Journées de méthodologie statistique de l'Insee 2025

DÉTECTION DES RÉPONDANTS SUSPECTS DANS UNE ENQUÊTE EN LIGNE

Diane MAILLOT-TCHOFO (*), Tom DEVYNCK (**), Fabienne LE SAGER (*), Louis MAREC (*)

(*) Médiamétrie
(**) Médiamétrie et Toulouse School of Economics

dmaillottchofodinda@mediametrie.fr flesager@mediametrie.fr

Mots-clés (6 maximum) : qualité de la donnée, enquête en ligne, questionnaire auto-administré, répondants négligents, erreur d'observation.

Domaines : Contrôle et correction des données, data editing - Identification et traitement d'erreurs d'observation.

Résumé

L'erreur d'observation la plus répandue dans les enquêtes est associée à l'incapacité ou au refus des répondants de fournir la bonne réponse [1]. Dans ce contexte, la volonté d'estimer correctement les niveaux de possession d'équipements multimédia (e.g. télévision, smartphone) a conduit Médiamétrie à développer une approche combinant deux méthodes pour détecter dans un dispositif multi-mode les répondants négligents, uniquement pour ceux qui répondent en ligne (CAWI). L'un des principaux défis de cette étude réside dans la nature des données, à savoir une enquête déjà en production.

Les travaux de Laura Gamble [2] et d'Anvita Mahajan [3] nous ont éclairés et amenés à élaborer une méthode hybride combinant deux approches. La première approche utilise les durées de complétion du questionnaire. En transformant l'inverse des temps de complétion de chaque module de l'enquête (7 dans notre cas), une somme pondérée peut être calculée pour chaque répondant. Les réponses négligentes sont déterminées en fonction d'un seuil de loi de probabilité Khi-carré.

La seconde consiste en un algorithme de partitionnement en deux étapes basées sur la possession d'équipements. Nous avons appliqué un algorithme K-Means sur les caractéristiques sociodémographiques des ménages des répondants. Dans la mesure où l'équipement des ménages dépend fortement des individus qui les composent (nombre, âge, etc.), l'objectif est d'examiner ensemble les répondants dont l'équipement et l'utilisation sont potentiellement similaires (c'est-à-dire les réponses au questionnaire). Ensuite, nous avons appliqué des modèles DBSCAN et Isolation Forest à chaque cluster pour détecter les répondants les plus distants de leur groupe tout en limitant les insuffisances respectives des modèles. Nous avons constitué la liste définitive de répondants suspects en combinant les résultats issus des deux méthodes.

Les données n'étant pas labelisées, nos résultats ont été évalués sur la base des caractéristiques des répondants présumés négligents par rapport à celles de l'ensemble de la population étudiée.

Les analyses présentées dans l'article reposent sur deux jeux de données correspondants à deux vagues de l'enquête. L'ensemble de données ayant servi à élaborer les modèles comptait près de 8 000 répondants, dont environ 180 étaient classés comme négligents. Un autre ensemble de données d'une autre vague de même taille a donné 304 répondants suspects. Les caractéristiques des répondants négligents et les écarts statistiquement significatifs par rapport à l'ensemble de la population étudiée étaient cohérents avec les attentes et les hypothèses formulées en amont. En effet, les personnes de moins de 35 ans sont surreprésentées (respectivement 27% contre 13%), tandis que les retraités sont sous-représentés (21% contre 29%).

L'étude réalisée nous a permis d'identifier les profils des répondants suspects, ce qui constitue un pas de plus pour prévenir les erreurs d'observation et veiller à la qualité des données recueillies.

Abstract

The most familiar observational error within surveys is associated with respondents' inability or unwillingness to provide the correct answer. In this context, the will to correctly estimate digital devices ownership (e.g. TV, smartphone) drove Médiamétrie to develop an ambivalent method to detect neglectful respondents in a Computer-Aided Web Interview (CAWI). We drew from the literature (Laura Gamble, 2023 and Anvita Mahajan, 2023) works and derived an ambivalent method combining both approaches. Our first approach makes use of the questionnaire's completion times. The second approach is a two-step clustering algorithm focused on the ownership of digital equipment. A K-Means algorithm was applied on the respondent's household socio-demographic characteristics. Then, machine learning models were applied to each cluster to contain the models' shortcomings. Our final list of sloppy respondents was obtained by combining the results of the two approaches.

1 Introduction

Les données d'enquête jouent un rôle crucial dans de nombreux processus décisionnels au sein des institutions publiques et privées. En particulier, l'enquête fournissant des données sur les équipements et abonnements audiovisuels des foyers français est indispensable au cadrage de plusieurs mesures, dont celles de la télévision et d'internet. La qualité de la donnée et donc des résultats produits dépend fortement de la fiabilité des réponses obtenues. Plusieurs types et sources d'erreurs peuvent affecter la qualité des données recueillies, qu'ils soient induits par l'agent enquêteur, le non-respect des consignes de passation, la saisie erronée des réponses obtenues, ou introduites par l'enquêté, le refus de participer à l'enquête, les réponses inexactes données sciemment ou encore la mauvaise compréhension de la question ou des réponses possibles.

Dans cette étude, nous nous sommes focalisés sur l'identification des répondants négligents, qui auraient, volontairement ou non, donné des réponses qui ne reflètent pas la réalité de leur quotidien ou de leur foyer. Notre cadre de recherche est une enquête multi-mode à la fois séquentielle et concurrentielle. De ce fait, un des enjeux de ce travail réside dans la nature de nos données. En effet, elles sont issues d'une enquête en production, et nous ne disposons pas de données d'entraînement ou d'information préalable concernant les mauvais répondants.

Aujourd'hui de nombreux contrôles de cohérence sont effectués entre les réponses d'un même enquêté, à la fois pendant la passation du questionnaire et après la récolte des réponses. Cependant, ces contrôles sont tous issus de l'expertise acquise par les collaborateurs travaillant sur le dispositif et sont implémentés sous forme d'algorithmes "si-alors" basiques. Le cœur de notre travail réside

dans la mise en place de contrôles automatiques, après la récolte de l'ensemble des données.

La plupart des méthodes existantes résolvent le problème en utilisant les données numériques et la notion d'*outlier* ou anomalie. La principale problématique dans la détection de réponses frauduleuses est que celles-ci ne se présentent pas toujours comme des données extrêmes. Au contraire, elles apparaissent parfois comme des réponses moyennes, cela peut-être le cas lorsqu'un enquêté sélectionne des réponses aléatoirement pour aller vite, ou systématiquement le même item (e.g. réponse A) tout au long du questionnaire . L'objectif est donc de croiser au maximum nos types de données afin de dégager les comportements et réponses dites négligentes, voire frauduleuses.

