

LE MONTANT DU PRÉJUDICE DÛ AUX ESCROQUERIES ET AUX FRAUDES AUX MOYENS DE PAIEMENT

Laurent DUVERNET (*)

(*) *Ministère de l'Intérieur, Service statistique de la Sécurité intérieure (lors de la réalisation de l'étude) et Insee, Département des Méthodes statistiques*

laurent.duvernet@insee.fr

Mots-clés : escroqueries, délinquance économique, non-réponse, asymétrie.

Domaines : correction de la non-réponse, intégration des données.

Résumé

Ce travail présente un premier chiffrage du préjudice dû aux escroqueries et aux fraudes aux moyens de paiement en France métropolitaine entre 2016 et 2023. Il s'appuie sur un recensement exhaustif des infractions dont ont eu connaissance les forces de police et de gendarmerie, ainsi que sur une estimation du préjudice subi par les personnes physiques et non déclaré aux forces de l'ordre, estimation obtenue grâce à une enquête de la statistique publique, Cadre de vie et sécurité (CVS - Insee-ONDRP-SSMSI).

Selon nos évaluations, le préjudice dû aux escroqueries et aux fraudes aux moyens de paiement est en augmentation rapide. Entre 2016 et 2023, il double presque pour les personnes physiques, passant de 2,3 milliards d'euros à 4,5 milliards d'euros. Concernant les personnes morales, le préjudice des atteintes qu'elles ont subies et qui a donné lieu à dépôt de plainte oscille sur la même période. Il est proche de 820 millions d'euros en 2016 et de 670 millions d'euros en 2023.

Ces évaluations sont entachées d'incertitude, du fait que le montant n'est disponible que pour une partie des préjudices déclarés aux forces de police et de gendarmerie. Les données d'enquête quant à elles ne peuvent renseigner sur les préjudices élevés les plus rares, qui comptent pour une part non négligeable du total. Il en résulte qu'il est particulièrement ardu de fournir un intervalle de confiance rigoureux concernant le chiffrage obtenu. Une annexe présente une brève caractérisation du lien entre l'asymétrie des données et le domaine de validité des intervalles de confiance classique en correction de la non-réponse.

Abstract

This study presents a first evaluation of the total financial losses due to scams and bank fraud in metropolitan France between 2016 and 2023. It relies on a comprehensive record of all offenses reported to police and gendarmerie forces, as well as on an estimate on the losses suffered by individuals who did not report the offense. This estimate was obtained from a public statistics survey, Cadre de vie et sécurité (CVS – Insee–ONDRP–SSMSI).

According to our results, the financial losses due to scams and bank fraud are increasing rapidly for individuals. They almost double between 2016 and 2023, rising from €2.3 billion to €4.5 billion. For firms and other legal entities, the total financial losses due to offenses that were reported to law enforcement fluctuate over the same period, amounting to around €820 million in 2016 and €670 million in 2023.

These estimates are subject to uncertainty, as the amount of financial loss is available only for some of the cases reported to the police and gendarmerie. Moreover, survey data cannot account for the highest and rarest losses, which represent a non-negligible share of the total. As a result, it is particularly difficult to provide a rigorous confidence interval for the estimated figures. In the appendix, we provide a short discussion of the link between the asymmetry of data and the validity range for the usual confidence intervals when dealing with nonresponse.

1 Introduction

Le chiffrage du montant économique de l'activité criminelle est un sujet encore peu étudié en France, et limité à des domaines circonscrits. Un rapport commandé par la MILDECA (Mission interministérielle de lutte contre les drogues et les conduites addictives) évalue le marché français de la drogue en 2016 à 2,3 milliards d'euros [16]. La fraude fiscale et sociale a fait l'objet de quelques évaluations partielles : ainsi la Caisse nationale des allocations familiales chiffre également à 2,3 milliards d'euros la fraude aux prestations familiales en 2018 [6], et l'Insee estime que la fraude à la TVA due par les entreprises en 2012 représente un montant compris entre 20 et 25 milliards d'euros [21]. À notre connaissance, les escroqueries et les fraudes aux moyens de paiement n'ont quant à eux pas encore fait l'objet d'un chiffrage global en France, à l'exception du rapport de l'Observatoire de la sécurité des moyens de paiement qui présente une évaluation annuelle du montant des fraudes aux moyens de paiement dont les établissements bancaires ont connaissance [19] (1,2 milliards d'euros en 2022).

La littérature internationale propose des études légèrement plus nombreuses sur le sujet, études qui mettent en évidence des problèmes méthodologiques récurrents [17]. En premier lieu, il est difficile de délimiter le champ infractionnel qu'on veut étudier : par exemple le même terme de *fraud* en anglais recouvre des périmètres variables selon les études qui y font figurer ou non les fraudes aux moyens de paiement (*bank fraud*, *payment card fraud*), la fraude à l'encontre de la collectivité (*tax fraud*), certaines formes de cybercriminalité (*computer misuse*), etc. Par ailleurs, le propre de ces comportements est que leur auteur cherche à tromper la victime, laquelle sert en général de source première pour constituer les données alors même qu'il se peut que son interprétation de la situation soit faussée par la tromperie potentielle : la victime peut ainsi sous-déclarer par rapport aux faits réellement commis car elle n'a pas détecté l'escroquerie, même a posteriori, mais aussi sur-déclarer, dans le cas par exemple où elle qualifie d'escroquerie ce qui ne serait interprété par quelqu'un d'autre que comme un rapport qualité/prix décevant lors d'une transaction commerciale [7]. La question même de savoir qui est la victime de l'infraction n'est pas non plus toujours évidente, dans la mesure où des mécanismes d'assurance ou de remboursement existent pour certains types de fraudes, en particulier concernant le remboursement par l'établissement bancaire de nombreux cas de fraudes aux moyens de paiement [17]. Enfin, si on s'intéresse non simplement aux nombres d'infractions commises, mais également aux montants des préjudices subis, on court le risque d'abord que les victimes minimisent leurs pertes, par exemple par peur du ridicule qu'il peut y avoir à « être tombé dans le panneau » et s'être fait arnaquer d'une somme importante, mais aussi qu'elles grossissent leurs pertes, par exemple dans l'espoir d'obtenir un remboursement plus important, ou encore tout simplement qu'elles se rappellent mal les montants réels plusieurs mois après les faits.

Afin de produire des réponses plus fiables, on peut chercher à décrire précisément à la personne qui se signale comme victime les différents modes opératoires possibles (escroquerie sentimentale,

fraude à l'investissement, chèque en bois...), ce qui a également l'avantage de fournir une mesure fine de la prévalence de ces différents types. La *Federal trade commission* publie ainsi un rapport annuel sur le nombre et le montant des escroqueries signalées aux autorités des États-Unis et ventilées selon plusieurs dizaines de catégories [4]. Cependant, outre le risque de mal classer des infractions qui relèveraient soit d'aucun des types proposés, soit de plusieurs à la fois, les effets de mobiliser une catégorisation détaillée des différentes escroqueries ne sont pas très clairs, ce que deux exemples illustrent. Dans [15], les auteurs mentionnent comment en 2016, le passage dans l'enquête nationale de victimation suédoise d'une question unique sur les escroqueries à une différenciation entre escroqueries à la consommation et escroqueries bancaires a provoqué un doublement de la prévalence. À l'inverse, le *Bureau of Justice Statistics* des États-Unis a produit en 2017 un supplément à son enquête de victimation consacré à la mesure des escroqueries dans lequel sept catégories précises d'escroqueries sont décrites en détail aux enquêtés. Ce protocole a abouti à la mesure d'une prévalence globale de 1,25% [18], soit un chiffre nettement plus faible que les prévalences entre 3% et 10% usuellement mesurées en Europe [15].

Enfin, les enquêtes de victimation ont l'inconvénient de ne mesurer le préjudice subi que par un petit échantillon de la population, et elles ne peuvent donc mesurer les préjudices les plus élevés et les plus rares. Cependant, comme on le détaille ci-dessous, il semble crédible que ces préjudices les plus élevés comptent pour une part importante du préjudice total, part qui ne peut donc être appréhendée de manière fiable par une enquête par sondage. Cette question est peu abordée dans la littérature existante.

Il résulte de tout cela que les enquêtes de victimation sur les escroqueries montrent une sensibilité particulièrement élevée à la formulation des questions et au protocole d'enquête [7]. Au Royaume-Uni, l'utilisation d'un nouveau protocole par téléphone de l'enquête de victimation en Angleterre et au Pays de Galles coïncide avec une augmentation de 25% de la prévalence des escroqueries entre 2020 et 2022 [20], sans qu'il soit clair d'évaluer ce qui dans cette augmentation est dû à un phénomène réel ou au changement de mesure.

