Mediametrie

DETECTION ET TRAITEMENT DES VALEURS ATYPIQUES DE LA MESURE INTERNET GLOBAL

Magdalena Auvinet, Médiamétrie, France, mauvinet@mediametrie.fr

Mathieu Hostin, Médiamétrie, France, mhostin@mediametrie.fr

Résumé. Cette note explique la méthode pour détecter et traiter les observations atypiques au sein du surf des panélistes sur les trois écrans : ordinateur, mobile et tablette, sur lesquels est basée la mesure Internet Global. La méthode choisie pour détecter les observations atypiques est l'Isolation Forest, méthode qui détecte les anomalies via isolement. Graphiquement, l'algorithme va séparer les données par des lignes orthogonales et va attribuer un score d'anomalie qui sera plus ou moins élevé selon la difficulté que l'algorithme aura eu à séparer une observation des autres. Une observation qui est facilement séparable des autres est une observation qui est logiquement plus atypique que les autres observations. La méthode choisie pour traiter les observations atypiques traite ces anomalies au cas par cas afin de s'adapter à nos besoins métiers. Elle a pour objectif de diminuer l'impact de ces données sans supprimer entièrement les informations de la mesure d'audience qu'elles peuvent apporter.

Mots-clés. Mesures d'Audience et Estimation Robuste, Détection valeurs atypiques, Traitement valeurs atypiques, Isolation Forest

Abstract. This note explains the method for detecting and processing outliers within the surfing of panelists on three screens: computer, smartphone, and tablet, on which the Global Internet measurement is based. The method chosen to detect outliers is the Isolation Forest, a method that detects anomalies through the data's isolation. Graphically, the algorithm will separate the data by orthogonal lines and will assign an anomaly score which will be higher or lower depending on the difficulty the algorithm has had in separating one observation from the others. An observation that is easily separable from others is an observation that is logically more atypical that the other observations. The method chosen to process outliers processes these anomalies on a case-by-case basis to adapt the different cases to our business needs. Its purpose is to reduce the impact of this data without entirely removing the audience measurement information that it may provide.

Keywords. Audience Measurement and Robust Estimation, Outliers detection, Outliers processing, Isolation Forest

Contexte

Une observation atypique trop éloignée du reste des données peut avoir un effet néfaste sur l'analyse de ces dernières car il peut être mal interprétée et peut biaiser l'ensemble de l'analyse, en particulier lors d'une mesure d'audience. C'est pourquoi il est important de détecter et de traiter ces observations atypiques. Cette étape intervient tôt dans le processus de la mesure d'audience afin de traiter les données en amont des étapes qui risqueraient de multiplier la dimension atypique de ces observations. La méthode qui est présentée dans cette note a été pensée et conçue au sein du Pôle Internet de la Direction Data Science de Médiamétrie.

Médiamétrie est l'acteur référent de la mesure d'audience en France de la Télévision, de la radio et d'Internet. Le Pôle Internet produit chaque mois les données d'audience de la consommation internet des Français à travers trois écrans : l'ordinateur, le mobile et la tablette. La mesure internet fonctionne à travers le suivi de la consommation internet d'environ 25 000 panélistes. Tout un processus (composé d'étapes de nettoyage des données, de traitements statistiques, etc.) est déroulé tous les mois afin d'obtenir, à partir de l'activité de ces panélistes, la consommation internet française globale. Durant le processus, on a un traitement spécifique à chacun des trois écrans : la méthode de détection et de traitement des observations atypiques se déroule au cours de chacun d'entre eux afin de corriger les valeurs atypiques de chaque écran indépendamment des autres. Lors de ces traitements, on travaille sur une base de tickets agrégée, la partie suivante de cette note aborde la composition de cette base.

1. Mise en place : bases utilisées et définitions

Il existe une base de tickets pour chacun des trois écrans. Chacune de ces bases de données listent toute la consommation internet de chacun des panélistes, c'est-à-dire toutes les applications et sites qu'ils visitent.