2 Contexte & Cadrage

Au sein de Médiamétrie, l'enquête du baromètre des équipements (BDE) sert de référence pour l'ensemble des équipements multimédias, des abonnements vidéo et des types d'accès internet. Elle permet de suivre l'évolution du parc d'équipements, de recueillir le multi-équipements et de détecter et suivre l'émergence de nouveaux équipements. Elle est aussi l'une des bases de cadrage des mesures d'audience de la télévision et d'internet produites par l'entreprise (qui font référence sur le marché).

2.1 Contexte de l'enquête en ligne

L'étude se déroule en deux phases : recrutement des foyers par téléphone (à partir de numéros générés aléatoirement et de numéros fixes issus d'annuaires), au cours duquel la composition du foyer ainsi que ces caractéristiques sociodémographiques sont récoltées. A la suite de quoi un guide d'accompagnement présentant les équipements avec des visuels est envoyé à l'enquêté pour lui permettre d'identifier ses équipements de la manière la plus juste possible. La seconde phase constitue le recueil des réponses par téléphone ou en ligne. Le mode de collecte est choisi par le répondant à la fin de la première phase ; environ 80% des répondants choisissent de répondre en ligne au questionnaire, cf Fig.1. Le tout se déroule sans aucune rétribution pour l'enquêté.

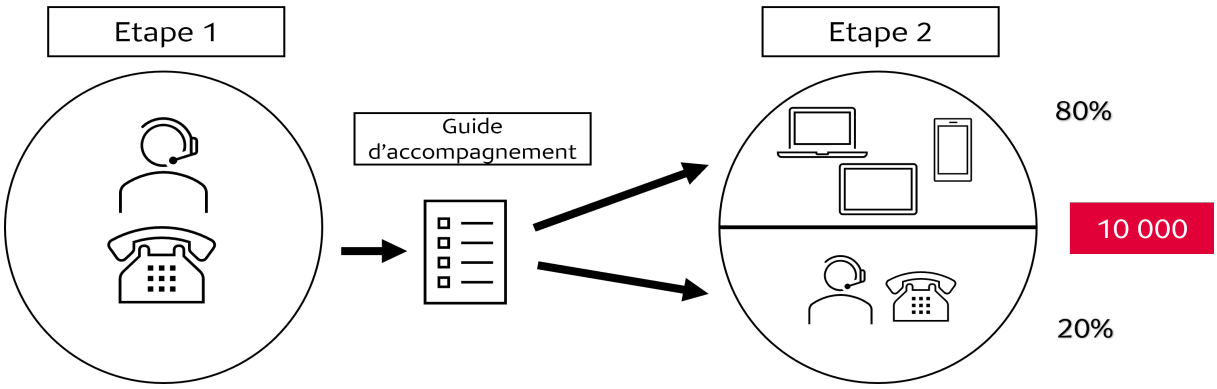


FIG. 1 – *Schema du processus de l'enquête du Baromètre des équipements*

Nous nous intéresserons ici uniquement aux 80% qui répondent au questionnaire Cawi. Ce dernier est composé de sept modules, tous concernent une catégorie d'équipement spécifique. On retrouve, dans cet ordre, des modules : de réception internet, télévision, consoles de jeux, ordinateur, tablette, téléphonie mobile et enfin le module "autres" équipements (ex : enceinte connectée). De plus, dans le module télévision, une boucle est présente, en effet tout un ensemble de questions (comme la

fréquence d'usage) est posé pour chaque télévision du foyer. Ce questionnaire est donc long, avec un temps de passation médian supérieur à 12 minutes. Cette longueur peut être un facteur de lassitude et perte d'attention chez le répondant.

Les données sont récoltées auprès de 10 000 foyers constitutifs d'un échantillon représentatif de la population des ménages français (en se basant sur les données de l'Insee), et ce de manière semestrielle, soit environ 20 000 foyers par an.

2.2 Cadrage de l'étude & données disponibles

Notre étude porte sur les données du second semestre 2024 et le premier semestre 2025, en deux jeux de données distincts. Nous disposons des réponses des enquêtés mais aussi des paradonnées récoltées par les progiciels d'enquêtes, ainsi que des métadonnées.

Les paradonnées sont des données recueillies en parallèle d'un dispositif de collecte, et qui en décrivent le processus, là où les métadonnées fournissent des informations contextuelles sur la manière dont les données ont été collectées. Les horodatages peuvent révéler des réponses données trop rapidement, suggérant un manque de réflexion ou une tentative de compléter l'enquête de manière expéditive. Le moyen de réponse au questionnaire Cawi, collecté via les métadonnées, peut également offrir des informations pertinentes.

Les jeux de données à notre disposition contiennent 8515 et 8159 réponses pour respectivement 2024 et 2025. Ils comprennent les valeurs recueillies via les questionnaires Cawi, les métadonnées et paradonnées.

Les données socio-démographiques sont récupérées lors de la première phase de collecte, ces données ne sont donc pas utilisées dans la détection, cependant elles seront utiles pour distinguer des types de foyers. On retrouve parmi elles, l'âge du répondant au questionnaire, son activité professionnelle, le nombre de personnes au sein du foyer, leurs âges ou encore leurs activités professionnelles.

Concernant les données d'équipement, elles représentent l'ensemble des équipements du foyer qui ont été déclarés. On retrouve un très grand nombre de variables, des plus communes, comme le nombre de télévisions, de téléphones fixes ou mobiles à des questions très spécifiques, comme la présence de certains types de box internet ou de modèles spécifiques de TV connectée. Les paradonnées et métadonnées, récupérées de manière automatiques lors du remplissage du Cawi, renseignent l'heure ou les heures de connexion questionnaire, les horodatages exacts des réponses aux modules, l'appareil utilisé pour répondre (e.g. smartphone, ordinateur). De cette manière, nous disposons des temps de réponses propres à chacun des modules. Cette information sera l'un de nos leviers majeurs pour la détection des répondants négligents.

3 Méthodologies de détection

De nombreux modèles non supervisés ont été testés dans la littérature. Des chercheurs (Najeeb Jebreel, 2020) utilisent un jeu de données où des valeurs atypiques sont simulées, cela leur permet d'évaluer la performance de différents modèles tels que l'Isolation Forest (IF) et le Density Based Spatial Clustering of Applications with Noise (DBSCAN).