À ces difficultés inhérentes à la menée d'une enquête de victimation par sondage on pourrait opposer la stabilité et l'exhaustivité des données administratives : plaintes enregistrées par les services de sécurité, signalement de transactions frauduleuses auprès des établissements bancaires, etc. Néanmoins ces données peuvent présenter également des inconvénients majeurs, au premier rang desquels le fait qu'elles ne correspondent qu'aux infractions que la victime a choisi de signaler ; or les enquêtes de victimation indiquent toutes un taux de signalement auprès des forces de l'ordre particulièrement faible pour les escroqueries et infractions voisines, en général de 10% à 30%, et parfois encore inférieurs [7, 15, 18]. Par ailleurs, les données administratives obéissent à leurs propres nomenclatures et règles de constitution qui peuvent orienter l'analyse et biaiser les résultats [5]. Enfin, en ce qui concerne les montants, les données de plaintes auprès des forces de l'ordre correspondent souvent à la transcription des montants indiqués par les victimes : il s'agit donc de données déclaratives qui, de ce fait, partagent certaines des mêmes fragilités que les données de victimation concernant l'exactitude avec laquelle la victime décrit les faits survenus. Elles peuvent également différer de la qualification qui aura été retenue par un magistrat amené à se prononcer, le cas échéant, sur les mêmes faits.

Il est donc difficile de comparer les différents chiffrages qui ont pu être réalisés sur l'ampleur économique des escroqueries et des infractions proches. Citons cependant [18] et [12] qui trouvent un préjudice total aux États-Unis de respectivement 3,2 milliards de dollars en 2017 pour les escroqueries et de 15 milliards de dollars en 2018 pour les vols d'identité (y compris les usurpations de moyens de paiement et certaines fraudes aux prestations sociales), ou [13] qui évalue à 1,8 milliards de livres sterling le préjudice économique direct des escroqueries en Angleterre et au pays de Galles en 2016 (le préjudice total, comprenant également le préjudice moral, le coût de la prévention, celui de la répression, etc. étant quant à lui évalué à 4,7 milliards de livres sterling). Pour le cas américain comme pour le cas britannique, il s'agit de chiffres produits à partir d'enquêtes de victimation, sur une définition plutôt restreinte des escroqueries qui exclut notamment les fraudes

aux moyens de paiement (à l'exception du chiffrage des vols d'identité aux Etats-Unis), et pour lesquels les victimes se limitent aux personnes physiques. Il existe peu ou pas d'études qui proposent des chiffrements globaux pour le préjudice subi par l'ensemble des personnes morales, bien que des chiffrements partiels par secteur aient été proposés [17].

Nous proposons dans ce qui suit de mobiliser différentes sources de données pour évaluer de manière aussi précise et exhaustive que possible le montant du préjudice total dû aux escroqueries et aux fraudes aux moyens de paiement en France. La première source est l'ensemble des infractions que les services de police et de gendarmerie ont enregistrées entre 2016 et 2023 avec un lieu de commission situé en France métropolitaine. Les victimes de ces infractions peuvent être aussi bien des personnes physiques que des personnes morales. Pour une partie de ces infractions, nous disposons du montant du préjudice lié à l'infraction, et dans cette étude nous inférons les montants manquants à partir des montants renseignés. La deuxième source que nous utilisons est une enquête par sondage réalisée en France métropolitaine en 2018 et 2019, au cours de laquelle un échantillon tiré aléatoirement dans la population de métropole a été interrogé sur les atteintes subies au cours de l'année précédente, y compris sur les pertes encourues. Cela nous permet d'estimer le montant du préjudice subi par les personnes physiques du fait d'atteintes qui n'ont pas fait l'objet d'un dépôt de plainte. En revanche, nous ne sommes malheureusement pas en mesure d'établir un chiffrage, même conjectural, pour le préjudice subi par les personnes morales qui n'ont pas déposé de plainte : les sources en la matière sont trop peu nombreuses ou inadaptées à ce stade.

2 Délimitation du champ

L'analyse statistique de la délinquance se heurte invariablement à la question de la catégorisation des comportements considérés. Délimiter l'ensemble des agissements délictueux qui relèvent de la famille des escroqueries ne va pas de soi, comme le montre la diversité des périmètres retenus pour la notion de *fraud* dans les études anglo-saxonnes ([17], [18], [4]). Le code pénal français définit quant à lui l'escroquerie comme « le fait, soit par l'usage d'un faux nom ou d'une fausse qualité, soit par l'abus d'une qualité vraie, soit par l'emploi de manœuvres frauduleuses, de tromper une personne physique ou morale et de la déterminer ainsi, à son préjudice ou au préjudice d'un tiers, à remettre des fonds, des valeurs ou un bien quelconque, à fournir un service ou à consentir un acte opérant obligation ou décharge. » (article 313-1). À l'aide de données judiciaires, il serait envisageable d'utiliser la catégorisation effectuée par les juges sur la base des textes légaux, dont notamment cette définition de l'escroquerie par le code pénal. Cependant une telle démarche serait très loin d'un recensement exhaustif des escroqueries puisqu'elle ne saurait prendre en compte que les infractions que la justice a eu à traiter.

Dans le cadre de cette étude, on se propose de solliciter deux sources de données : d'une part les infractions constatées et enregistrées par les services de police et de gendarmerie, et d'autre part par les données des enquêtes de victimation réalisées par la statistique publique en population générale. Le périmètre des infractions que l'on retient est, de manière pragmatique, celui qui résulte des classements inhérents à ces données, classements opérés soit par les policiers et les gendarmes lors de la saisie des données recueillies dans l'exercice de leurs fonctions, soit par les répondants aux enquêtes de victimation concernant les faits qu'ils ont eux-même subis.

Comme nous l'expliquons ci-dessous, cela nous amène à retenir un périmètre légèrement plus large que celui des seules escroqueries au sens de l'article 313-1 du code pénal, puisqu'il inclut également des fraudes aux moyens de paiement qui peuvent relever d'autres incriminations telles que la contrefaçon ou falsification d'un instrument de paiement (article L. 163-3 du code monétaire et financier), ou des infractions commerciales telles que la publicité mensongère.

La police et la gendarmerie classifient les infractions dont elles ont connaissance selon la nomenclature dite des *natures d'infractions* ou NATINF. Cette nomenclature compte plusieurs dizaines de milliers de NATINF, qui sont regroupées en grandes catégories au sein de la NFI (nomenclature

française des infractions, établie en 2021 par les ministères de l'Intérieur et de la Justice). On a choisi de retenir trois catégories de la NFI : la catégorie 07.A1 (*Escroquerie*), la catégorie 07.B1.2 (*Contrefaçon de moyens de paiement autres que la monnaie*) qui englobe un certain nombre de fraudes aux chèques et aux cartes bancaires tels que la falsification ou la contrefaçon de chèques ou d'autres instruments de paiement, le vol ou l'usage frauduleux d'instrument perdu, ou encore l'usurpation de numéro de carte ou d'identifiants bancaires, et la catégorie 08.D3.4 (*Infractions aux réglementations commerciales ou protection du consommateur*). La principale motivation pour adjoindre la catégorie 07.B1.2 à la catégorie 07.A1 est que ces deux groupes sont poreux dans la pratique de terrain des forces de police et de gendarmerie, au sens où certaines fraudes à la carte bancaire sont parfois étiquetées avec des NATINF relevant de la catégorie NFI 07.A1, et parfois des NATINF relevant de la catégorie NFI 07.B1.2 [8]. Prendre en compte également la catégorie 08.D3.4 ajoute une quantité négligeable de données tout en assurant une meilleure homogénéité avec les données d'enquête, cf. infra.

Deux questions posées dans le cadre de l'enquête de victimation « *Cadre de vie et sécurité* » (CVS) concernent les deux ensembles que constituent les escroqueries et les fraudes aux moyens de paiement. En effet, les enquêtés ont été interrogés d'une part sur les « arnaques » dont ils auraient pu être victimes au cours de l'année écoulée, et d'autre part sur les « débits frauduleux » qu'ils auraient pu constater sur leurs comptes bancaires pendant la même période (dans les deux cas il s'agit des termes exacts employés par le questionnaire de l'enquête). Ces deux questions ne coïncident qu'approximativement avec les catégories de la NFI : ainsi la question sur les arnaques inclut explicitement non seulement les escroqueries, mais également les fraudes commerciales dont a pu être victime la personne interrogée, si bien que nous prenons en compte dans les données administratives la catégorie NFI 08.D3.4 (*Infractions aux réglementations commerciales ou protection du consommateur*). Notons à ce sujet qu'entre 2016 et 2023, la catégorie NFI 08.D3.4 représente 50 à 100 fois moins d'infractions que la catégorie NFI 07.A1 dans les données administratives (et 15 à 35 fois moins d'infractions que la catégorie NFI 07.B1.2). Par ailleurs, tous les résultats présentés dans ce document de travail ont été répliqués sans inclure dans les données la catégorie NFI 08.D3.4 : cela n'aboutit à aucun changement notable.

