DETECTION ET TRAITEMENT DES VALEURS EXTREMES ET INFLUENTES DANS LA MESURE D'AUDIENCE INTERNET

Magdalena Auvinet[1] & Lucie Cellier[2]

[1] *Médiamétrie, 70 rue Rivay, 92532 Levallois-Perret Cedex, mauvinet@mediametrie.fr*
[2] *Médiamétrie, 70 rue Rivay, 92532 Levallois-Perret Cedex, lcellier@mediametrie.fr*

Résumé. Médiamétrie mesure l'audience de l'internet fixe (i.e. connexion via un ordinateur). Cette mesure est estimée sur la base d'un panel de 22 000 individus ayant accès à Internet depuis leur domicile ou leur lieu de travail. Des indicateurs de référence sont publiés mensuellement : le nombre de visiteurs, de pages vues, le temps passé et le nombre de visites. La nature de ces variables d'intérêt, et en particulier leur caractère fortement asymétrique, favorise la présence de valeurs atypiques et/ou influentes. De nombreuses méthodes de détection existent, basées notamment sur le calcul de distances et de seuils ou sur le calcul de différences. Des méthodes multivariées peuvent aussi être envisagées comme les K-means. Après détection, la valeur atypique pourra être soit modifiée soit supprimée. Après redressement de l'échantillon, une observation peut devenir influente du fait d'une combinaison valeur et/ou poids de redressement élevés. Un traitement en aval est donc nécessaire. La méthode de traitement testée est la winsorization. Une méthode de détection et de traitement des valeurs atypiques a été mise en place début 2014, elle s'appuie sur la contribution des individus sur les indicateurs estimés. L'objet de cette communication sera de présenter les alternatives possibles à cette approche pour la détection et le traitement des valeurs atypiques et influentes sur Internet.

Mots-clés. Outlier, donnée atypique, donnée extrême, observation influente, distribution asymétrique, méthode de Tukey, K-means, Winsorization

1 Introduction

Les sources d'erreurs dans les enquêtes sont multiples. Certaines sont dues à l'échantillonnage : elles se produisent parce que l'information désirée n'est observée que pour une partie de la population. Les autres types d'erreurs sont les erreurs de couverture (base de sondage imparfaite), les erreurs dues à la non-réponse et les erreurs de mesure ou de traitement.

Les observations atypiques sont des observations dont les caractéristiques diffèrent de celles de la majorité des données (P. J. Rousseuw et A. M. Leroy [6] (1983)). En sondage, les données atypiques peuvent être sélectionnées ou non dans l'échantillon et leur impact diffère selon le cas. On introduit également le concept d'observation influente. Une observation peut être définie comme influente si son exclusion de la population (et de l'échantillon le cas échéant) a un effet important sur l'erreur de prédiction (J.-F. Beaumont et D. Haziza [1] (2012)). La présence possible d'unités influentes peut provenir :

- d'une distribution asymétrique de y,
- de poids de sondage élevés.

De par la nature des dispositifs d'enquête et des variables d'intérêt, la mesure d'audience est propice à la présence de valeurs influentes.

La mesure d'audience de l'internet fixe (internet via les ordinateurs) est basée sur un panel de 22 000 individus accédant à internet depuis leur domicile et/ou leur lieu de travail. L'activité de ce panel représentatif de la population internaute française âgée de 2 ans et plus est observée de manière détaillée à l'aide d'un logiciel spécifique installé chez les panélistes et chargé de mesurer la

fréquentation des sites Internet. A partir de cette base de panélistes, les résultats d'audience Internet sont publiés mensuellement, les indicateurs de référence étant l'audience, les pages vues, le temps passé et le nombre de visites.

La méthodologie utilisée pour cette mesure d'audience comprend une fusion, une modélisation et un redressement. Suite à ce processus complexe, les observations se voient attribuer des poids et certaines sont dupliquées lors des phases de fusion et modélisation. Les comportements extrêmes sont ainsi potentiellement démultipliés.

L'enjeu est de trouver une méthode qui permettent d'assurer une plus grande robustesse des indicateurs produits tout en gardant l'accès à la donnée la plus fine, en maintenant la cohérence entre les indicateurs et dans une optique d'intégration dans la chaîne de production « industrielle » des résultats.

