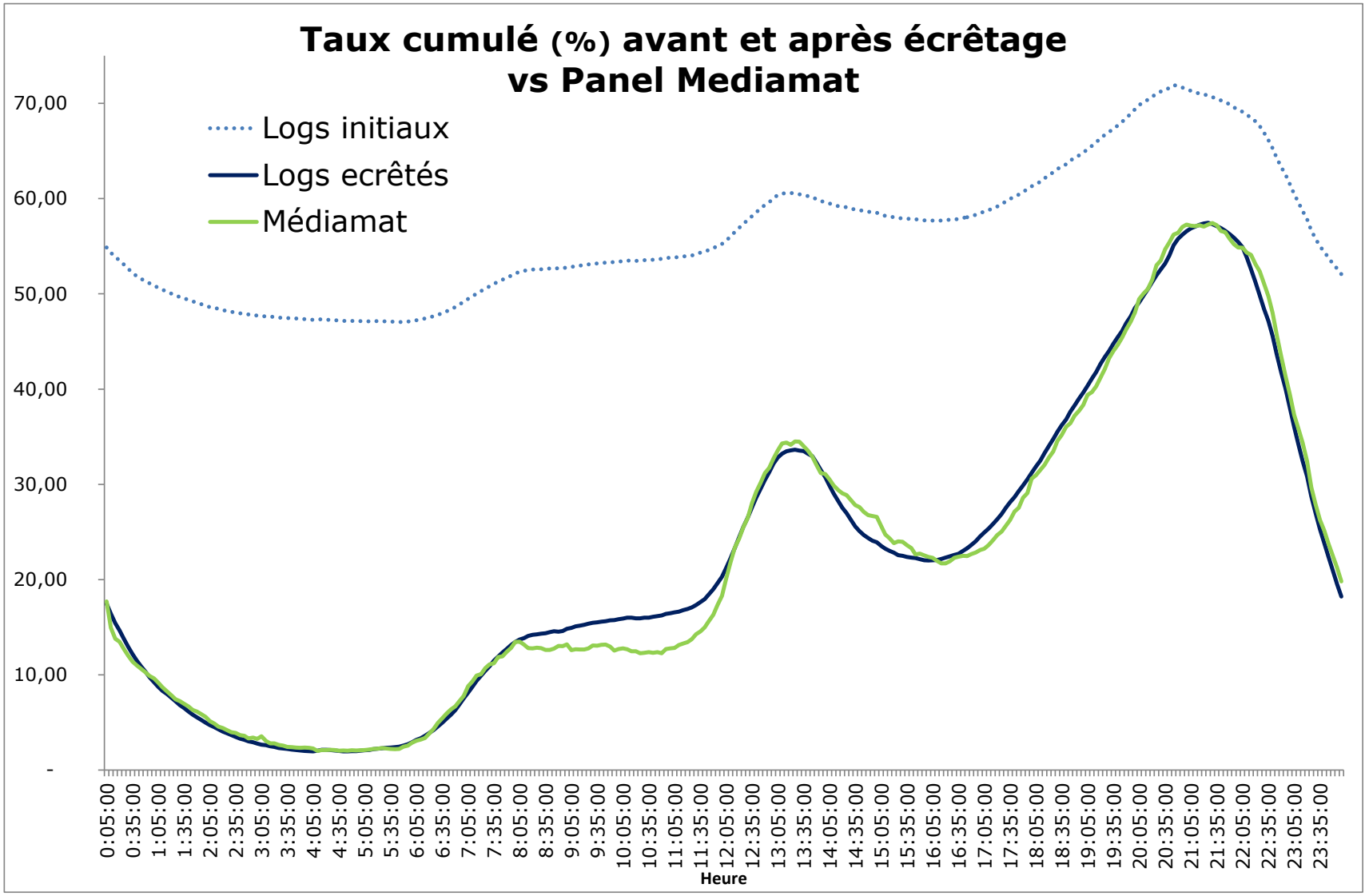
Durée des derniers tickets – Ajustement loi Γ :

Les paramètres θ et ν sont estimés sans difficulté par la méthode des moments à partir des observations faites dans Médiamat (sur la base des postes reliés au décodeur).

Le choix de la loi Gamma est validé car la loi s'ajuste parfaitement à la distribution :



Application aux données TV



3

Du niveau foyer au niveau individu

Approche retenue

Une approche comportementale et sociodémographique

Nous travaillons sur un échantillon d'abonnés pour lesquels nous disposons :
- des données d'usage (logs de connexion après écrêtage),
- des données sociodémographiques (composition du foyer avec sexe, âge, CSP et lien de parenté de chacun des individus du foyer).

Les individus du foyer étant « connus », le modèle doit déterminer à chaque instant qui regarde la TV (quand celle-ci est allumée).