De plus, la question de l'enquête CVS sur les débits frauduleux demande au répondant d'écarter les vols de cartes bancaires, qui sont pourtant comptabilisés en général par les forces de police et de gendarmeries avec des NATINF de la catégorie NFI 07.B1.2 (*Contrefaçon de moyens de paiement autres que la monnaie*). Selon les données collectées par l'Observatoire de la sécurité des moyens de paiements, en 2022 [19], les transactions frauduleuses qui résultaient d'une carte bancaire perdue, volée ou non parvenue représentaient 15% du volume des transactions relevant d'une fraude à un moyen de paiement (1,1 millions de transactions sur 7,1 millions). En mesurant en valeur de ces transactions frauduleuses et non plus en volume, cette proportion était de 8% (96 millions d'euros sur 1,2 milliards d'euros). Il apparaît donc comme vraisemblable que les possibles écarts entre les périmètres des données administratives et des données d'enquête ne concernent que des faits minoritaires par rapport à l'ensemble des faits dont il est question ici.

Nous nous limitons enfin à l'évaluation du préjudice direct subi par la victime, et nous ne cherchons pas à chiffrer les autres coûts liés à l'infraction, tels que ceux de sa répression ou de sa prévention. Enfin, il a été retenu le choix de limiter l'étude à la France métropolitaine car les enquêtes de victimation réalisées en outre-mer ne l'ont été que ponctuellement dans chaque département ou région d'Outre-mer (DROM) : La Réunion en 2011, Guadeloupe, Guyane et Martinique en 2015 et Mayotte en 2020.

3 Données

3.1 Données administratives

3.1.1 Organisation des données

Les policiers et les gendarmes qui ont connaissance d'une infraction (souvent à la suite d'un dépôt de plainte d'une victime, mais pas uniquement) enregistrent les informations dont ils disposent à l'aide d'un logiciel de rédaction des procédures : nature de l'infraction (NATINF), localisation géographique et temporelle, nombre de victimes, caractéristiques sociodémographiques de ces dernières, etc.

Ces infractions sont regroupées par procédures : à une même procédure peut correspondre une ou plusieurs infractions, et chaque infraction peut compter zéro, une ou plusieurs victimes. Pour des affaires complexes, une même procédure peut donc regrouper plusieurs infractions de NATINF différentes ou identiques, et de multiples victimes. En outre, ces victimes peuvent être des personnes physiques ou des personnes morales.

La base statistique Infractions du Service statistique ministériel de la sécurité intérieure (SSMSI) recense toutes les infractions commises en France et relevées lors de l'établissement du procès verbal ou de l'enregistrement de la plainte par les services de police et de gendarmerie [23]. Elle contient des informations sur la date et le lieu de commission de l'infraction, ainsi que la NATINF qui a été saisie et un identifiant de la procédure dont elle est issue. La base statistique Victimes du SSMSI décrit quant à elle les victimes de ces infractions : notamment leur nature (personne physique ou morale), et leur sexe, leur âge et leur nationalité s'il s'agit de personnes physiques.

Comme on l'explique ci-dessous, les données sur le montant du préjudice subi sont disponibles au niveau de la procédure. On a donc fait ici le choix de travailler sur toutes les procédures qui comportent au moins une infraction étiquetée avec une NATINF des catégories NFI 07.A1 (*Escroquerie*), 07.B1.2 (*Contrefaçon de moyens de paiement autres que la monnaie*) et 08.D3.4 (*Infractions aux réglementations commerciales ou protection du consommateur*). Au sein de ces procédures, on a considéré uniquement les victimes d'infractions qui relèvent de ces trois catégories NFI. **Dans ce qui suit, par souci de brièveté, on s'autorisera à parler des « procédures d'escroquerie » ou du « préjudice des escroqueries » pour renvoyer en fait à l'ensemble des données administratives qui relèvent de ces trois catégories.**

L'utilisation de la nomenclature des NATINF s'est généralisée à partir de 2016¹, ce qui nous contraint à retenir la période 2016-2023 pour ces données administratives. En outre, on se restreint à la métropole afin d'avoir un champ géographique homogène à celui des données de l'enquête CVS qui n'a été réalisée qu'en métropole.

La répartition entre police et gendarmerie est principalement territoriale, la police étant compétente essentiellement sur les grandes agglomérations, et la gendarmerie sur les villes petites et moyennes et les zones rurales. Une victime peut cependant choisir de porter plainte devant le service de son choix, ce qui explique que la gendarmerie traite également un petit volume d'infractions commises en zone police, et inversement. Au sein des procédures considérées, la proportion des procédures enregistrées par la gendarmerie s'élève à 39%. Si l'on regarde en termes d'infractions ou de victimes, cette proportion ne se modifie qu'à peine (37% des infractions et 39% des victimes).

Les logiciels de rédaction des procédures et les pratiques d'enregistrement varient quelque peu entre les services de police et de gendarmerie. Notamment, les gendarmes et plus rarement les agents de police rédigent un descriptif de l'affaire appelé *MANOP* pour manière d'opérer : il s'agit d'un texte d'une ou plusieurs phrases courtes qui parfois contient une information sur le préjudice — en général il s'agit du montant déclaré par les victimes. Par ailleurs, les gendarmes peuvent également saisir ce montant du préjudice dans un champ dédié, de manière redondante ou non à ce qui est saisi dans la *MANOP*. Ces informations sont saisies au niveau de la procédure. Pour les

1. en 2015 sur le périmètre de la police nationale et en 2016 sur celui de la gendarmerie nationale.

procédures d’escroqueries enregistrées par la gendarmerie, la quasi-totalité (99%) des MANOPS sont renseignées, ainsi que 52% des champs préjudice. En revanche, ces informations sont beaucoup moins souvent disponibles dans les données enregistrées par la police : seulement 6% des MANOPS d’escroqueries en police contiennent un descriptif des faits incriminés, et le champ préjudice n’est à peu près jamais rempli. Combiner les deux champs permet de reconstituer ainsi le préjudice enregistré au niveau de la procédure pour 70% des procédures en gendarmerie et pour 1% des procédures en police, soit pour 28% de l’ensemble des procédures.

Outre ces données issues des enregistrements effectués par les services de police et de gendarmerie, nous utilisons également des données directement saisies par les victimes sur deux plateformes en ligne. La plateforme Perceval, ouverte en juin 2018, permet de signaler une fraude aux moyens de paiement. Elle est ouverte aux victimes qui disposent toujours de leur instrument de paiement (les utilisations frauduleuses de carte de paiement volées ou perdues sont donc exclues). Un signalement sur Perceval permet le cas échéant de se faire rembourser par l’établissement bancaire sans avoir à déposer une plainte. La plateforme THESEE, ouverte en mars 2022, permet un signalement ou un dépôt de plainte en cas d’escroquerie commise en ligne.

3.1.2 Description des préjudices dans les données administratives

Nous disposons de 2,16 millions de procédures d’escroqueries sur la période 2016-2023 en France métropolitaine, ce qui représente 2,01 millions de victimes personnes physiques et 260 000 victimes personnes morales (hors données des plateformes Perceval et THESEE). Parmi ces procédures, 600 000 comportent un préjudice lisible, pour 570 000 victimes personnes physiques et 63 000 victimes personnes morales.