Pour l'ordinateur :

- Seuil S de 40 panélistes[1]
- Seuls les sous-domaines[2] ayant un nombre de panélistes égal ou supérieur à S sont concernés
- On **agrège** ensuite les informations de pages et de temps au niveau sous-domaine.
- Le nombre de visites est aussi calculé et dépend de la durée entre deux pages visitées.
- La base obtenue donne alors le nombre de pages, la durée et le nombre de visites distinctes que chaque panéliste a effectué dans chaque sous-domaine.

Définition pour l'ordinateur :

Pages = Somme des pages visitées par un panéliste sur un sous-domaine

Temps = Somme du temps passé par un panéliste sur un sous-domaine

*Visites = Somme des visites d'un panéliste sur un sous-domaine. Les visites sont définies à partir des tickets. Une visite correspond à un ticket sauf si le temps entre deux tickets est supérieur à **30 minutes** : dans ce cas-là, une nouvelle visite est comptée à partir de 30 minutes depuis le dernier ticket.*

[1] Pourquoi 40 panélistes ? Des travaux internes ont été mis en œuvre il y a quelques années pour déterminer ce seuil optimal pour traiter des sous-domaines qui ont suffisamment de visiteurs uniques nécessaires au traitement des atypiques.

[2] Sous-domaine : sous-ensemble du domaine
exemple : sousdomaine.domaine.com

Pour le mobile et la tablette :

- Seuil S de 40 panélistes pour le mobile, 20 panélistes pour la tablette[3]
- Seules les Brands[4] ayant un nombre de panélistes égal ou supérieur à S sont concernés
- On distingue deux paramètres
    a. SITE : tickets associés à une activité sur un site
    b. APP : tickets associés à une activité sur une application

Cette séparation permet de traiter équitablement les parties SITE et APP, parties qui ne sont pas consommées de la même manière par les panélistes

- Pour chaque périmètre :
    a. On **agrège** ensuite les informations de pages et de temps au niveau Brands.
    b. Le nombre de visites est aussi calculé et dépend de la durée entre deux pages visitées.
    c. La base obtenue donne alors le nombre de pages, la durée et le nombre de visites distinctes que chaque panéliste a effectué dans chacune des Brands.
- Les deux bases sont ensuite regroupées

Définition pour le mobile et la tablette :

Pages = Somme des pages surfés par un panéliste sur une brand

Temps = Somme du temps passé par un panéliste sur une brand

*Visites = Somme des visites d'un panéliste sur une brand. Les visites sont définies à partir des tickets. Une visite correspond à un ticket sauf si le temps entre deux tickets est supérieur à **5 minutes** : dans ce cas-là, une nouvelle visite est comptée à partir de 5 minutes depuis le dernier ticket.*

Finalement, la base obtenue, pour chacun des écrans, est composée de cinq variables : deux variables permettant l'identification du panéliste et du sous-domaine (ou Brand pour mobile et tablette), le nombre de pages vues, le temps passé dessus et le nombre de visites effectuées. Une ligne correspond donc à la somme des pages, du temps et du nombre de visite dans le mois qu'un panéliste a effectué au sein d'un sous-domaine/Brand (voir exemple ci-dessous).

La méthode de détection et de traitement des observations atypiques pour l'ordinateur est effectuée sur cette base de données. Pour le mobile et la tablette, on utilise le niveau Brand à la place du niveau sous-domaine car c'est le niveau le plus fin pour ces deux écrans.

[3] La consommation internet sur le mobile se rapproche de l'ordinateur. Cependant sur tablette, on a moitié moins de panélistes, donc on divise le seuil par deux sinon beaucoup trop de Brands échapperaient au traitement des atypiques.