Cette approche permet de tester et de valider les performances des modèles dans un contexte contrôlé, fournissant ainsi des indications sur leurs transposition à des ensembles de données réelles.

À travers ce contrôle, une méthode hybride a été développée (Anvia Anjay Mahajan, 2023), des modèles tels que l'IF ou Local Outlier Factor (LOF) sont utilisés. La combinaison des résultats des deux modèles permet d'atténuer les limitations respectives de ces mêmes modèles considérés en isolation dans le cadre de la détection d'anomalies ou valeurs extrêmes. Dans ces méthodes les durées de remplissage de questionnaire ne sont pas utilisées. Une approche basée uniquement sur ces dernières, dérivée de la littérature (Laura Gamble, 2023), transforme les durées de chaque module en indicateur auquel des pondérations sont appliquées pour être traduit en critère de réponse dite falsifiée.

Cette section sera donc dédiée à ces deux méthodes, dont les résultats ont été combinés pour obtenir nos répondants négligents finaux.

3.1 Méthode 1 : Durées de remplissage

Les durées de remplissages doivent être manipulées avec précaution, en effet, il faut faire face aux temps anormalement longs causés, entre autres par le fait que certains répondants peuvent laisser le questionnaire ouvert sur leur appareil, puis le reprendre plus tard, et ce sans qu'il n'y ait de déconnexion. Quand il n'y a pas de déconnexion, nous ne disposons pas d'information si une pause a eu lieu ou non. Une des conséquences est que, dans la détection d'anomalies, ces comportements peuvent ressortir voire invisibiliser les autres, alors que ce ne sont pas forcément ceux que l'on souhaite identifier comme négligents. Afin de pallier ce phénomène, nous utilisons comme valeur de durée son inverse. Soit $x_{i,j}$ la durée de remplissage en seconde de l'individu j pour le module i alors on définit $v_{i,j} = 1/x_{i,j}$.

En pratique, certains temps sont très courts (seulement quelques secondes), cela est dû à la nature particulièrement modulable du questionnaire. Sa longueur est directement corrélée au nombre de personnes et équipements media présents dans le foyer. De fait, pour chaque module, nous pondérons la durée de passation (inverse) par le nombre d'équipements déclarés sur le module en question. Nous associons donc à chaque individu j la valeur de $v_{i,j}$ centrée et réduite par rapport à son nombre d'équipement k et mis au carrée sur le module i. Ceci correspond à la valeur du χ^2_{ik} où

$$\chi^2_{ik} = \left(\frac{v_{i,j,k} - \overline{v_{ik}}}{\sigma_{ik}} \right)^2$$

avec $v_{i,j,k}$ l'inverse de la durée de passation de l'individu j sur le module i ayant déclaré k équipements sur ce même module, $\overline{v_{ik}}$ la moyenne de l'inverse des durées sur le module i des répondants ayant déclarés k équipements et σ_{ik} la variance de l'inverse des durées sur le module i parmi ceux ayant déclarés k équipements. Cette valeur est calculée pour tout k de chaque module i.

Avant de construire le score de "négligence" final sur les durées, une pondération sur les durées médianes de remplissages est également effectuée. Ce, dans le but d'attribuer, à chaque module, l'importance adéquate, en effet, les modules ne sont pas égaux en taille. Nous la définissons comme suit :

$$P_i = \frac{med\{x_{i1},x_{i2},...,x_{in}\}}{\sum_{i=1}^{7} med\{x_{i1},...,x_{in}\}}$$

où P_i le poids du module i, $x_{i,j}$ la durée de remplissage en seconde de l'individu j pour le module i, et $med\{x_{i1},x_{i2},...,x_{in}\}$ la médiane des durées de remplissage du module i pour les individus allant de 1 à n, enfin, le dénominateur est la somme des médianes de nos 7 modules. Le score final est donc une approximation simplifiée utilisant la somme pondérée des variables indépendantes suivant une loi du χ^2. Soit, pour chaque individu (différencié par leurs nombres d'équipements k), la somme $\sum_{i=1}^{7} P_i * \chi^2_{ik}$ est déterminée. Le caractère négligent ou non d'un individu j est statué à l'aide de

la comparaison des valeurs obtenues au score à un quantile du χ^2 de degré 1. Cela nous permet d'être flexible quant au seuil, en effet, des p-valeurs < 0.05 ou < 0.01 peuvent être choisies à la convenance du décideur.

Les individus associés aux p-valeurs inférieures au seuil fixé seront considérés comme suspects au regard de leurs durées de remplissage, conformément aux nombres d'équipements déclarés.

3.2 Méthode 2 : Déclaration d'équipements

Cette deuxième méthode fait usage des variables socio-démographiques en pré-requis de comparaisons des déclarations d'équipements. Ces données sont récoltées lors de la première phase téléphonique de l'étude. Comme mentionné précédemment, l'équipement des ménages dépend fortement des individus qui les composent. En effet, au deuxième semestre 2024, Médiamétrie estime que le nombre moyen d'écrans change en fonction du profil de la personne de référence (PDR) du foyer. Lorsque cette dernière se situe dans la tranche d'âge 35-49 ans, le nombre d'écrans moyen est de 8.1 alors que si elle est plutôt dans la tranche d'âge 65 ans et plus, ce chiffre chute à 5.1.

Un algorithme de classification non supervisée classique (K-Means), est notre première étape. Le but étant de regrouper des foyers qui devraient avoir des niveaux et types d'équipements similaires. Les variables suivantes ont été incluses dans le modèle : le nombre de personne dans le foyer, la catégorie socio-professionnelle (CSP) de la PDR, le nombre de personnes CSP-[1] du foyer, le nombre de personnes CSP+[2] du foyer, le nombre de retraités dans le foyer, le nombre de personnes inactives[3] dans le foyer, le nombre de personnes entre 0 et 15 ans, entre 16 et 24 ans, entre 35 et 49 ans, au delà de 50 ans au sein du foyer, l'âge du répondant à l'enquête ainsi que l'âge de la PDR. Dans notre cas, cette étape a abouti à quatre clusters distincts. Ce choix s'est fait à l'aide des méthodes usuelles, notamment la somme des variances intra-clusters (selon les différents nombre de clusters) et l'analyse des clusters résultants.

Une fois ces clusters définis, on leur applique deux modèles de machine learning (ML) en parallèle d'après une méthode hybride (Anvita Anjay Mahajan, 2023). Nous avons sélectionné les deux modèles DBSCAN et IF pour la différence de leurs caractéristiques.