	2016	2017	2018	2019	2020	2021	2022	2023
Nb proc.	43 971	48 733	53 533	61 845	76 758	94 409	88 219	87 271
Nb vic.	46 321	51 276	56 070	63 708	78 747	96 474	89 642	88 937
moyenne	3 236	3 423	3 944	4 028	3 669	4 679	4 799	4 859
q0.25	155	160	164	172	170	200	199	213
q0.5	500	500	530	550	545	620	664	812
q0.75	1 500	1 500	1 550	1 753	1 800	2 000	2 062	2 500
q0.9	4 231	4 400	5 000	6 000	5 500	6 000	6 500	7 700
q0.99	45 000	48 664	52 436	63 134	51 209	60 000	60 573	65 201
q0.999	280 000	240 000	290 000	300 000	280 000	280 000	400 000	330 000
max	4 110 000	18 390 000	2 900 000	6 450 000	5 980 000	43 970 000	40 150 000	6 500 000

TAB. 1 – *Distribution des montants de préjudice (personnes physiques, données administratives, infractions commises en France métropolitaine)*

Note : La première ligne correspond au nombre de procédures avec un préjudice lisible et où au moins une victime est une personne physique, la deuxième au nombre de victimes personnes physiques de ces procédures, la troisième à la moyenne du montant par victime du préjudice associé à chaque procédure pondérée par le nombre de victimes personnes physiques. Les lignes suivantes correspondent aux quantiles et au maximum de la distribution pondérée du montant.

Lecture : pour 75% des personnes physiques qui avaient été enregistrées comme victime au sein d’une procédure ouverte en 2016 et où le préjudice était lisible, ce préjudice était d’un montant par victime inférieur ou égal à 1 500 euros.

Champ : France métropolitaine.

Source : SSMSI, bases statistiques des infractions et des victimes de crimes et délits enregistrés par la police et la gendarmerie entre 2016 et 2023.

Dans les tables 1 et 2, on peut voir les grandes caractéristiques des valeurs renseignées du préjudice dans les données administratives. Elles sont ventilées par année, et selon la nature des victimes

(personnes physiques ou personnes morales). Les chiffres indiqués correspondent à la distribution du préjudice moyen par victime au sein d'une même procédure, que nous pondérons par le nombre de victimes (personnes physiques ou personnes morales) des procédures. Par exemple si le préjudice renseigné pour une procédure est de 300, et que cette procédure compte deux victimes personnes physiques et une victime personne morale, alors cette procédure sera comptabilisée comme un préjudice de 100 pour deux victimes personnes physiques, et un préjudice de 100 pour une victime personne morale. Les deux premières lignes de chacune des deux tables correspondent respectivement au nombre total de procédures dont le préjudice est lisible, et au nombre de victimes de ces procédures. Notons que la plupart des procédures (98% du total) ne comptent qu'une seule victime.

	2016	2017	2018	2019	2020	2021	2022	2023
Nb proc.	4 389	5 436	6 075	6 587	6 670	7 035	8 030	8 838
Nb vic.	5 883	7 198	7 684	8 168	7 769	7 934	8 781	9 318
moyenne	12 886	14 303	13 701	16 129	16 329	15 114	17 137	20 356
q0.25	168	157	240	239	245	400	437	484
q0.5	840	750	990	1 350	1 336	1 985	1 942	2 000
q0.75	3 998	3 600	4 200	5 000	6 000	6 705	7 202	7 662
q0.9	15 000	14 808	15 312	19 927	22 483	25 000	24 549	27 342
q0.99	300 000	230 695	183 774	326 728	300 469	276 434	251 847	300 000
q0.999	820 000	900 000	1 370 000	980 000	1 120 000	900 000	1 300 000	1 490 000
max	2 800 000	16 610 000	9 100 000	6 970 000	4 000 000	1 950 000	7 550 000	14 000 000

TAB. 2 – *Distribution des montants de préjudice (personnes morales, données administratives, infractions commises en France métropolitaine)*

Note : La première ligne correspond au nombre de procédures avec un préjudice lisible et où au moins une victime est une personne morale, la deuxième au nombre de victimes personnes morales de ces procédures, la troisième à la moyenne du montant par victime du préjudice associé à chaque procédure pondérée par le nombre de victimes personnes morales. Les lignes suivantes correspondent aux quantiles et au maximum de la distribution pondérée du montant.

Lecture : en 2016, pour 75% des personnes morales qui avaient été enregistrées comme victime au sein d'une procédure ouverte en 2016 et où le préjudice était lisible, ce préjudice était d'un montant par victime inférieur ou égal à 3 998 euros.

Champ : France métropolitaine.

Source : SSMSI, bases statistiques des infractions et des victimes de crimes et délits enregistrés par la police et la gendarmerie entre 2016 et 2023.

On peut remarquer que la distribution de ces montants est particulièrement asymétrique et étalée vers la droite. Si la grande majorité des préjudices sont inférieurs à quelques milliers d'euros pour les personnes physiques, et à quelques dizaines de milliers d'euros pour les personnes morales, des préjudices bien supérieurs sont également présents dans les données, avec des chiffres allant jusqu'à plusieurs dizaines de millions pour les cas les plus élevés. Le préjudice moyen est ainsi cinq à sept fois supérieur au préjudice médian pour les personnes physiques, et huit à dix-huit fois supérieur pour les personnes morales. Ce type de comportement renvoie à la théorie des distributions à queues épaisses dont le caractère asymétrique est mesuré par un paramètre appelé indice de queue. Les estimateurs usuels de cet indice de queue tels que les estimateurs de Hill ou de Pickands prennent des valeurs entre 0,8 et 1,4, ce qui est compatible avec des distributions théoriques qui n'admettent pas de variance, voire pas d'espérance (cf. Annexe A pour une présentation détaillée de ces points). Cela invite à manipuler avec grande prudence la plupart des méthodes statistiques usuelles. En particulier, toutes les méthodes fondées sur la minimisation d'une variance, telles que la régression, peuvent s'avérer mal adaptées à de telles données.

Le nombre de préjudices recensés augmente entre 2016 et 2023, alors même que leur distribution se décale vers la droite. Cela est vrai aussi bien pour les personnes physiques que pour les personnes morales. La crise sanitaire de 2020-2021 n'a pas d'incidence marquée sur cette évolution à la hausse du nombre comme du montant des préjudices dus à des escroqueries.

3.2 Données d'enquête

3.2.1 Nombre d'atteintes et dépôt de plainte

L'enquête « *Cadre de vie et sécurité* » (CVS) a été réalisée annuellement entre 2007 et 2021 (à l'exception de 2020 du fait de la crise sanitaire). Un échantillon aléatoire de 20 000 à 25 000 ménages en France métropolitaine étaient interrogés chaque année sur les faits dont ils avaient pu être victimes au cours de l'année précédant l'enquête, qu'ils aient ou non porté plainte. Cette enquête était réalisée par l'Insee, en partenariat avec l'Observatoire national de la délinquance et des réponses pénales (ONDRP, supprimé fin 2020) et le SSMSI (créé fin 2014) au ministère de l'Intérieur.

Entre 2011 et 2021, les répondants ont répondu à des questions sur les débits frauduleux qu'ils auraient pu constater sur les comptes bancaires de leurs ménages. En 2018 et 2019 uniquement, des questions supplémentaires ont porté sur les arnaques subies par les individus interrogés. Nous nous restreignons donc ici aux résultats des enquêtes CVS de 2018 et 2019, portant respectivement sur les faits commis en 2017 et 2018. La prévalence des débits frauduleux parmi les ménages de France métropolitaine s'élève selon ces données à 4,2% en 2017 et 4,3% en 2018². Celle des arnaques parmi les individus s'élève à 3,3% en 2017 et 2,4% en 2018. Dans ce qui suit, nous confondons ces deux familles d'atteintes, et nous considérons qu'elles délimitent ensemble un périmètre d'infractions similaire à celui des champs combinés des catégories NFI 07.A1, 07.B1.2 et 08.D3.4.

Concernant le montant du préjudice subi par des personnes physiques, les données issues de CVS fournissent les informations suivantes pour chacune des deux années 2017 et 2018: le nombre d'atteintes n_i subies par le répondant i au cours de l'année, le montant m_i de la dernière atteinte de l'année, si cette dernière atteinte a donné lieu au dépôt d'une plainte (variable indicatrice p_i qui vaut 1 s'il y a eu dépôt de plainte et 0 sinon), et enfin un poids de pondération w_i . En revanche, s'agissant des répondants qui ont subi plusieurs atteintes, nous ne disposons pas de l'information sur le nombre de plaintes déposées, mais uniquement sur la présence ou non d'un dépôt de plainte pour la dernière atteinte subie.