Cette communication a pour but de présenter les solutions retenues par Médiamétrie pour la détection et le traitement des valeurs atypiques et influentes sur Internet.

2 Quelles méthodes pour l'internet fixe ?

Un processus de détection des enregistrements atypiques a été mis en place début 2014. Il repose sur la contribution mensuelle d'un individu sur le total d'un site internet sur les trois indicateurs pages vues, temps passé et visites. L'ensemble des sites ont été regroupés en 15 classes établies à partir du nombre de panélistes ayant visité le site. Un seuil limite de contribution en pages, temps et visites, repose sur la méthode de l'écart interquartile expliquée par J. Bernier et K. Nobrega [2] (1998), est défini pour chaque classe de sites:

Seuil de contribution pages de la classe de sites *cl*

$Seuil\_P_{cl} = Med\_contrib\_max\_P_{cl} + C*(Q3\_contrib\_max\_P_{cl} - Med\_contrib\_max\_P_{cl})$

avec :

- $Med\_contrib\_max\_P_{cl}$: médiane de la contribution maximale en pages de la classe de sites *cl*
- C = 4 (ce seuil a été déterminé par des tests)
- $Q3\_contrib\_max\_P_{cl}$: au 3e quartile de la contribution maximale en pages de la classe de sites *cl*

Seuil de contribution temps de la classe de sites *cl*

$Seuil\_T_{cl} = Med\_contrib\_max\_T_{cl} + C*(Q3\_contrib\_max\_T_{cl} - Med\_contrib\_max\_T_{cl})$

Seuil de contribution visites de la classe de sites *cl*

$Seuil\_V_{cl} = Med\_contrib\_max\_V_{cl} + C*(Q3\_contrib\_max\_V_{cl} - Med\_contrib\_max\_V_{cl})$

Dès que le seuil de contribution est dépassé, l'observation est jugée comme atypique sur le site (en pages et/ou temps et/ou visites).

La correction des enregistrements atypiques s'appuie sur la suppression de pages d'un panéliste sur un site. Les panélistes n'ayant qu'une seule visite sur un site ne sont pas corrigés afin que ce processus n'affecte pas le nombre de visiteurs uniques d'un site, c'est-à-dire l'audience. Les corrections sont effectuées séparément sur les panélistes considérés comme atypiques :

- Uniquement au niveau des visites
- Uniquement au niveau des pages vues
- Uniquement au niveau du temps passé
- Uniquement au niveau du temps passé et des visites
- Uniquement au niveau des pages vues et des visites
- Uniquement au niveau des pages vues et du temps passé

en supprimant en priorité les visites les plus atypiques selon les cas. Par exemple, si un panéliste est atypique uniquement au niveau des pages vues, les visites sont supprimées par ordre décroissant sur les pages vues et par ordre croissant sur le temps passé. Le nombre de visites supprimées est fixé à partir de la somme des pages vues de ces visites afin qu'elle soit sensiblement supérieure ou égale au nombre de pages vues du panéliste atypique sur le sous-domaine concerné dépassant le seuil de détection.

Une deuxième itération partielle de détection et de traitement est effectuée sur les panélistes considérés comme atypiques sur les pages et le temps après la première correction. En effet, les panélistes initialement atypiques sur les trois indicateurs ont tendance à devenir atypiques sur les pages et le temps après le premier traitement.

Après ces corrections, il reste des panélistes atypiques selon la règle sur la contribution :

- les panélistes n'ayant qu'une seule visite sont de nouveau atypiques en fin de processus
- les panélistes peuvent rester atypiques : la nouvelle contribution dépasse à nouveau le seuil sur un ou plusieurs indicateurs
- des panélistes non atypiques avant correction peuvent le devenir si leur nouvelle contribution devient trop élevée.

Les deux derniers points sont dus au fait que la détection et le traitement des atypiques ne sont faits qu'en une itération pour la majorité des cas. De plus, les valeurs non atypiques peuvent devenir influentes après l'étape du redressement et celles-ci ne sont pas traitées actuellement.