DBSCAN repose sur la détection de zones dites de "haute densité" dans un ensemble de données. Les observations classifiées comme étant du bruit, car se situant dans des zones très peu denses voire sans voisin, sont les anomalies (ou *outliers*). Le paramètre du nombre minimum de voisins nécessaire à DBSCAN a été choisi conformément à la littérature (Martin Ester, 1996), soit $k = 4$. Le paramètre de la taille du rayon, qui correspond à la distance maximum au-delà de laquelle les points ne sont plus considérés comme voisins, est déterminé à l'aide du tracé des distances des k-plus proches voisins (ici, 4-plus proches voisins). Soit dans notre cas, 0.477 au premier semestre de 2025 et 0.445 au second semestre 2024.

Le second modèle sélectionné (IF), contrairement à d'autres méthodes, isole les points de données en construisant des arbres de décision de manière aléatoire. Cette méthode est plus efficace dans la détection d'anomalies internes qu'externes, autrement dit, lorsque les données présentent des îlots d'observations, la détection sera plus efficace sur les observations dissidente au sein de chaque îlot, plutôt qu'une observation extrême à plusieurs îlots (Najeeb Jebreel, 2020). Ce phénomène est dû au fait que des scores sont attribués en fonction du nombre de décisions nécessaires pour isoler une

1. Inclus les agriculteurs, employés, ouvriers qualifiés.

2. Inclus les artisans, commerçants, cadres, professions intellectuelles supérieures, chefs d'entreprise, professions intermédiaires.

3. inclus les écoliers, étudiants et chômeurs.

observation.

Ici aussi, les paramètres du modèle ont été défini conformément à la littérature (Yousra Chabchoub, 2022), le nombre d'arbre de décision à générer est fixé à 100 et la taille de l'échantillon (pour chaque arbre) à 256, et ce pour les deux semestres de notre étude. Après application du modèle, chaque observation se voit attribuer un score. Ici aussi, un seuil doit être fixé pour isoler les anomalies des observations "normales". Nous avons utilisé, conjointement, la fonction de distribution empirique cumulative des scores ainsi que le paramètre de forme[4]. Ce dernier caractérise la nature et la forme des queues d'une distribution de probabilité. Finalement, les seuils ont été fixés à 0.667 et 0.663 pour, respectivement, 2024 semestre deux et 2025 semestre un.

Les variables natives suivantes ont été prises en compte pour les deux modèles de ML : le nombre de TV, de consoles de jeux portable, de consoles de jeux associées à une TV, d'ordinateurs, de tablettes, de numéro de téléphone fixe, le nombre de téléphones mobiles, le nombre de fournisseurs d'accès à internet (FAI).

Deux variables supplémentaires, construites, ont également été ajoutées. À partir de la première question de chaque module qui consiste à recueillir le nombre total d'un équipement donné (posé à tous les répondants), nous calculons un indicateur « Ne Sait Pas » (NSP), qui est le ratio du nb de de réponse NSP sur le nombre de questions où cet item est proposé.

La seconde variable permet d'appréhender la sous déclaration du nombre de téléphone mobile des individus du foyer. Elle est calculée en faisant la soustraction du nombre de personnes du foyer déclarées comme possédant un mobile avec le nombre de téléphones mobiles total du foyer déclaré en tout début du module. Ce module étant l'avant dernier du questionnaire, nous considérons cet indice comme révélateur de lassitude et de compréhension de la mécanique du questionnaire. En effet, si l'enquêté comprend la structure du questionnaire au fur et à mesure du remplissage, cette sous déclaration au niveau de l'équipement individuel de téléphone mobile est un moyen de réduire la longueur totale du questionnaire.
Chaque observation considérée comme une anomalie par l'une des deux approches sera classifiée comme une réponse négligente. Ce choix a été fait de part le caractère complémentaire des deux modèles qui n'ont pas du tout les mêmes mécaniques et donc ne ciblent pas les mêmes comportements ou profils. Cette deuxième méthode basée sur les équipements est résumée par le schéma 2.

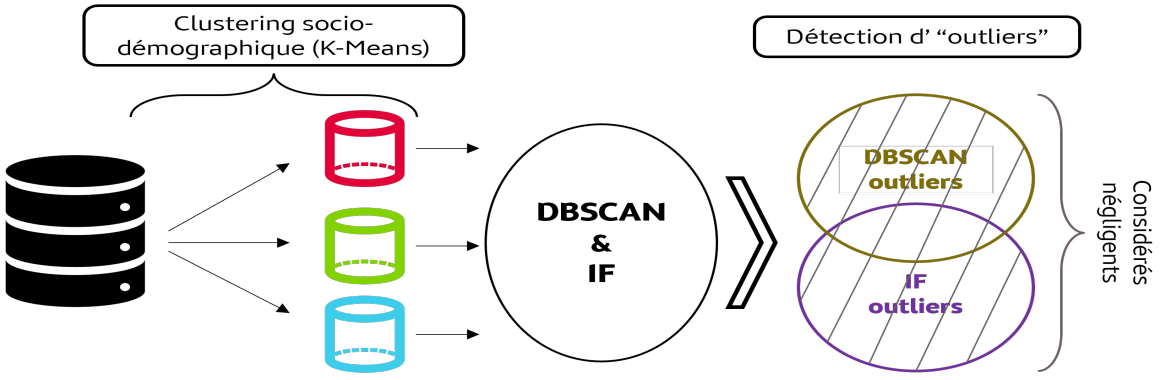


FIG. 2 – *Schéma du processus de notre deuxième méthode de détection de répondants négligents*

Finalement, nous combinons les résultats des méthodes une et deux, donc de 3 approches différentes

4. *Shape Parameter* en anglais.

au total. Le fait d'avoir trois seuils et de multiples paramètres sur lesquels jouer pour être restrictif ou non est un avantage certain en terme de flexibilité.

4 Résultats empiriques

Les différentes méthodes présentées ont permis la détection de répondants dits négligents à l'égard de plusieurs points de vue. Sur les deux semestres auxquels nous avons appliqués la méthode, 652 répondants ont été identifiés comme tels pour des jeux composés de 16 674 répondants au global. Parmi ces 652, 380 proviennent du deuxième semestre 2024 sur une base de 8515 répondants, soit 3.76% du total. Les 332 autres proviennent du premier semestre 2025 pour une base de 8159 répondants, soit 4.07% du total.