[4] Brand : Agrégat de domaines, sous-domaines, d'URLS identifiés par un logo unique. Exemple : la marque Google.

id_ss_dom	RN_ID	nb_pages	nb_temps	nb_visites
981	509826913480102	2	2	2
981	509849886630201	17	749	2
981	509896430680102	1	4	1
981	509917832620101	1	24	1
981	509917832620301	1	20	1
981	509921171130101	18	535	2
1033	133014065380301	6	91	2
1033	133089883560102	30	182	11
1033	133092533440101	7	191	1
1033	133096434520401	85	1936	23
1033	133239608530101	3	69	1
1033	133246613840201	1	44	1
1033	133257832700101	15	759	5
1033	133258213960101	14	340	5
1033	133268597210201	1	44	1
1033	509159372690101	3	47	1
1033	509231599600101	6	201	3

Exemple 1 : Extrait de quelques lignes de la base utilisée. Le panéliste ordinateur dont le RN_ID est 509849886630201 a visité 2 fois le sous-domaine dont l'ID est 981 où il a au total vu 17 pages pour un total de temps passé sur ce sous-domaine de 749 secondes.

A l'issue de cette étape on a une base par écran, composée des informations des individus (Pages / Temps / Visites) sur chacun des sous-domaines/Brand. Ces tables seront enrichies par la suite par une nouvelle variable qui donne le nombre d'individu par sous-domaine/Brand : cela permettra à l'algorithme de distinguer les sous-domaines/Brand qui sont visités par énormément de panélistes que ceux très peu fréquentés (car la façon de les consommer peut-être différente). On parlera de cette variable comme de la taille du sous-domaine/Brand.

Avant de commencer la méthode de détection et traitement des observations atypiques, nous avons donc une table composée :

- Pour l'ordinateur :

ID sous-domaine	ID individu	NB pages	NB temps	NB visites	Taille du sous-domaine

- Pour le mobile et la tablette (deux tables distinctes) :