Cela permet d'estimer d'abord le nombre total d'atteintes subies par la population sur l'année :

$$\hat{n} = \sum_i n_i w_i,$$

ensuite le nombre d'atteintes qui ont donné lieu à une plainte (avec donc une possible sous-estimation concernant les répondants qui auraient effectué plusieurs dépôts de plainte dans l'année) :

$$\hat{p} = \sum_{i, n_i > 0} p_i w_i,$$

puis la distribution du préjudice pour les atteintes qui ont donnée lieu à une plainte $(m_i)_{i, p_i=1}$, et enfin la distribution du préjudice pour les atteintes qui n'ont pas donné lieu à une plainte $(m_i)_{i, p_i=0}$. Les données montrent cependant que les chiffres de CVS concernant le dépôt de plainte sont relativement fragiles. En premier lieu, la proportion des atteintes qui donnent lieu à une plainte est faible : le rapport entre \hat{p} et \hat{n} tels que définis ci-dessus vaut environ 20% si on se limite aux débits frauduleux, et seulement 7% si on se limite aux arnaques. Il en résulte que les montants

2. Cette prévalence des débits frauduleux est par ailleurs restée stable sur toute la période de 2016 à 2021. Elle avait en revanche sensiblement augmenté entre 2011 et 2016 [25]

des préjudices donnant lieu à un dépôt de plainte sont estimés à partir d'un nombre restreint de réponses à l'enquête CVS.

De manière plus problématique, on constate des discordances importantes entre d'un côté le nombre \hat{p} d'atteintes donnant lieu à des plaintes mesuré grâce à l'enquête CVS, et de l'autre le nombre de victimes qui sont effectivement enregistrées par les forces de police et de gendarmerie. Rappelons que \hat{p} devrait être une minoration du nombre réel de plaintes puisqu'il estime en fait le nombre de plaintes portées pour la dernière atteinte subie dans l'année, et non pour toutes les atteintes. De plus, d'autres questions de l'enquête CVS, non prises en compte ici, portaient sur l'éventuel dépôt d'une main courante, et, après l'entrée en service de la plateforme de signalement en ligne Perceval, sur un possible signalement passé par cette plateforme, si bien qu'en principe, ces différentes démarches ont bien été distinguées par les répondants et ne devraient pas être confondues dans les chiffres de dépôts de plainte que nous proposons ici³.

La table 3 montre en fait que ces nombres, qui devraient être grossièrement identiques dans les données administratives et dans les données CVS, ou en tout cas légèrement inférieur pour le nombre issu des données CVS, varient du simple au double, le plus faible étant le nombre issu des données administratives. Pour expliquer cette disparité, on peut penser que les personnes interrogées dans le cadre de l'enquête CVS pourraient avoir eu tendance à surévaluer dans leurs réponses au questionnaire de l'enquête leur recours effectif à un dépôt de plainte, ou encore simplement à se tromper quant à la date de leur dépôt de plainte dans leurs réponses données plusieurs mois après les faits. Par ailleurs, on peut remarquer que selon CVS, le nombre de plaintes pour 2017 est un tiers plus élevé que celui pour 2018. Cette variabilité d'une année sur l'autre invite également à une certaine prudence. Notons qu'un appariement à venir entre les données CVS et les bases du SSMSI devrait permettre de répondre à ces interrogations.

	2017	2018
Débites frauduleux	278 000	227 000
Arnaques	130 000	92 000
Total plaintes selon CVS	408 000	319 000
Total plaintes selon données admin.	208 000	211 000

TAB. 3 – *Nombre de plaintes dans les données de victimation et nombre de victimes dans les données administratives*

Note : Les trois premières lignes correspondent au nombre de plaintes déposées d'après les résultats de l'enquête CVS, la dernière au nombre de victimes enregistrées par la police et la gendarmerie.

Lecture : en 2017, selon l'enquête CVS, 278 000 personnes ont porté plainte suite à un débit frauduleux.

Champ : France métropolitaine.

Sources : Insee-ONDRP-SSMSI, enquêtes Cadre de vie et sécurité 2018 et 2019, et SSMSI, bases statistiques des infractions et des victimes de crimes et délits enregistrés par la police et la gendarmerie entre 2016 et 2023.

3.2.2 Description des préjudices dans les données CVS

Nous indiquons dans la table 4 les caractéristiques de la distribution du montant m_i du dernier préjudice subi telle qu'elle apparaît dans les données CVS, pondérée par les poids de sondage de l'enquête. La première ligne correspond au nombre de répondants, c'est-à-dire au nombre non pondéré d'observations disponibles.

La comparaison de la distribution des préjudices selon les données de la police et de la gendarmerie et selon les données CVS (du moins pour les atteintes qui selon les répondants ont donné lieu à une plainte) fait apparaître un autre phénomène. Si les trois quartiles et le quantile d'ordre 0,9

3. Néanmoins des travaux méthodologiques sont en cours pour confirmer ou infirmer ce point.

	2017 avec plainte	2018 avec plainte	2017 sans plainte	2018 sans plainte
Nb répondants	175	112	941	707
moyenne	828	1 256	449	348
q0.25	160	125	30	40
q0.5	400	600	70	100
q0.75	1 000	1 800	285	296
q0.9	2 100	4 000	800	800
max	8 000	20 000	30 000	12 000

TAB. 4 – *Distribution des montants de préjudice (CVS)*

Note : La première ligne correspond au nombre de répondants qui se déclarent victimes dans l'échantillon CVS, la deuxième à la valeur moyenne (pondérée selon les poids de sondage de l'enquête CVS) du montant du préjudice dû à la dernière atteinte subie. Les lignes suivantes correspondent aux quantiles et au maximum de la distribution pondérée de ce montant. Pour les deux premières colonnes, il s'agit des répondants qui déclarent avoir porté plainte pour la dernière atteinte subie, pour les deux dernières colonnes, il s'agit des répondants qui déclarent n'avoir pas porté plainte pour cette dernière atteinte.

Lecture : pour 75% des victimes d'un débit frauduleux ou d'une arnaque en 2017 qui n'avaient pas porté plainte pour la dernière atteinte subie en 2017, le préjudice encouru lors de cette atteinte était inférieur ou égal à 285 euros.

Champ : France métropolitaine.

Source : Insee-ONDRP-SSMSI, enquêtes Cadre de vie et sécurité 2018 et 2019.

sont similaires selon les deux sources de données, on observe en revanche un moindre étalement vers la droite pour la distribution des données CVS. En effet, le décile supérieur de la distribution, soit la plage de valeurs entre le quantile d'ordre 0,9 et le maximum, est nettement plus resserré dans les données CVS que dans les données administratives. Cela peut correspondre au fait que les répondants auraient tendance à sous-déclarer dans leur réponse à l'enquête les montants des préjudices subis lorsqu'ils sont très élevés, ou cela peut aussi simplement résulter de ce que les rares victimes qui ont subi ces préjudices très importants, de l'ordre de cent mille euros ou plus, n'ont pas été sélectionnées dans l'échantillon de l'enquête CVS. Cependant, pour des distributions telles que celles qu'on observe dans la table 1, ces préjudices importants mais minoritaires peuvent compter pour une part significative du total, comme on le mentionne dans l'annexe A.

Il résulte de ce qui précède que le montant moyen ou total du préjudice des atteintes qui ont donné lieu à une plainte selon CVS semble donc un estimateur relativement fragile du montant moyen ou total réel. Les résultats qu'on pourrait obtenir en se fondant sur les seules données CVS seraient sensiblement différents de ce que l'on pourrait obtenir à partir des données de la police ou de la gendarmerie, qui sont a priori plus exhaustives et plus précises. En effet, dans le premier cas, on obtient par exemple un préjudice moyen d'environ 800 euros pour les escroqueries qui ont donné lieu à un dépôt de plainte en 2017 (table 4), alors qu'on trouve un chiffre presque quatre fois supérieur dans le deuxième cas (table 1).

On se restreint par conséquent à l'utilisation des données CVS pour la seule estimation du préjudice causé par les atteintes qui n'ont pas donné lieu à une plainte. On fait en effet l'hypothèse que ces préjudices ne présentent que peu de valeurs extrêmes, du moins pas sensiblement plus que ce qui est accessible dans les données CVS. Il semble en effet crédible que les pertes subies les plus élevées incitent fortement les victimes à porter plainte.

On court donc cependant le risque que ce préjudice sans dépôt de plainte soit sous-estimé par les données CVS, pour deux raisons. D'une part, comme on l'a vu, il semble plausible qu'une partie des répondants à l'enquête CVS qui n'ont pas porté plainte pendant l'année passée aient déclaré le contraire, et qu'ils soient donc écartés à tort des données que nous prenons en compte. D'autre part il est possible qu'on ne prenne pas en compte de rares préjudices très élevés qui n'auraient

pas donné lieu à un dépôt de plainte et qui ne seraient pas représentés dans l'échantillon CVS.