4.1 Un an d'historique

Le tableau 1 détaille la provenance des détectés par méthode et par semestre. Il est important de remarquer que certains répondants sont détectés par plusieurs, voire toutes les méthodes. Par ailleurs, notez que le total ne contient aucun doublon.

Méthode	**Semestre 1 2024**	**Semestre 2 2025**	**Total**
Clustering + DBSCAN	172	188	360
Clustering + IF	72	105	177
Union DBSCAN + IF	198	228	426
Méthode des Temps	126	107	233
Total	320	332	**652**

TAB. 1 – *Nombre de détectés par méthode et semestre*

Les résultats empiriques couplés à l'analyse statistique et métier permettant d'identifier les profils des répondants négligents. Mais également de tirer des conclusions et enseignements si l'on considère cette donnée comme un indicateur d'évolution de questionnaire ou de pratiques passées et futures.

Les deux approches, bien que combinées, sont très différentes dans la manière dont elles ont été pensées et quels comportements elles sont supposées reconnaître. De ce fait, nous allons, dans un premier temps les analyser séparément.

La méthode 1, basée sur la durée de remplissage de questionnaires, vise à isoler les répondants ayant des temps de réponse aux différents modules anormaux par rapport au nombre d'équipements qu'ils déclarent sur ce même module. Les figure 5,8 & 9 décrivent la distribution des durées de remplissage, en comparant la population des répondants détectés à sa complémentaire.
On remarque, d'après les deux histogrammes, qui concernent les modules mobiles et consoles de jeux (portables et sur télévision), que les distributions des détectés est sensiblement décalé vers la gauche. Autrement dit, un décalage vers des temps plus bas. Par ailleurs, plusieurs pics de concentrations sont visibles ici. Ceci nous conforte dans l'idée que les répondants négligents ont des temps de réponse intra-module plus erratiques que les autres, d'autant plus que c'est un phénomène qui est stable sur les deux périodes d'intérêt étudiées. La figure9 récapitule les temps médians de remplissage par module dans l'ordre d'apparition dans le questionnaire. On peut constater que les répondants détectés sont systématiquement notablement inférieurs à la population complémentaire et la population totale. Les différences les plus marquées se trouvent sur les modules de télévision et

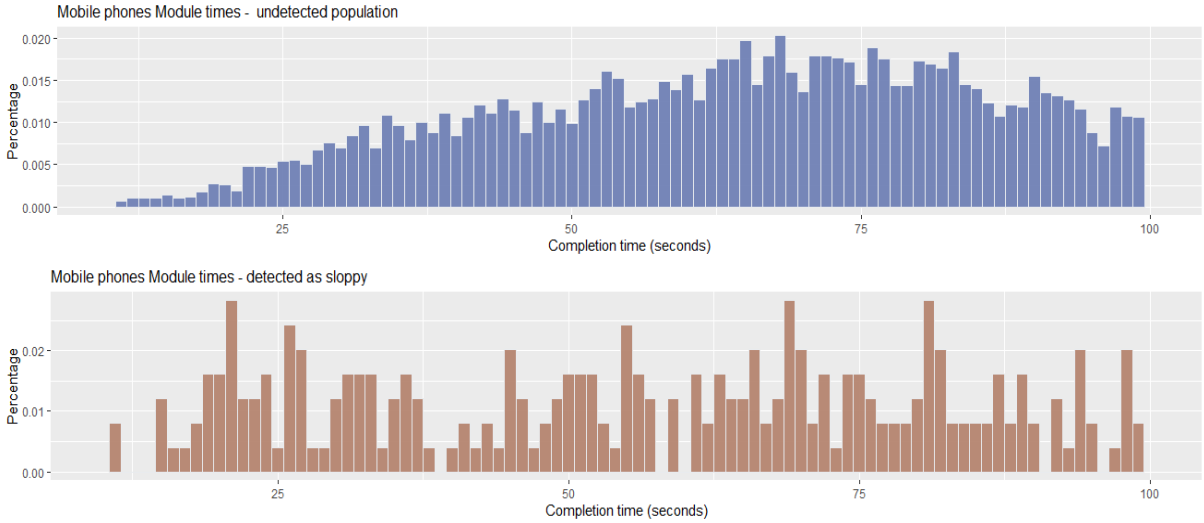


FIG. 3 – *Semestre 2 2024*

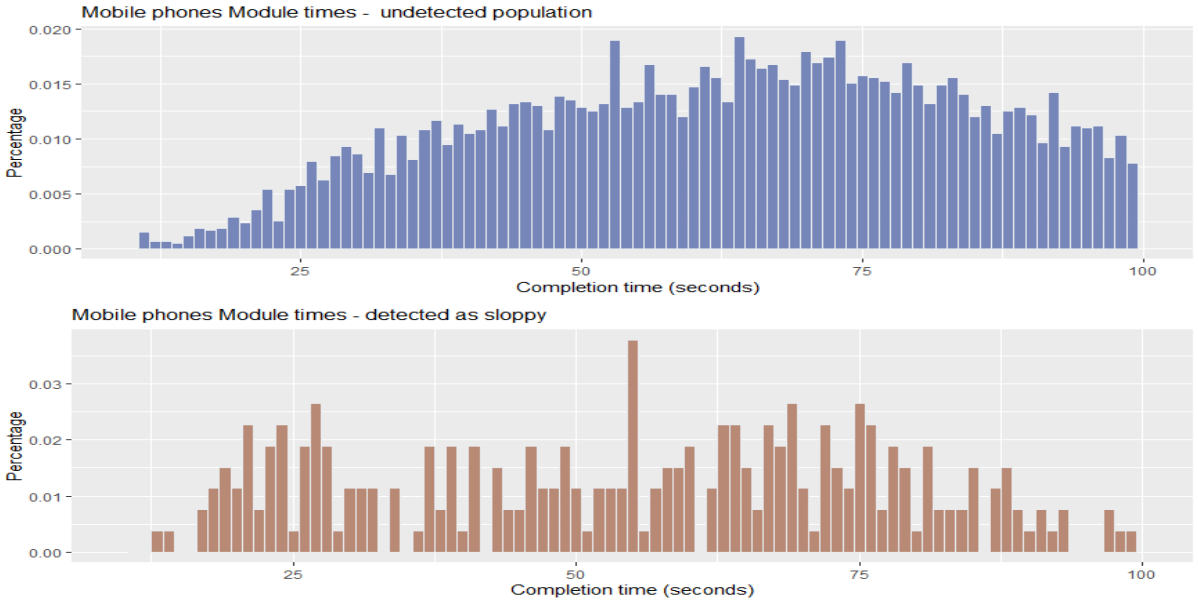


FIG. 4 – *Semestre 1 2025*

FIG. 5 – *Distribution des durées de remplissage du module de téléphonie mobile par type de population et par semestre*

console de jeux (portable et télévision). Il est important de noter que le module TV est le plus long en terme de nombre de questions posées, mais aussi car il comprend une boucle avec un ensemble de questions qui est répété pour chaque télévision déclarée. Une mécanique facilement identifiable et pourrait amenée un enquêté avec plusieurs télévision à aller trop vite sur ce module.