ID brand	ID individu	NB pages	NB temps	NB visites	Taille de la brand

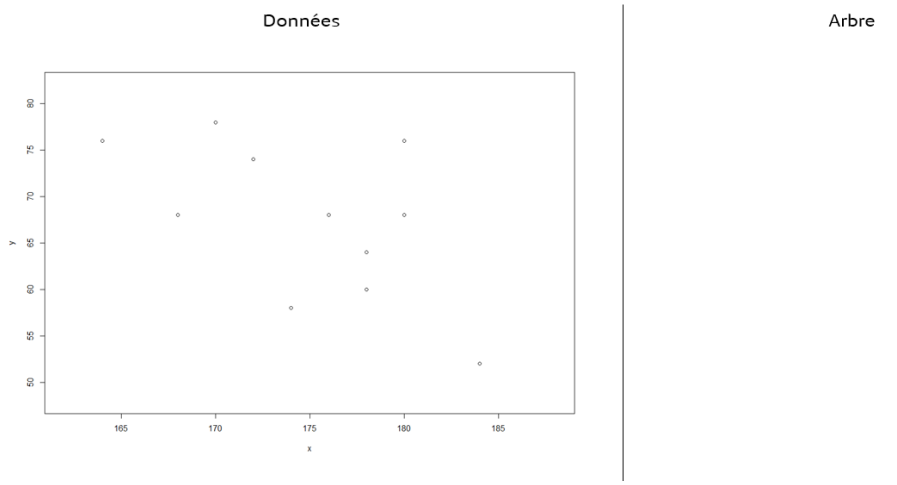
2. Méthode

2.1. Détection

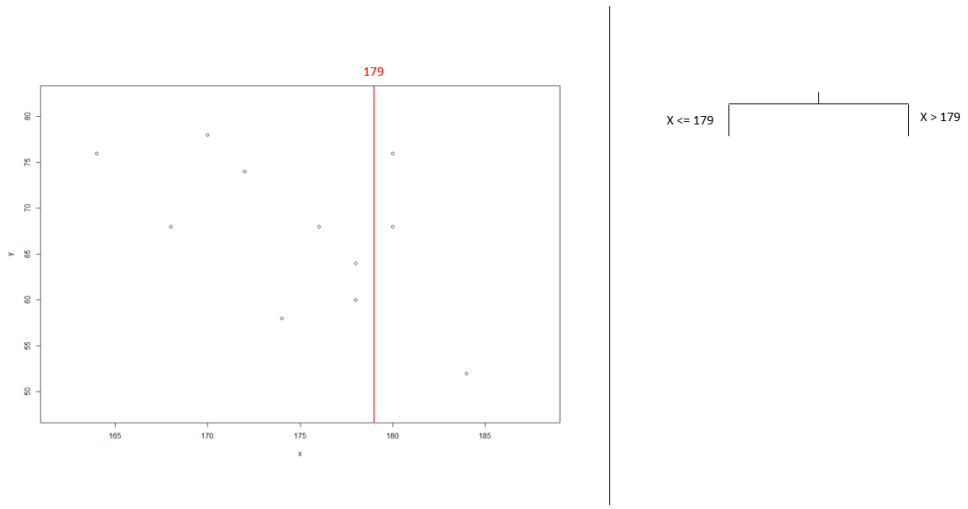
Pour détecter une observation atypique, on utilise la méthode de *l'Isolation Forest*. Cette dernière attribue un score d'anomalie pour chaque observation, ce score représente leur tendance atypique au sein du jeu de données. L'algorithme isole toutes les observations : pour cela il choisit une des variables (pages, temps, visites et nombre de panéliste du sous-domaine/brand de l'observation) et fixe un seuil, au hasard, pour séparer les valeurs qui dépassent le seuil de celles qui ne le dépassent pas. C'est comme un arbre de décision où il existe une branche pour chacune des observations. Ci-dessus une illustration du fonctionnement de l'algorithme.

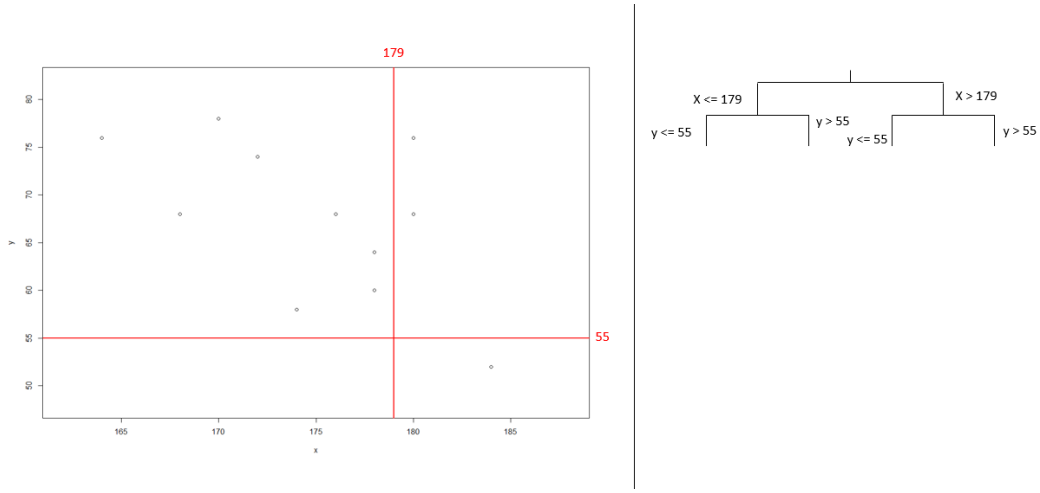
<u>Illustration du fonctionnement de l'Isolation Forest :</u>

Par exemple, un jeu de données aléatoire composé de deux variables x et y.