4 Méthodes

4.1 Estimation du préjudice lié aux infractions qui ont donné lieu à un dépôt de plainte

Les bases Infractions et Victimes du SSMSI contiennent toutes les plaintes enregistrées par les forces de police et de gendarmerie et constituent donc une source d'information exhaustive. La difficulté ici est que pour la quasi-totalité des procédures enregistrées par la police, et pour une partie importante de celles enregistrées par la gendarmerie, l'information sur le montant du préjudice n'a pas été saisie dans le logiciel de rédaction des procédures. Il faut donc reconstituer ces montants manquants (72% du total) à partir des montants lisibles (28%).

Pour traiter ce problème, on s'inscrit dans le cadre méthodologique du traitement de la non-réponse dans une enquête par sondage. On considère qu'on est dans un cas similaire à celui où une enquête a été menée par recensement exhaustif d'une population (les procédures pour escroquerie — au sens des procédures qui contiennent au moins une infraction qui relève des catégories NFI 07.A1, 07.B1.2 ou 08.D3.4— enregistrées par les forces de police et de gendarmerie), mais qu'on est confronté à un phénomène de non-réponse, puisque pour 72% des observations, la variable correspondant au montant du préjudice n'est pas renseignée. Les autres variables disponibles le sont cependant. On va donc chercher à imputer les valeurs manquantes à partir des données qui sont disponibles : les montants de préjudice renseignés et les autres variables disponibles, telles que l'âge des victimes, leur localisation géographique, ou le nombre d'infractions par procédure.

Il nous faut cependant examiner au préalable d'une part le lien entre ces autres variables et le fait que le montant du préjudice soit lisible, et d'autre part, le lien entre ces autres variables et la valeur de ce montant lorsqu'il est lisible. Comme on l'expose ci-dessous, cet examen nous permet de conclure que nos données sont compatibles avec l'hypothèse que les valeurs manquantes sont *missing at random* [14] : c'est-à-dire qu'il est raisonnable de supposer que le fait que le montant du préjudice soit renseigné ou non lorsqu'une plainte est déposée est sans lien avec la valeur même de ce montant. Cela nous permet dans un deuxième temps de choisir, parmi les méthodes d'imputation de non-réponse dans un cadre *missing at random*, celle qui semble la plus appropriée à notre cas précis.

4.1.1 Nature des données manquantes

On cherche ici à examiner si on peut mettre en évidence un lien entre le fait que le montant du préjudice soit lisible et les autres variables. Rappelons que les données administratives se partagent entre données enregistrées par la police nationale, principalement concernant les infractions localisées dans les grandes zones urbaines, et celles enregistrées par la gendarmerie, principalement dans le reste du territoire, et rappelons de plus que les montants lisibles se trouvent très majoritairement dans les données de la gendarmerie, comme le montre la table 5. Le lien entre la taille d'unité urbaine et le fait que le montant du préjudice soit lisible est donc tout à fait apparent.

Pour aller plus loin, on procède selon la démarche classique en apprentissage supervisé et en machine learning qui consiste à prédire si le montant est ou non lisible à partir des autres variables (ou *features*) dont nous disposons. Parmi ces *features* figurent pour toutes les procédures le nombre d'infractions de la procédure, la date de l'infraction la plus ancienne de la procédure, la région et la taille d'unité urbaine les plus fréquentes au sein de la procédure. Pour les procédures dont les victimes sont toutes des personnes physiques, il s'agit également de l'âge moyen des victimes, de la part d'hommes et de la part d'étrangers parmi les victimes. Pour les procédures avec au moins une victime personne morale, il s'agit aussi de la part de personnes physiques parmi les victimes.

TUU	Préj. lisible		Préj. non lisible	
	Gendarmerie	Police	Gendarmerie	Police
Commune rurale	190 900	300	71 100	28 000
2 000 à 4 999 habitants	66 400	100	26 700	9 400
5 000 à 9 999 habitants	69 200	200	29 200	10 400
10 000 à 19 999 habitants	56 300	300	25 400	15 600
20 000 à 49 999 habitants	32 200	1 600	15 200	72 600
50 000 à 99 999 habitants	27 400	2 200	13 000	110 300
100 000 à 199 999 habitants	17 900	1 600	9 500	75 800
200 000 à 1 999 999 habitants	100 300	6 100	48 600	487 300
Unité urbaine de Paris	24 600	4 600	12 000	492 900

TAB. 5 – Nombre de procédures par taille d'unité urbaine, par service et par lisibilité du préjudice de la procédure.

Lecture : Pour 190 900 procédures enregistrées par la gendarmerie, les infractions qui composent la procédure ont été majoritairement commises dans une commune rurale et le montant du préjudice associé à la procédure est lisible.

Champ : France métropolitaine.

Source : SSMSI, bases statistiques des infractions et des victimes de crimes et délits enregistrés par la police et la gendarmerie entre 2016 et 2023.

Nous utilisons trois méthodes de classification : une régression logistique standard, un arbre de décision et une classification par *support vector machines* avec noyau gaussien, ces deux dernières méthodes permettant de prendre en compte d'éventuelles dépendances non-linéaires. Les données sont divisées entre un échantillon d'entraînement (80% des données) et un échantillon de test (20% des données). Chaque méthode est calibrée par validation croisée à dix blocs sur l'échantillon d'entraînement, puis ses performances sont évaluées sur l'échantillon de test. Nous prenons comme critère d'évaluation le pourcentage de bien classés (montant lisible ou non lisible) parmi l'ensemble des données. Nous indiquons les résultats de la meilleure des méthodes utilisées au sens de ce critère sur l'échantillon de test dans la table 6. Outre la proportion de bien classés, nous faisons également figurer l'indicateur kappa de Cohen (qui mesure l'écart entre le pourcentage de bien classés obtenus et celui d'un résultat au hasard : une valeur 0 correspondant à une classification purement aléatoire, et 1 à une classification optimale), ainsi que la sensibilité (proportion des montants lisibles bien classés parmi l'ensemble des montants lisibles) et la spécificité (proportion des montants non lisibles bien classés parmi l'ensemble des montants classés).

On réitère cette approche soit pour les procédures dont les victimes sont toutes des personnes physiques ($N \approx 1\,930\,000$), soit pour celles où au moins une victime est une personne morale ($N \approx 220\,000$). On distingue également selon les *features* utilisées pour bien mettre en avant le rôle de la taille d'unité urbaine (TUU), principal critère de partage entre zone de police et zone de gendarmerie. On procède ainsi à la classification selon que l'ensemble des *features* est donné par la TUU seule, par la région seule (qui est l'autre variable géographique dont nous disposons), par la TUU et par la région, par toutes les variables à l'exception de la TUU, ou par toutes les variables. Cela permet de mettre en évidence le fait que les meilleures prédictions sont systématiquement établies à l'aide de la seule TUU, les autres variables n'améliorant aucunement les performances de la prédiction.

Ce qu'on prédit ainsi est en fait essentiellement le fait qu'une procédure ait été saisie par les services de police ou de gendarmerie. Pour confirmer qu'il n'y a pas d'autre déterminant de la lisibilité du montant, nous effectuons une classification des seules procédures saisies par la gendarmerie, selon la même approche que ci-dessus ($N \approx 760\,000$ ou $N \approx 72\,000$ selon que les procédures comptent ou non des victimes personnes morales). Cela ne permet pas de mettre en évidence de

Toutes procédures où toutes les victimes sont des personnes physiques					
	TUU	Région	TUU et Région	Tout sauf TUU	Tout
Proportion de bien classés	0,799	0,715	0,799	0,722	0,800
Kappa de Cohen	0,506	0,000	0,505	0,155	0,503
Sensibilité	0,646	0,000	0,641	0,191	0,639
Spécificité	0,860	1,000	0,864	1,000	0,870
Toutes procédures où au moins une victime est une personne morale					
	TUU	Région	TUU et Région	Tout sauf TUU	Tout
Proportion de bien classés	0,794	0,764	0,798	0,765	0,799
Kappa de Cohen	0,415	0,000	0,403	0,039	0,394
Sensibilité	0,527	0,000	0,481	0,036	0,457
Spécificité	0,906	1,000	0,899	0,993	0,905
Procédures en gendarmerie où toutes les victimes sont des personnes physiques					
	TUU	Région	TUU et Région	Tout sauf TUU	Tout
Proportion de bien classés	0,704	0,704	0,704	0,705	0,705
Kappa de Cohen	0,000	0,000	0,000	0,010	0,011
Sensibilité	1,000	1,000	1,000	0,999	0,999
Spécificité	0,000	0,000	0,002	0,011	0,011
Procédures en gendarmerie où au moins une victime est une personne morale					
	TUU	Région	TUU et Région	Tout sauf TUU	Tout
Proportion de bien classés	0,656	0,656	0,656	0,656	0,656
Kappa de Cohen	0,000	0,000	0,000	0,037	0,032
Sensibilité	1,000	1,000	1,000	1,000	1,000
Spécificité	0,000	0,000	0,003	0,061	0,053

TAB. 6 – Prédiction de la lisibilité du montant du préjudice

Lecture : lorsqu'on considère toutes les procédures où les victimes sont toutes des personnes physiques, dans la classification à partir de la TUU et de la région, la sensibilité, donc la proportion de bien classées au sein des procédures ayant un préjudice lisible, était de 0,641. Cette proportion était de 0,864 au sein des procédures ayant un préjudice non lisible (spécificité).