A l'instar de précédemment, les deux semestre donnent les mêmes conclusions, bien que certains écarts soient moins marqués sur le semestre 1 2025.

On peut maintenant tourner notre attention vers les discriminations de la deuxième méthode, basée sur les déclarations d'équipements tout en intégrant la notion de composition de foyers. Les tableaux 2 et 3 font l'état des lieux des variables utilisées dans la deuxième phase de la méthode ainsi que des variables critiques de composition de foyer.

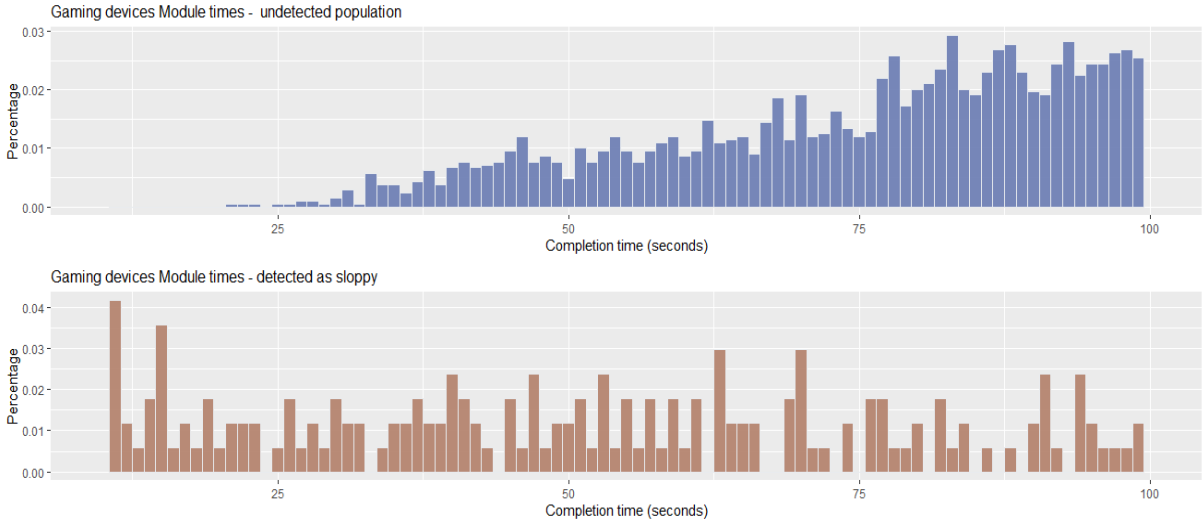


FIG. 6 – *Semestre 2 2024*

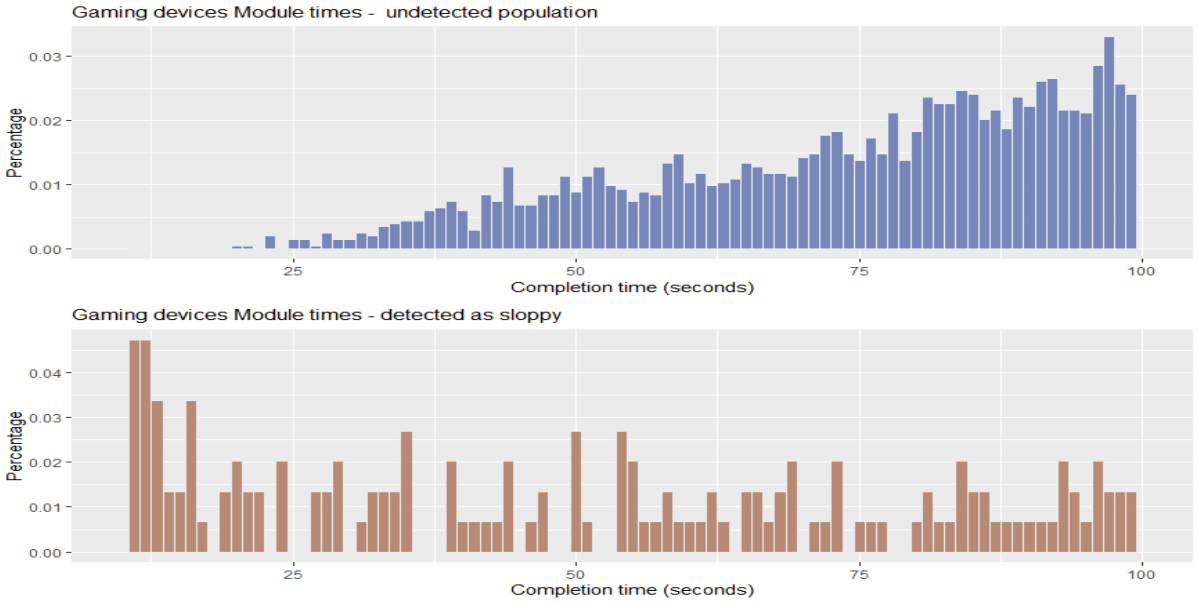
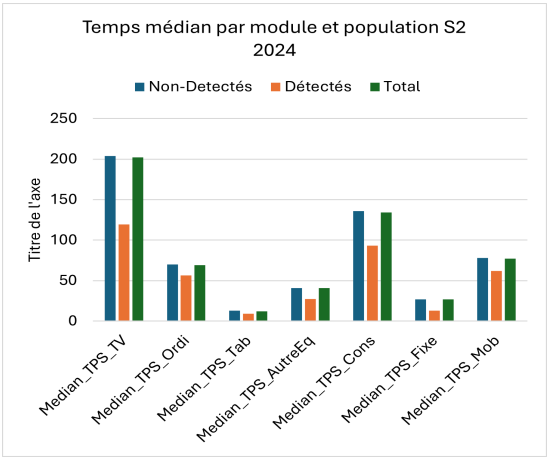