Modèles de Markov cachés par typologie de foyer

Le modèle d'individualisation que nous avons retenu est un modèle de type Markov Cachés.

On construit une typologie de foyers (couple avec enfant(s), parent isolé avec enfant(s), etc…) : au sein de chaque type, un modèle est construit.

Les types de foyers rares ou non présents au sein du panel Médiamat sont regroupés avec les types les plus proches (un parent vivant avec son enfant et un autre adulte est par exemple assimilé à un couple avec enfant).

Modèles de Markov - généralités

Propriété 1 : mémoire du processus

Un processus stochastique en temps discret et horizon fini $\{X_n\}_{1\leq n\leq N}$ à valeur dans un espace d'états fini X, qui vérifie :

$$P(X_{n+1} = x_{n+1}|X_n = x_n, \dots, X_1 = x_1) = P(X_{n+1} = x_{n+1}|X_n = x_n) \quad \text{(E.1)}$$

pour tout n=1, …N-1 est appelé chaîne de Markov.

La propriété E.1 caractérise en fait la « mémoire » du processus : pour savoir où le processus va se déplacer à l'instant n+1 il n'est pas nécessaire de connaître tout le passé du processus, il suffit de connaître sa position à l'instant présent n.

On définit les probabilités de transition : $\pi_{x,x'}^{(n)} = \mathrm{P}(\mathrm{X_n} = \mathrm{x'}|\mathrm{X_{n-1}} = \mathrm{x})$,

et la loi initiale : $\mu_x = \mathrm{P}(\mathrm{X_1} = \mathrm{x})$.

Propriété 2 : chaîne de Markov homogène

Lorsque la probabilité $\pi^{(n)}$ ne dépend pas de l'indice n, la chaîne de Markov est dite homogène. Dans notre cas, nous allons considérer une chaîne de Markov homogène.

La loi d'une chaîne de Markov homogène est entièrement caractérisée par sa loi initiale μ et ses probabilités de transition π :

$$P(X_{1:n} = x_{1:n}) = \mu_{x_1}\pi_{x_1,x_2} \dots \pi_{x_{n-1},x_n} \quad \text{(E.2)}$$

Modèles de Markov cachés

Il existe de nombreux modèles d'estimation d'un état inconnu au vu d'observations partielles.

Ici, on considère le cas où l'état inconnu est modélisé par une chaîne de Markov $\{X_n\}_{1\leq n\leq N}$ à valeur dans un espace d'états fini X.

Cette chaîne n'est pas directement observée mais on dispose d'observations $\{Y_n\}_{1\leq n\leq N}$ à valeur dans un espace d'états fini Y. On suppose que ces observations sont réalisées à travers un canal sans mémoire, c'est à dire que conditionnellement aux états X_n, les observations Y_n sont mutuellement indépendantes, et que chaque observation Y_n ne dépend que de l'état X_n au même instant.

Cette propriété s'exprime de la manière suivante :

$$P(Y_{1:n} = y_{1:n} | X_{1:n} = x_{1:n}) = \prod_{l=1}^{n} P(Y_l = y_l | X_l = x_l) \qquad \text{(E.3)}$$

On définit les probabilités d'observation : $\varphi_x^y = P(Y_n = y | X_n = x)$.

Les probabilités d'observations ne dépendent pas de l'indice n.

La matrice $\varphi = (\varphi_x^y)$ vérifie les propriétés suivantes :

- pour tout x et tout y, $\varphi_x^y \geq 0$
- pour tout x, $\sum_y \varphi_x^y = 1$

Application aux données TV

Estimation des paramètres sur le panel Médiamat

La loi initiale μ décrit la probabilité d'un état du foyer.

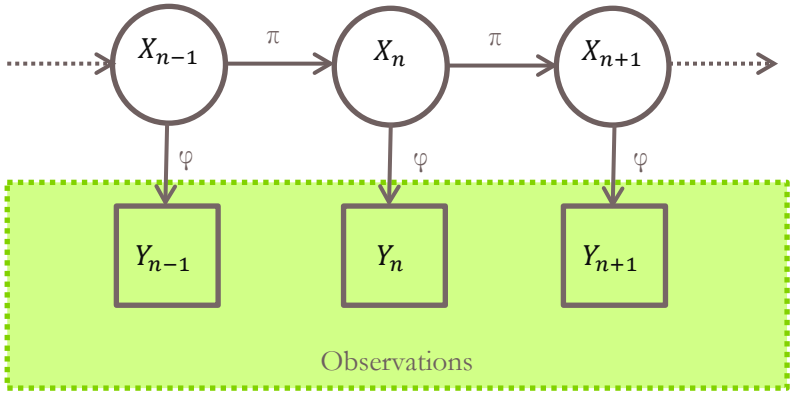
Ex : Pour un couple, on calcule la probabilité des 3 états possibles : chef seul, conjointe seule, couple.