Champ : France métropolitaine.

Source : SSMSI, bases statistiques des infractions et des victimes de crimes et délits enregistrés par la police et la gendarmerie entre 2016 et 2023.

déterminant au fait qu'un montant soit ou non lisible dans les procédures enregistrées par la gendarmerie, puisqu'on obtient des performances du même ordre qu'une classification aléatoire, quels que soient la méthode de classification ou l'ensemble de *features* retenus. Nous mettons ainsi en évidence que parmi nos covariables, seule la taille d'unité urbaine présente une dépendance mesurable avec le fait que le montant du préjudice soit lisible, et ce uniquement via la répartition des procédures entre police et gendarmerie.

4.1.2 Lien entre les covariables et la valeur du montant du préjudice

On examine ici les liens dans les données administratives entre le montant du préjudice de la procédure, lorsqu'il est lisible, et les autres variables disponibles. La liste des variables considérées est la suivante. Pour toutes les procédures, il s'agit du nombre d'infractions de la procédure, de l'année de l'infraction la plus ancienne de la procédure, de la région (treize modalités) et de la taille d'unité urbaine⁴ (neuf modalités) les plus fréquentes parmi les infractions de la procédure. Pour les pro-

4. La modalité « Unité urbaine de Paris » de la variable Taille d'unité urbaine constitue un sous-ensemble strict de la modalité « Ile de France » de la variable Région.

cédures qui ne comptent que des victimes qui sont des personnes physiques, il s'agit également de l'âge moyen des victimes de la procédure, de la part d'hommes parmi les victimes et de la part d'étrangers parmi les victimes. Pour les procédures qui comptent au moins une victime qui est une personne morale, il s'agit de la part de personnes physiques parmi les victimes de la procédure.

On cherche donc à savoir si le montant du préjudice y est en lien avec ces différentes variables x_j . Au vu des valeurs extrêmes que peut prendre y , on choisit de travailler sur le montant logarithmique $\ln(y+1)$ qu'on régresse sur les autres variables. Cela donne des résultats significatifs pour la plupart de ces variables, cf. tables 7 et 8. Cependant, il se peut qu'un nombre restreint de données affecte la significativité de certaines variables. Pour vérifier cela, on effectue la même régression, mais en se limitant aux données qui correspondent à une certaine modalité soit pour la taille de l'unité urbaine (neuf modalités), soit pour l'année de la procédure (entre 2016 et 2023, soit huit modalités). Notons qu'en ce qui concerne la taille de l'unité urbaine, qui est la seule variable dont nous disposons qui est liée au fait qu'un préjudice soit lisible ou non, calibrer un modèle de régression différent pour chaque modalité revient en fait à pratiquer la méthode dite des groupes de contrôle homogènes, cf. infra.

Nos résultats sont regroupés dans les tables 7 et 8. Seulement un petit nombre de variables explicatives x_j sont corrélées de manière consistante avec le montant du préjudice exprimé en logarithme (vu que nous estimons au total plusieurs centaines de coefficients sur de multiples modèles, nous fixons le seuil de significativité à 1%, et non au traditionnel seuil de 5% qui produirait trop de résultats significatifs même en l'absence de lien réel). Pour les procédures qui ne comptent que des victimes personnes physiques, il s'agit de l'âge des victimes, de la part d'hommes, de l'année de la procédure et dans une moindre mesure du nombre d'infractions, pour lequel le signe positif ou négatif du lien avec le montant du préjudice s'avère variable selon les tailles d'unités urbaines et les années. Pour les procédures dont les victimes comptent des personnes morales, il s'agit de la part de personnes physiques et de l'année de la procédure. Ni la région, ni la taille d'unité urbaine ne permettent de mettre en évidence de lien significatif et stable avec le montant du préjudice. Remarquons en particulier que les régressions ne permettent donc pas de trouver de dépendance stable entre le montant du préjudice et la taille d'unité urbaine, variable qui est elle-même en lien étroit avec le fait que ce montant soit renseigné. Cela conforte donc l'hypothèse que la valeur du montant peut être supposée sans lien avec le fait que le montant soit lisible ou non.

Pour terminer, il est notable que toutes les régressions effectuées donnent des coefficients de détermination R^2 qui sont au mieux de l'ordre de quelques centièmes. La part de la variabilité du montant du préjudice qui est expliquée par les données dont nous disposons est donc tout à fait minoritaire.

4.1.3 Méthode d'imputation retenue

Inférer le total d'une distribution qui présente à la fois des valeurs manquantes et de fortes valeurs extrêmes est un exercice délicat. Il se peut en effet que certaines valeurs très élevées, qui ont une contribution non-négligeable dans le total, ne soient pas observées et qu'on sous-estime donc le total, ou à l'inverse, bien que ce soit moins probable, que des valeurs particulièrement élevées fassent partie des valeurs observées et qu'on se livre donc à une sur-estimation (cf. annexe A).

Dans le cadre *missing at random*, et en l'absence de valeurs extrêmes, une approche standard pour imputer des données manquantes à partir de covariables consiste d'abord à séparer la population en groupes dits groupes de réponses homogènes, c'est-à-dire au sein desquels la probabilité qu'une observation donne lieu à une valeur manquante est à peu près constante, puis à régresser au sein de chacun de ces groupes la variable d'intérêt quand elle est observée (ici le montant du préjudice) sur les autres variables de manière à pouvoir la prédire quand elle n'est pas observée [24]. Les résultats de théorie des sondages permettent alors d'exhiber une approximation de la variance de l'estimateur obtenu, et par conséquent des intervalles de confiance sous hypothèse de gaussianité asymptotique. Cependant, les données dont nous disposons sont très fortement non-gaussiennes, ce qui indique une certaine fragilité des méthodes qui s'appuient sur une estimation de la variance,

	Ensemble	Par TUU		Par année	
	Valeur du coef. et significativité	Nombre de coef. significatifs (sur 9)		Nombre de coef. significatifs (sur 8)	
		Positif	Négatif	Positif	Négatif
Constante	-95 ***	0	8	8	0
Âge moyen	0,010 ***	9	0	8	0
Part d'hommes en %	0,27 ***	9	0	8	0
Part d'étrangers en %	0,12 ***	2	0	2	0
Nombre d'infractions	0,0096 **	2	6	4	3
Année de la procédure	0,050 ***	8	0		
Auvergne Rhône Alpes	-0,032 **	0	2	0	0
Bourgogne Franche-Comté	-0,061 ***	0	4	0	1
Bretagne	-0,020	0	2	1	2
Centre Val de Loire	-0,12 ***	0	5	0	4
Corse	0,16 ***	2	0	1	0
Grand Est	Réf.	Réf.	Réf.	Réf.	Réf.
Hauts de France	-0,063 ***	0	3	0	3
Ile de France	0,12 ***	3	0	2	0
Normandie	-0,036 *	0	1	0	1
Nouvelle Aquitaine	-0,096 ***	0	3	0	4
Occitanie	-0,051 ***	0	1	0	1
Pays de la Loire	-0,083 ***	0	2	0	3
Provence Alpes Côte d'Azur	0,11 ***	3	0	4	0
Commune rurale	0,063 ***			3	1
2 000 à 4 999 habitants	-0,020			0	0
5 000 à 9 999 habitants	-0,033 *			0	0
10 000 à 19 999 habitants	-0,031 *			0	0
20 000 à 49 999 habitants	Réf.			Réf.	Réf.
50 000 à 99 999 habitants	0,007			0	0
100 000 à 199 999 habitants	-0,007			0	0
200 000 à 1 999 999 habitants	-0,079 ***			0	3
Unité urbaine de Paris	0,043			1	0

TAB. 7 – Régression du logarithme du montant du préjudice. Données administratives, procédures où les victimes sont toutes des personnes physiques.