FIG. 7 – *Semestre 1 2025*

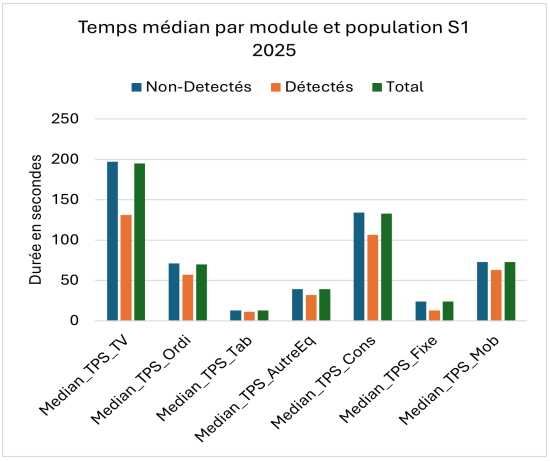
FIG. 8 – *Distribution des durées de remplissage du module de console de jeux par type de population et par semestre*

Concernant les variables utilisées pour les modèles DBSCAN & IF, la population des détectés possède de manière quasi-systématique sensiblement plus ou sensiblement moins d'équipements. C'est le cas pour les nombres moyens de TV, consoles TV, d'ordinateurs, de tablettes, entre autres, mais ils possèdent bien plus de téléphones fixes, et ont un indice de déclaration de téléphones mobiles tronquée plus élevé.

Notons que le phénomène est plus marqué au deuxième semestre de l'année 2024 par rapport au premier semestre 2025. Si on s'intéresse aux variables de composition de foyer, l'élément marquant est l'âge du répondant à l'enquête qui est en moyenne 6 ans plus jeune chez les détectés en 2024, et près de 10 ans au premier semestre 2025. Un second élément également clivant est la part de foyers comportant un seul membre au semestre 2 de l'année 2024, il y avait un écart de 9 points de pourcentage entre les détectés et les autres, et environ 12 points au premier semestre 2025.



(a) *Semestre 2 2024*



(b) *Semestre 1 2025*

FIG. 9 – *Repartition des temps médians de remplissage des modules par population*

	Semestre 1 2025		Semestre 2 2024	
Moyennes des variables	**Non Détectés**	**Détectés**	**Non Détectés**	**Détectés**
Ratio “Ne Sait Pas”	0,009	0,007	0,008	0,007
Nombre de TV	1,510	1,238	1,511	1,284
Nombre de console TV	0,303	0,289	0,261	0,303
Nombre de console portable	0,666	0,515	0,628	0,656
Nombre d’ordinateur	2,065	1,964	1,981	1,803
Nombre de tablette	0,753	0,639	0,723	0,562
Nombre de telephone fixe	1,868	2,377	1,781	2,319
Nombre de mobile	2,327	2,096	2,324	2,181
Declaration Mobile dévaluée	0,210	0,256	0,226	0,519

TAB. 2 – *Valeurs moyennes des populations par semestre sur les variables utilisées dans la modelisation DBSCAN et IF*

Variables	**Semestre 1 2025**		**Semestre 2 2024**	
Composition de Foyer	**Non Détectés**	**Détectés**	**Non Détectés**	**Détectés**
% moyen de 16 - 24 ans dans le foyer	10,9%	15,2%	10,8%	21,3%
% moyen de 25 - 34 ans dans le foyer	11,9%	18,0%	10,8%	19,4%
% moyen de 65+ dans le foyer	26,4%	18,9%	26,7%	14,2%
Age moyen du repondant	52,002	45,846	52,038	42,497
% de PDR - de 35 ans	13,7%	26,8%	13,5%	33,1%
% de PDR entre 35 et 50 ans	24,3%	26,5%	24,4%	23,8%
% de PDR 50 ans et +	61,9%	46,7%	62,2%	43,1%
% de NPF 1	25,3%	37,0%	25,8%	34,7%

TAB. 3 – *Compositon du foyer par population et semestre*

La part des PDR âgées de 35 ans ou moins est également beaucoup plus importante (environ le double) chez les détectés et ce sur les deux semestres. De ces informations, nous pouvons déduire que, sur ces deux jeux de données, les répondants identifiés comme négligents sont plus jeunes que la population globale, résident dans des foyers avec peu de membres, et s’il y en a plusieurs, les

membres appartiennent à des tranches d'âge jeunes (16-24 ans, 25-34 ans).
Effectivement, si l'on analyse maintenant les clusters créés en amont de la modélisation pour la méthode 2, nous faisons le même constat. Le tableau 4 donne les composants socio-démographiques majeurs par cluster. Le tableau 5 énumère quant à lui le nombre de cas détectés par cluster.

2024 Semestre 2								
CLUSTER	**N**	**NPF Moyen**	**Age Moyen Répondant**	**PDR CSP+**	**PDR CSP-**	**PDR Retraité**	**Détectés**	**% Détectés**
1	2411	3,218	41,777	0,634	0,321	0,025	89	3,69%
2	1551	2,100	27,83	0,580	0,266	0	123	7,93%
3	1989	1,753	73,626	0,045	0,007	0,937	39	1,96%
4	2564	2,265	58,393	0,543	0,247	0,183	69	2,69%
2025 Semestre 1								
CLUSTER	**N**	**NPF Moyen**	**Age Moyen Répondant**	**PDR CSP+**	**PDR CSP-**	**PDR Retraité**	**Détectés**	**% Détectés**
1	1471	1,981	28,345	0,589	0,257	0	111	7,55%
2	2608	2,301	58,886	0,564	0,217	0,198	80	3,07%
3	2337	3,252	41,365	0,678	0,276	0,021	90	3,85%
4	1743	1,748	74,757	0,04	0,007	0,944	51	2,93%