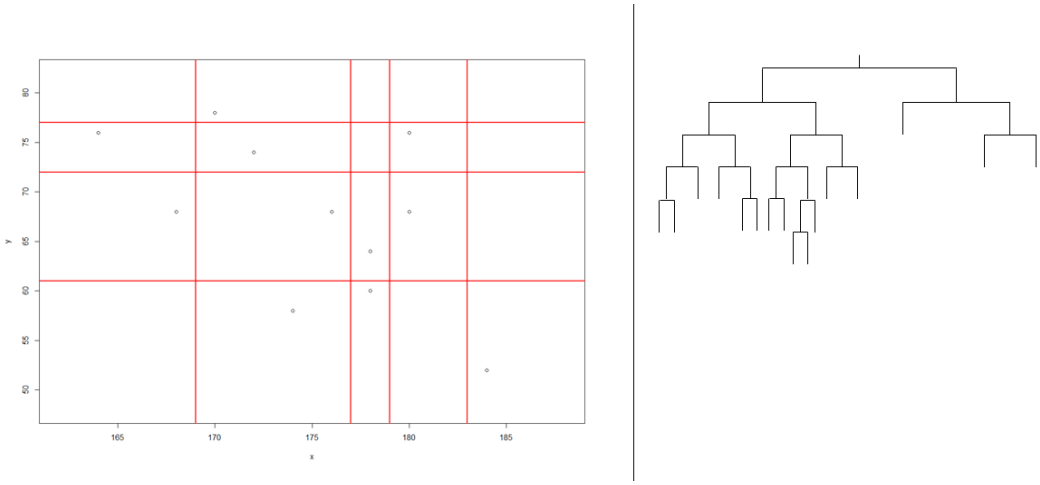


L'algorithme va séparer aléatoirement les données, ce découpage est représenté par une branche dans l'arbre. Chaque nouvelle séparation sera une nouvelle branche dans l'arbre.





L'algorithme s'arrête une fois que toutes les données ont été séparées.



Ce processus est répété N = 500 fois pour, à la fin, obtenir un score d'anomalie $S(x_i, N)$ selon le calcul suivant[5] :

$$S(x_i, N) = 2^{\frac{-E(h(x_i))}{c(N)}}$$

Sachant que :

x_i est une observation (une ligne) étant composée de 4 variables : Pages, Temps, Visites et Nombre de panéliste.

N est le nombre d'arbres créés

h(x) est la profondeur[6] du chemin que l'observation *x* a suivit

[5] La valeur N = 500 a été choisie après de nombreux tests et en suivant les recommandations d'articles scientifiques.

[6] Profondeur : se mesure selon le nombre de nœuds (qui représentent chaque étage de séparations dans l'arbre de décessions) nécessaires pour isoler une observation

Et que :

$$c(N) = 2H_{N-1} - \frac{2(N-1)}{N}$$

Avec :

$c(N)$ est la profondeur moyenne, d'un arbre, nécessaire pour isoler les observations

H_{N-1} est le nombre harmonique et il peut être estimé par ln(N-1) + 0.5772156649 (constante d'Euler–Mascheroni).

Chaque observation obtient un score allant de 0.5 à 1.

$S(x_i, N) = 0.5$ est une observation dite « normale » (elle suit la tendance globale des observations) et $S(x_i, N) = 1$ est une observation totalement atypique par rapport à l'ensemble des données.

Finalement, les observations ayant un score d'anomalie supérieur ou égal à 0.7 seront définis comme atypiques.[7]

Avant de passer à la phase de traitement, on détermine la variable (ou les variables) à « l'origine » du caractère atypique des observations. Pour cela, on utilise un procédé inspiré de la méthode de *winsorisation* : si l'observation (atypique) est supérieure en nombre de pages, de temps ou de visites à la plus grande observation (non atypique) dans le même sous-domaine, alors cette observation est atypique en pages, en temps ET/OU en visites. En détail :

	$x_{ap} > \max(x_p)$	$x_{at} > \max(x_t)$	$x_{av} > \max(x_v)$
Atypique Pages	**x**		
Atypique Temps		**x**	
Atypique Visites			**x**
Atypique Pages et Temps	**x**	**x**	
Atypique Pages et Visites	**x**		**x**
Atypique Temps et Visites		**x**	**x**
Atypique Pages et Temps et Visites	**x**	**x**	**x**

Tableau 1 : Processus de définition de l'observation atypique observée.

[7] Le seuil à 0.7 a été choisi d'un commun accord avec l'équipe Business. Il permet de ne pas trop, ni trop peu, détecter et corriger des valeurs atypiques.