C'est pour ces raisons que nous avons été amenés à chercher d'autres méthodes permettant de détecter et traiter les valeurs atypiques et influentes. De nombreuses méthodes existent (méthodes algébriques, graphiques ou encore probabilistes appliquées par O. A. Vasyechko, N. Benlagha et M. Grun-Rehomme [10] (2005) et O. A. Vasyechko et M. Grun-Rehomme [11] (2010)) Dans le cas des valeurs atypiques, les différentes méthodes de détection suite à nos recherches bibliographiques et à nos tests effectués deux méthodes ont été appliquées :

- la méthode de Tukey [9] (1977) est une méthode basée sur le calcul de seuils au-delà duquel les observations sont détectées atypiques. Soit l'intervalle suivant :

$$[Q_1 - \mathrm{c} \times \mathrm{IQR}\,,\ Q_3 + \mathrm{c} \times \mathrm{IQR}] \qquad (1)$$

  avec $IQR = Q_3 - Q_1$ l'écart interquartile et c un réel positif

  Si la valeur de la variable d'intérêt n'appartient pas à l'intervalle, elle est déclarée atypique. Tukey détermine une frontière basse (inner fences) pour c = 1.5 et une frontière haute (outer fences) pour c = 3 comme l'explique Songwon Seo [8] (2002).

  Cette méthode a été appliquée sur la variable « temps de connexion ». En amont, une transformation Box-Cox a été appliquée dans le but de rapprocher notre variable d'une distribution normale. Cette détection est très restrictive et peu d'observations sont détectées comme étant atypiques.

Comme introduit par A. Ruiz-Gazen [7] (2009), des méthodes multivariées peuvent être envisagées pour détecter les données atypiques. La méthode présentée ci-dessous en fait partie.

- Les K-means : il s'agit d'une méthode multivariée permettant de détecter les individus atypiques au regard des trois indicateurs (pages, temps et visites). Cette méthode de partitionnement des données est très sensible aux valeurs atypiques ; elle peut donc être utilisée pour une détection automatique.

  L'ensemble des sites ont été regroupés en 16 classes établies à partir du nombre de panélistes ayant visité le site. La détection par K-means a été lancée sur 16 classes de sites.

Les comportements internet sont différents sur chaque classe. En jouant sur une détection par classe et non au global, une observation avec un nombre de pages, temps et visites élevé pour un site à faible audience pourra être détecté. La détection a été lancée sur des données normalisées.

La méthode de Tukey va permettre de détecter l'inactivité des sessions internet tandis que la méthode des K-means va permettre de détecter des comportements « inhabituels ».

En ce qui concerne le traitement des valeurs détectées par les deux méthodes présentées ci-dessus, nous avons préconisé :

- la modification des valeurs : ramener les valeurs détectées par la méthode de Tukey au seuil de détection.
- la suppression de pages pour les valeurs détectées par les K-means. Les pages à supprimer sont obtenues en utilisant l'algorithme des plus proches voisins

Lors des traitements de fusion et modélisation, des observations sont dupliquées. Afin de ne pas accentuer les observations influentes, une réduction du nombre de duplications pour les valeurs atypiques/influentes a été préconisée. Les observations sont détectées par une méthode avec un seuil ou par winsorization selon le cas.

Pour mesurer l'influence en fin de processus, le biais conditionnel sera étudié ainsi que la winsorization pour le traitement des valeurs influentes. Dans le but de réduire l'impact des individus influents, la winsorization consiste à réduire la valeur et/ou le poids de ces individus de la manière suivante (C. Favre Martinoz, D. Haziza et J.-F. Beaumont [4] (2012)):

Winsorization de type 1 :

Soient : K une réel positif, n la taille de l'échantillon, i = 1 à n, d_i le poids de l'individu, y_i la valeur de la variable d'intérêt pour l'individu i, les valeurs de la variable d'intérêt après winsorization sont défini par :

$$\widetilde{Y}_i = \begin{cases} \frac{K}{d_i}, \ si\ y_i d_i > K \\ y_i, \ sinon \end{cases} \quad (2)$$

Le poids après winsorization est donnée par :

$$\widetilde{d}_i = \begin{cases} \frac{K}{y}, \ si\ y_i d_i > K \\ d, \ sinon \end{cases} \quad (3)$$

Winsorization de type 2 :