Les probabilité de transition π correspondent aux probabilités d'avoir un état au temps t+1 sachant l'état au temps t.

Ex : Pour un couple et l'état « chef seul » au pas p, on calcule la probabilité des 3 états possibles au pas de temps p+1.

Les probabilité des observations φ correspondent aux probabilités de regarder une thématique (actualités, cinéma, sport, etc.) sachant l'état du foyer.

Ex : Pour un couple, on calcule la probabilité de regarder chacune des thématiques pour chacun des 3 états possibles.

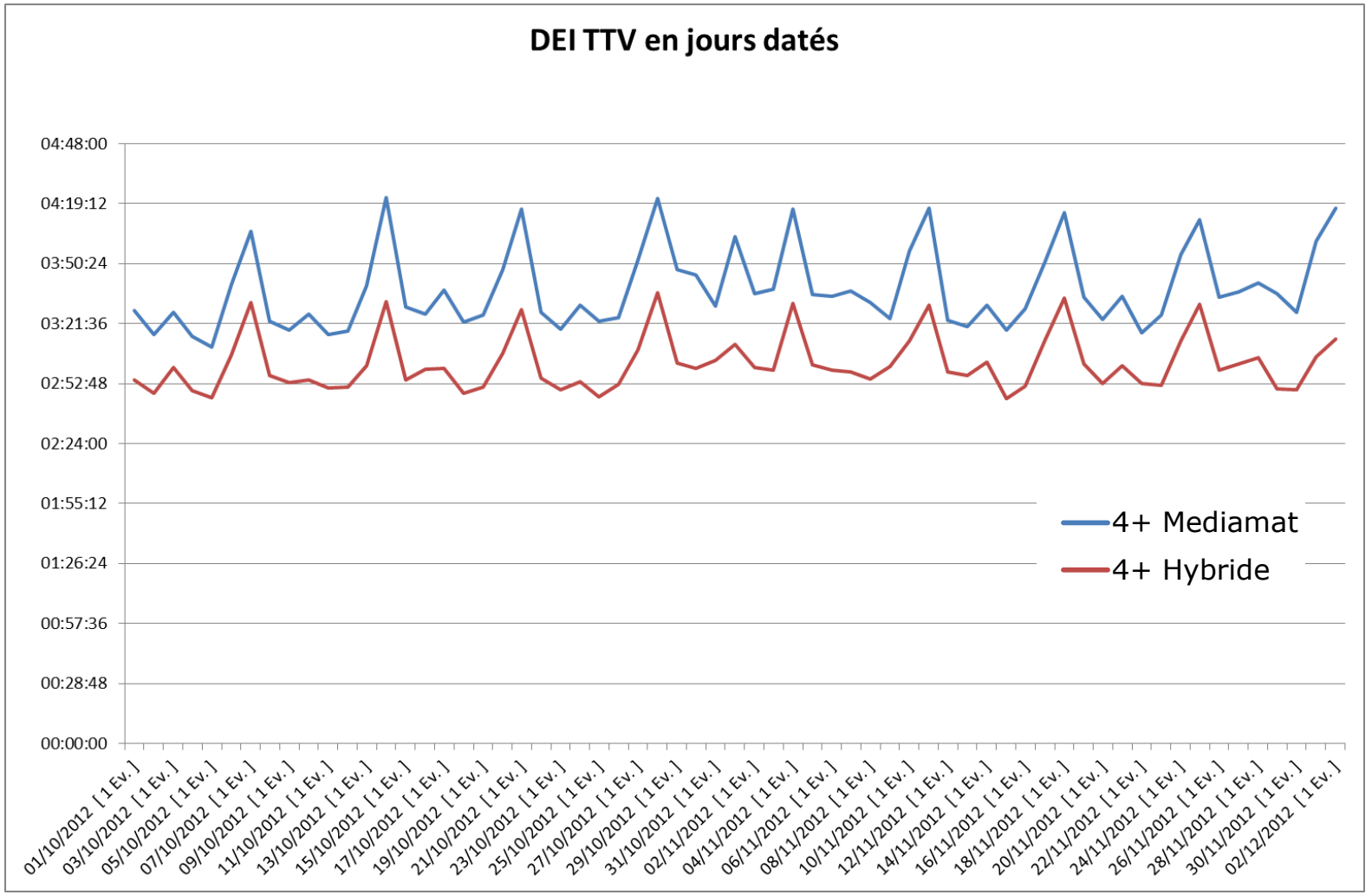


Mise en œuvre de l'individualisation

Chaque session TV est traitée indépendamment. L'algorithme de Viterbi permet de déterminer, en fonction des paramètres, du type du foyer et de l'audience du poste, les individus présents devant l'écran par pas de temps.