Sachant que :

x_a est une observation atypique d'un sous-domaine

x_{ap} est le nombre de pages d'une observation **atypique** d'un sous-domaine

x_{at} est le temps passé d'une observation **atypique** d'un sous-domaine

x_{av} est le nombre de visites d'une observation **atypique** d'un sous-domaine

$max(x_p)$ est le plus grand nombre de pages parmi les observations **non atypiques** d'un sous-domaine

$max(x_t)$ est le temps passé le plus élevé parmi les observations **non atypiques** d'un sous-domaine

$max(x_v)$ est le plus grand nombre de visites parmi les observations **non atypiques** d'un sous-domaine

2.2. Traitement

La partie de traitement se base sur la suppression de visites entières. On supprime des visites selon un ou des objectifs définis, on arrête de supprimer à partir du moment où le ou les objectifs de suppression ont été atteints ou dépassés. Il existe aussi certains cas spéciaux quand l'observation n'a qu'une visite car on ne souhaite pas supprimer d'audience (donc on ne peut pas supprimer la dernière visite d'une observation).

La suppression se fait toujours sur les visites (sauf cas de visite unique). Le choix de quelles visites sont supprimées en priorité se fait selon la variable (ou les variables) à l'origine du caractère atypique des observations. Une observation dite « Atypique **Pages** » verra ses visites ayant le plus de **pages** être supprimées en priorité, jusqu'à l'atteinte de l'objectif de suppression de **pages**.

Les différents cas de traitement des observations atypiques :

CAS 1 « Atypiques Visites Uniquement » : On supprime, jusqu'à objectif, prioritairement les visites ayant le moins de temps et de pages.

CAS 2 « Atypiques Pages Uniquement » :

- Si Visite > 1 : On supprime, jusqu'à objectif, prioritairement les visites ayant le plus de pages.
- Si Visite = 1 : On supprime les pages ayant le moins de temps, jusqu'à objectif.

CAS 3 « Atypiques Temps Uniquement » :

- Si Visite > 1 : On supprime, jusqu'à objectif, prioritairement les visites ayant le plus de temps.
- Si Visite = 1 : On supprime les pages ayant le plus de temps, jusqu'à objectif.

CAS 4 « Atypiques Temps et Visites » : On supprime les visites ayant un temps le plus proche du temps moyen par visite à supprimer.

CAS 5 « Atypiques Pages et Visites » : On supprime les visites ayant un nombre de pages le plus proche du nombre de pages moyen par visite à supprimer.

CAS 6 « Atypiques Pages et Temps » :

- Si Visite > 1 : On supprime les visites ayant un temps par page le plus proche du temps par page moyen à supprimer.
- Si Visite = 1 : On supprime les pages avec le plus de temps jusqu'à objectif temps puis on supprime les pages avec le moins de temps pour finir l'objectif temps restant.

CAS 7 « Atypiques Pages, Temps et Visites » : On supprime les visites ayant un temps par page le plus proche du temps par page moyen à supprimer.

CAS 8 dit Spécial « Non Atypiques Pages, Temps et Visites » : On supprime les visites ayant un temps par page le plus proche du temps par page moyen à supprimer.

Les objectifs de suppression sont calculés selon la différence entre les 1% d'observations[8] les plus grandes du sous-domaine avec la variable atypique à corriger.

Pages à supprimer = pages – médiane(1% des plus grandes observations non atypiques)

Temps à supprimer = temps – médiane(1% des plus grandes observations non atypiques)

Visites à supprimer = visites – médiane(1% des plus grandes observations non atypiques)

Ce traitement permet de réduire l'effet atypique de ces variables, tout en gardant l'information qu'elles sont parmi les plus élevées du sous-domaine en question (point important lors d'une mesure d'audience : on veut garder l'information qu'un panéliste derrière un comportement atypique reste un panéliste qui a une forte consommation). Ci-dessous, un exemple graphique d'un traitement des valeurs atypiques des pages d'un sous-domaine.