Même notations. Les valeurs de la variable d'intérêt après winsorization sont défini par :

$$\widetilde{Y}_i = \begin{cases} \frac{K}{d_i} + \frac{1}{d_i} * \left(y_i - \frac{K}{d_i}\right), \ si\ y_i d_i > K \\ y_i \qquad\qquad , \ sinon \end{cases} \quad (4)$$

Le poids après winsorization est donnée par :

$$\widetilde{d}_i = \begin{cases} 1 + (d_i - 1) * \left(\frac{K}{y_i d_i}\right), \ si\ y_i d_i > K \\ d_i \qquad\qquad , \ sinon \end{cases} \quad (5)$$

Les indicateurs de références (pages, temps et visites) étant corrélés positivement, cette méthode a été implémentée en fin de processus sur la variable « nombre de pages vues » (méthode inspirée de la méthode utilisée à l'INSEE [3] (2013)). Les valeurs de la variable ne sont pas modifiées car le processus de modification en interne est trop complexe. Le poids a donc été réduit

selon la formule (5). Peu d'individus sont impactés, la répartition du poids n'est modifiée que très légèrement et la structure des variables de redressement est encore respectée après winsorization. De surcroît, il parait évident que les usages internet ne sont pas les mêmes selon que l'on regarde les navigations d'un site de banque, d'une boite de messagerie, d'un réseau social ou d'un site de streaming. Un individu peut être influent sur une sous-catégorie de site et ne pas l'être sur les autres sous-catégories de sites, ce qui au global ne se verrait pas. Ainsi, pour prendre en compte ce type de comportement, la winsorization a été réalisée par sous-catégorie de site. Une approche conditionnelle présentée par E. Lesage [5] (2012), reposant sur le fait d'utiliser une solution auxiliaire aurait pu être envisagée.

Le travail sur les observations atypiques et influentes est extrêmement complexe et demande un travail minutieux. Nous avons choisi des méthodes cohérentes avec nos données et avec les résultats à produire. Les méthodes utilisées traitent les formes d'atypisme et d'influence présentes dans le panel de l'internet fixe. Il n'existe pas de solution unique, mais des précautions sont à prendre.

Bibliographie

[1] Beaumont J.-F. et Haziza D. (2012), Estimation en présence de valeurs influentes dans les enquêtes, *Septième Colloque Francophone sur les Sondages ENSAI (Rennes, France)*

[2] Bernier J. and Nobrega K. (1998), Outlier detection in asymmetric samples: a comparaison of an inter-quartile range method and a variation of sigma gap method, *SSC Annual Meeting*.

[3] Brion P., Gros E.et Guggemos F. (2013), La gestion des unités influentes dans l'ESA par winsorisation, Insee.

[4] Favre Martinoz C., Haziza D. et Beaumont J.-F. (2012), Une méthode simple de détermination du seuil pour la winsorisation. [5] Lesage E. (2012), Utilisation d'une approche conditionnelle pour traiter les outliers, *7e Colloque Francophone sur les Sondages Ensai*, Bruz.

[5] Lesage E. (2012), Utilisation d'une approche conditionnelle pour traiter les outliers, *7e Colloque Francophone sur les Sondages Ensai*, Bruz.

[6] Rousseeuw P. J. and Leroy A. M. (1987), Robust Regression and Outlier Detection, *New York: John Wiley and Sons.*

[7] Ruiz-Gazen A. (2009), Détection d'observations atypiques multivariés au sein de données spatiales, Toulouse School of Economics et IMT.

[8] Songwon Seo (2002), A Review and Comparison of Methods for Detecting Outliers in Univariate Data Sets, *Master's Thesis*, University of Pittsburgh.

[9] Tukey J.W. (1977), Exploratoy Daa Analysis, *Ed. Addison-Wesley*.

[10] Vasyechko O. A., Benlagha N. et Grun-Rehomme M. (2005), Comparaison de méthodes de détection des valeurs extrêmes : Application en statistique d'entreprise, Paris.

[11] Vasyechko O. A et Grun-Rehomme M. (2010), Méthodes de détection des unités atypiques: Cas des enquêtes structurelles ukrainiennes, *"42èmes Journées de Statistique"*, Marseille.