TAB. 4 – *Variables de composition de foyer par cluster et semestre*

2024 Semestre 2 - Pourcentages Moyens								
CLUSTER	**N**	**0-15 ans dans Foyer**	**16-24 ans dans Foyer**	**25-34 ans dans Foyer**	**35-49 ans dans Foyer**	**50-64 ans dans Foyer**	**65ans et plus**	**NPF de 1**
1	2411	24,1%	12,0%	3,6%	46,8%	12,5%	1,0%	15,6%
2	1551	11,2%	27,2%	50,4%	10,7%	0,5%	0,1%	41,1%
3	1989	0,1%	0,4%	0,6%	0,8%	3,5%	94,6%	30,3%
4	2564	3,3%	9,1%	2,6%	2,5%	69,7%	12,7%	23,9%
2025 Semestre 1 - Pourcentages Moyens								
CLUSTER	**N**	**0-15 ans dans Foyer**	**16-24 ans dans Foyer**	**25-34 ans dans Foyer**	**35-49 ans dans Foyer**	**50-64 ans dans Foyer**	**65ans et plus**	**NPF de 1**
1	1471	9,2%	23,6%	55,3%	11,4%	0,3%	0,1%	44,1%
2	2608	3,0%	10,0%	3,4%	2,7%	64,3%	16,5%	23,7%
3	2337	24,2%	12,5%	3,6%	46,2%	12,2%	1,2%	14,2%
4	1743	0,1%	0,2%	0,4%	0,7%	3,1%	95,6%	29,0%

TAB. 5 – *Variables de composition de foyer par cluster et semestre*

La proportion la plus élevée de répondants détectés comme négligents (avec environ 8% de détectés (4)) se trouve dans les clusters contenant des foyers avec les mêmes profils socio-démographiques. Ce sont des individus jeunes, âgés en moyenne de 28 ans, avec une majorité de PDR CSP+ (aucun retraité), et de foyers composés d'un seul individu dans environ 40% des cas.
En comparaison, les autres clusters, indifféremment du semestre, produisent entre 2 et 3% de répondants négligents.

4.2 Évaluation d'évolution

La détection de répondants négligents peut également servir d'indicateurs dans le cas d'évolutions de l'enquête sur laquelle est appliquée la méthode. En effet, cette étude oriente nos travaux dans le cas de changements dans les questionnaires, ou dans l'adaptation du questionnaire selon le type d'écrans (tablette, ordinateur ou smartphone). Les profils des répondants négligents, leurs proportions par rapport au niveau de restriction qui est appliqué dans les seuils ou encore le medium utilisé pour répondre (s'il est discriminant) sont tous des éléments qui peuvent être révélateurs d'une amélioration nécessaire ou d'une démonstration de la valeur probante d'une évolution.

Dans notre cas, entre le second semestre 2024 et le premier semestre 2025, une refonte du module de téléphonie mobile a été implémentée.
Dans le cadre de cette étude, nous avons pu en apprécier les effets sur les répondants négligents détectés grâce à la variable de déclaration tronquée du nombre de mobile.

Au second semestre de 2024, qui intégrait partiellement cette refonte du module, la valeur de cet indicateur était plus de deux fois supérieur pour les individus détectés par rapport aux autres. Au premier semestre 2025, bien qu'elle reste plus élevée, la différence est minime, cela signifie que cette variable n'est plus aussi discriminante pour la négligence, cf. Tab2.

	Semestre 1 2025		**Semestre 2 2024**	
Appareil utilisé pour répondre à l'enquête	Non Détectés	Détectés	Non Détectés	Détectés
% Ordinateur	47,4%	44,0%	47,7%	36,9%
% Smartphone	52,2%	55,4%	51,8%	62,8%
% Tablette	0,4%	0,6%	0,5%	0,3%

TAB. 6 – *Medium de réponse à l'enquête par population et par semestre*

De la même manière, d'après les paradonnées, l'étude du semestre 2 de l'année 2024 pouvait suggérer qu'une amélioration était à faire sur le design du questionnaire lors du remplissage via un smartphone (Tab6), or les résultats du semestre suivant ne vont pas dans ce sens. Cette conclusion était donc peut-être ponctuelle et non démonstrative d'un problème systémique.

5 Conclusions & Perspectives

Nous avons exposé dans cet article les méthodes utilisées pour parvenir à identifier les répondants négligents à une enquête en ligne. Au total, 3 types d'approches ont été définis, chacune visant à détecter des répondants négligeants de manière différente mais complémentaire. Avec deux jeux de données de près de 16 700 répondants au total, environ 650 ont été identifiés comme négligents, soit près de 4% de la base totale. Les répondants détectés, bien que différents sur quelques points précis, sont sensiblement les mêmes sur les deux semestres étudiés, c'est-à-dire, des foyers jeunes, actifs, peu nombreux.

Les seuils de décision utilisés pour les deux semestres ont été choisis selon les mêmes critères et la même volonté de restriction pour avoir des résultats qui soient les plus comparables possibles. Cela dit, une modulation reste toujours possible en dehors du cadre d'une étude. De fait, déterminer ces seuils selon des critères différents peut être pertinent selon le type de négligence que l'on souhaite viser.

une modulation reste toujours possible en dehors du cadre d'une étude. De fait, selon l'usage des répondants négligents qui est planifié, faire reposer ces seuils sur des critères différents est tout aussi pertinent.

Dans le cas de notre étude, La finalité de cette détection est pour le moment limitée à une finalité d'évaluation de la qualité et l'identification d'améliorations du design du questionnaire. Une évolution sur le fond et la forme du questionnaire (module téléphonie mobile), a eu lieu entre le second semestre 2024 et le premier semestre de 2025; le résultat était probant : les détectés ont drastiquement diminué (en proportion) sur la variable mesurant au plus proche cette évolution. Cette analyse a servi de démonstration de la qualité et utilité de cette évolution. C'est donc un appui supplémentaire pour démontrer la qualité de l'évolution.

Dans le futur, cette approche pourrait être mise en production, c'est à dire être lancée automatiquement à l'issue de chaque semestre où le baromètre des équipements est produit. Nous pourrions également l'appliquer en cours de production, à certains points déterminés, dans le but d'ajuster les quotas des profils demandés selon la quantité et les profils de répondants négligents détectés. Par

ailleurs, elle pourrait également être réalisée dès qu'une évolution de l'enquête (questionnaire ou outils) doit être testée.

Bibliographie

[1] Groves, Robert M., Survey errors and survey costs, *John Wiley & Sons*, 2005.
[2] Gamble, Laura. Evaluation of Key Performance Indicators for Interviewer Falsification Suspicion from Paradata and Interview Data, *Ottawa 2023*.
[3] Mahajan, Anvita Anjay. Hybrid Model using LOF and iForest Algorithms for Detection of Insider Threats, *Authorea Preprints*, 2023.
[4] Jebreel, Najeeb. Detecting Bad Answers in Survey Data Through Unsupervised Machine Learning, 2020.
[5] Ester, Martin. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96). AAAI Press, 226–231, 1996.
[6] Chabchoub, Yousra. "An In-Depth Study and Improvement of Isolation Forest," in IEEE Access, vol. 10, pp. 10219-10237, 2022.