[8] On a observé, après le traitement, que la nouvelle distribution des observations est plus naturelle et ne crée pas d'effet de cluster autour du point le plus élevé si on utilise la médiane des 1% des plus grandes valeurs non atypiques (ce qui aurait comme problème la création de nouvelles observations atypiques).

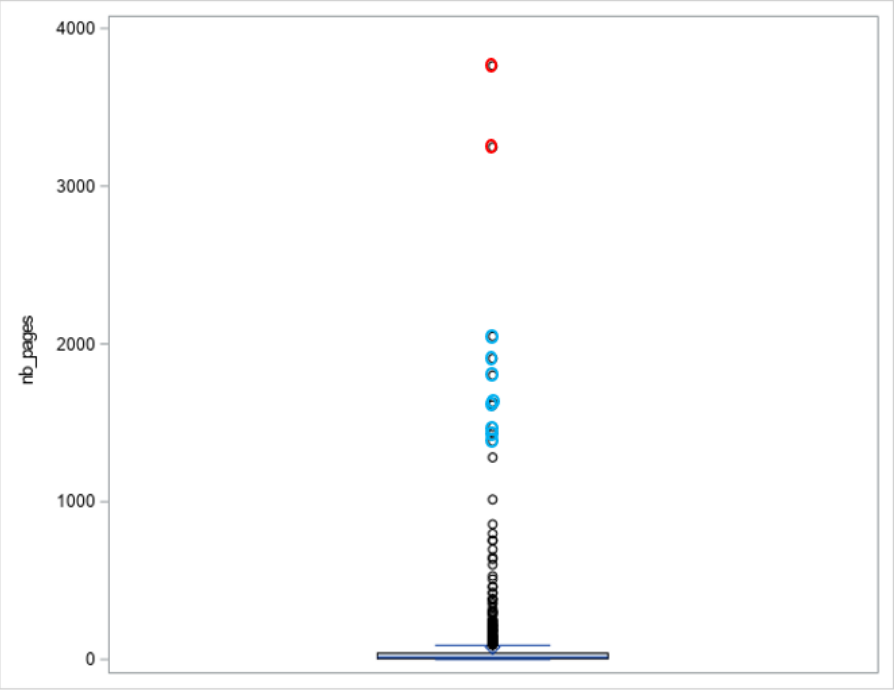


Figure 1 : Exemple sur un sous-domaine, avant traitement. En rouge : les observations atypiques du sous-domaine représenté. En bleu : les 1% d'observations non atypiques ayant le plus de pages du sous-domaine représenté. Après traitement, les valeurs anciennement atypiques (en rouge) seront réduites pour être aux environs de la médiane des 1% plus grandes valeurs non atypiques (en bleu).

La recherche de la meilleure méthode de détection et de traitement des observations atypiques est compliquée et doit se faire au cas par cas, selon les besoins et les problématiques apportées par les différentes sources de données.

La méthode de détection et traitement des observations atypiques est appliquée officiellement en production de la mesure d'Internet Global depuis les données de janvier 2022.

Bibliographie / Webographie

Maurras Togbe, Yousra Chabchoub, Aliou Boly, Raja Chiky. Etude comparative des méthodes de détection d'anomalies. Revue des Nouvelles Technologies de l'Information, Editions RNTI, 2020, Extraction et Gestion des Connaissances EGC 2020.

Rongfang Gao, Tiantian Zhang, Shaohua Sun, Zhanyu Liu. Research and Improvement of Isolation Forest in Detection of Local Anomaly Points, 2019 Journal of Physics: Conference Series.

Isolation Forest :
https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf?q=isolation-forest

Isolation Forest algorithm for anomaly detection
https://medium.com/@arpitbhayani/isolation-forest-algorithm-for-anomaly-detection-f88af2d5518d

Outlier Detection with Isolation Forest
https://towardsdatascience.com/outlier-detection-with-isolation-forest-3d190448d45e

Le package R :
https://cran.r-project.org/web/packages/solitude/solitude.pdf