Application aux données TV

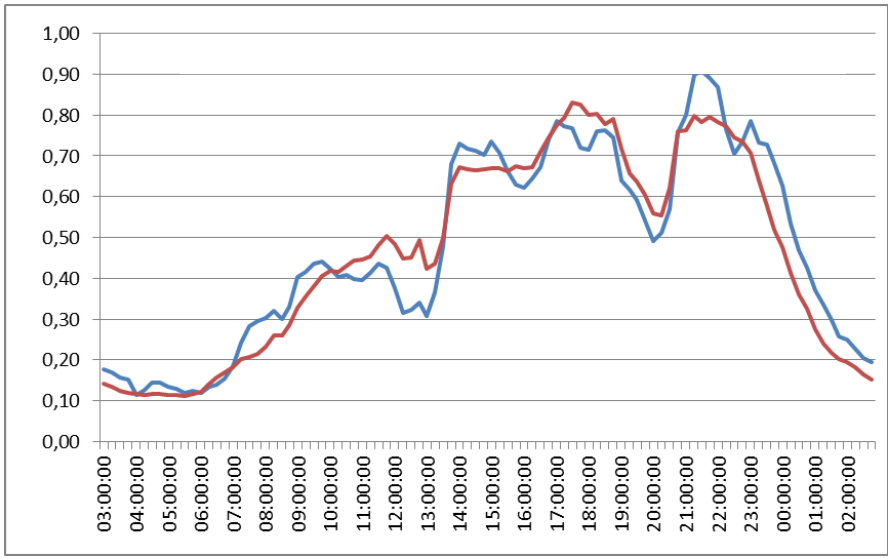
Quelques résultats



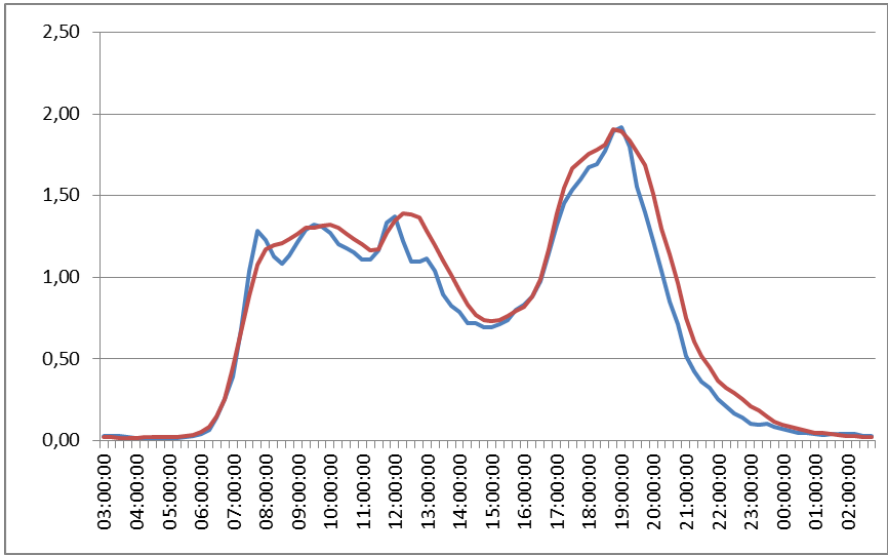
Application aux données TV

Quelques résultats

Chaines « découverte »



Chaines « jeunesse »



4+ Mediamat

4+ Hybride

4

Conclusion et perspectives

Un essai à transformer !

Un premier test concluant

La solution mise en œuvre pour répondre au besoin de résultats plus fiables et plus fréquents s'appuie sur trois sources d'information :

- les données d'usage issues de la voie de retour,

- la qualification d'un panel de foyers abonnés afin d'en connaître les caractéristiques de chacun de leurs membres,

- les données d'audience individuelle du panel Médiamat afin d'estimer les paramètres des modèles.

La mesure est donc à proprement parler hybride reposant à la fois sur de la donnée exhaustive et sur un panel de référence dont le rôle est central pour la modélisation.

Les résultats sont aujourd'hui très satisfaisants en termes de cohérence des résultats d'audience à l'issue de la modélisation.

Perspectives d'application à d'autres opérateurs

D'autres opérateurs sont aujourd'hui intéressés pour mener des tests sur leurs données d'usage.

Des tests seront ainsi menés d'ici 2014 afin de valider que la méthodologie peut être étendue à tout type d'opérateur.

Se posera alors la question d'une mesure élargie à l'ensemble des opérateurs et en particulier, l'estimation de la duplication : comment traiter le cas des foyers multi-abonnés ?

Merci de votre attention

Aurélie Vanheuverzwyn
avanheuverzwyn@mediametrie.fr

Lila Zydorczak
lzydorczak@mediametrie.fr

Mediametrie