Note : la première colonne contient les coefficients et leur significativité dans la régression effectuée avec l'ensemble des données ($N \approx 550\,000$, $R^2 \approx 1,8\%$). Les p-valeurs sont notées de manière suivante : *** pour les p-valeurs inférieures à 0,001, ** pour les p-valeurs entre 0,001 et 0,01 et * pour les p-valeurs entre 0,01 et 0,05. Les colonnes suivantes contiennent le nombre de fois où, lorsqu'on restreint les données à une seule taille d'unité urbaine ou à une seule année, le coefficient est significativement positif ou négatif au seuil de 1% (N varie entre 15 000 et 180 000, et R^2 entre 1,2% et 2,3%).

Lecture : avec des données restreintes à l'une des neuf catégories de taille d'unité urbaine, dans la régression, le nombre d'infractions se voit attribuer deux fois sur neuf un coefficient positif et une p-valeur inférieure à 1%, et six fois sur neuf un coefficient négatif et une p-valeur inférieure à 1%. Dans un cas sur neuf, la p-valeur était supérieure à 1%.

Champ : France métropolitaine.

Source : SSMSI, bases statistiques des infractions et des victimes de crimes et délits enregistrés par la police et la gendarmerie entre 2016 et 2023

	Ensemble	Par TUU		Par année	
	Valeur du coef. et significativité	Nombre de coef. significatifs (sur 9)		Nombre de coef. significatifs (sur 8)	
		Positif	Négatif	Positif	Négatif
Constante	-140 ***	0	8	8	0
Part de pers. phys. en %	-1,4 ***	0	9	0	8
Nombre d'infractions	0,0044	1	1	0	0
Année de la procédure	-0,072 ***	8	0		
Auvergne Rhône Alpes	0,047	1	0	0	0
Bourgogne Franche-Comté	-0,027	0	0	0	0
Bretagne	-0,13 *	0	0	0	1
Centre Val de Loire	-0,048	0	1	0	0
Corse	0,20	0	0	0	0
Grand Est	Réf.	Réf.	Réf.	Réf.	Réf.
Hauts de France	-0,009	0	0	0	0
Ile de France	0,18 *	0	0	1	0
Normandie	-0,16 *	0	1	0	1
Nouvelle Aquitaine	-0,25 ***	0	2	0	1
Occitanie	-0,038	1	1	0	0
Pays de la Loire	-0,089	0	0	0	0
Provence Alpes Côte d'Azur	0,21 ***	2	0	0	0
Commune rurale	0,17 ***			0	0
2 000 à 4 999 habitants	-0,007			0	0
5 000 à 9 999 habitants	-0,048			0	0
10 000 à 19 999 habitants	0,006			0	0
20 000 à 49 999 habitants	Réf.			Réf.	Réf.
50 000 à 99 999 habitants	0,056			0	0
100 000 à 199 999 habitants	-0,009			0	0
200 000 à 1 999 999 habitants	0,11 *			0	0
Unité urbaine de Paris	0,20 *			0	0

TAB. 8 – Régression du logarithme du montant du préjudice. Données administratives, procédures où au moins une victime est une personne morale.

Note : la première colonne contient les coefficients et leur significativité dans la régression effectuée avec l'ensemble des données ($N \approx 53\,000$, $R^2 \approx 1,6\%$). Les p-valeurs sont notées de manière suivante : *** pour les p-valeurs inférieures à 0,001, ** pour les p-valeurs entre 0,001 et 0,01 et * pour les p-valeurs entre 0,01 et 0,05. Les colonnes suivantes contiennent le nombre de fois où, lorsqu'on restreint les données à une seule taille d'unité urbaine ou à une seule année, le coefficient est significativement positif ou négatif au seuil de 1% (N varie entre 2 000 et 11 000, et R^2 entre 0,8% et 3,0%).

Lecture : avec des données restreintes à l'une des neuf catégories de taille d'unité urbaine, dans la régression, le nombre d'infractions se voit attribuer une fois sur neuf un coefficient positif et une p-valeur inférieure à 1%, et une fois sur neuf un coefficient négatif et une p-valeur inférieure à 1%. Dans les sept autres cas, la p-valeur était supérieure à 1%.

Champ : France métropolitaine.

Source : SSMSI, bases statistiques des infractions et des victimes de crimes et délits enregistrés par la police et la gendarmerie entre 2016 et 2023

telles que les régressions linéaires ordinaires. Même les méthodes plus sophistiquées, qui s'appuient sur une troncation des données ou sur des techniques de régression robustes [1], cherchent en général à optimiser un compromis biais-variance, si bien que leurs propriétés théoriques sont peu claires sur des données à la queue de distribution aussi épaisse que les nôtres et potentiellement sans variance.

On se contente donc ici d'appliquer la méthode d'imputation par régression au sein des groupes de réponses homogènes, sans chercher à produire d'intervalles de confiance gaussiens dont la validité serait contestable. Les groupes de réponses homogènes que nous utilisons sont simplement les différentes modalités de la taille d'unité urbaine, dont on a vu que c'était notre seul prédicteur fiable de la proportion de préjudices lisibles.

Pour les procédures qui ne comptent que des victimes qui sont des personnes physiques, les variables explicatives retenues pour la régression sont l'année de la première infraction de la procédure, le nombre d'infractions dans la procédure, l'âge moyen des victimes, ainsi que la proportion d'hommes parmi les victimes.

Pour les procédures qui comptent au moins une victime qui est une personne morale, les variables explicatives retenues pour la régression sont l'année de la première infraction de la procédure et la part de personnes physiques parmi les victimes de la procédure.

Une fois que nous disposons d'une valeur pour le préjudice de chaque procédure, soit que cette valeur était présente dans les données initiales, soit qu'elle ait été imputée selon la méthode décrite ci-dessus, la somme de ces préjudices pondérés par le nombre de victimes dans chaque procédure qui sont des personnes physiques permet d'obtenir une estimation du préjudice total subi par les personnes physiques. En pondérant par le nombre de victimes qui sont des personnes morales, on obtient de même une estimation du préjudice total subi par les personnes morales.

4.2 Estimation du préjudice lié aux infractions qui n'ont pas donné lieu à un dépôt de plainte

Faute d'avoir une source de données pour le préjudice subi par des personnes morales en l'absence de plainte, on se restreint ici aux personnes physiques. On se sert des données issues de l'enquête CVS pour estimer leur préjudice n'ayant pas donné lieu à un dépôt de plainte.

Le préjudice total, avec ou sans dépôt de plainte, peut être estimé en multipliant le préjudice moyen par atteinte (estimé à partir de la moyenne des préjudices de la dernière atteinte subie au moment de l'enquête) par le nombre total d'atteintes subies par la population, soit

$$\text{estimation du préjudice total} = \hat{n} \times \frac{\sum_{i, n_i > 0} m_i w_i}{\sum_{i, n_i > 0} w_i}$$

en reprenant les notations de la section précédente.

Nous voulons cependant estimer le préjudice des atteintes qui n'ont pas donné lieu à une plainte. Pour cela, nous faisons l'hypothèse simplificatrice que si un répondant i n'a pas porté plainte pour la dernière atteinte dont il a été victime (c'est-à-dire $p_i = 0$), alors il n'a pas porté plainte pour aucune des atteintes dont il a été victime, si bien que $\sum_{i, p_i=0} n_i w_i$ est un bon estimateur du nombre total des atteintes qui n'ont pas donné lieu à une plainte. On se donne alors comme estimateur du préjudice total de ces atteintes :

$$\text{estimation du préjudice sans plainte} = \frac{\sum_{i, p_i=0} n_i w_i \sum_{i, n_i > 0 \text{ et } p_i=0} m_i w_i}{\sum_{i, n_i > 0 \text{ et } p_i=0} w_i}$$

Cet estimateur donne des valeurs sensiblement différentes pour les deux années pour lesquelles les données CVS sont disponibles : 2,04 milliards en 2017 et 1,41 milliard en 2018. Cette baisse sensible ne se retrouve pas dans les données administratives en ce qui concerne les atteintes qui ont donné lieu à une plainte. Confronté à cette différence sensible entre les estimations pour

2017 et pour 2018, ainsi qu'au fait que nous n'avons pas de sources directes de données pour le préjudice des personnes physiques sans dépôt de plainte pour les autres années de la période, nous choisissons d'estimer un ratio entre le préjudice des personnes physiques avec et sans dépôt de plainte, en nous fondant sur les deux années dont nous disposons : 2017 et 2018

$$\text{ratio estimé} = \frac{\text{estimation du préjudice sans plainte en 2017 et 2018}}{\text{estimation du préjudice avec plainte en 2017 et 2018}} \quad (1)$$

Cela nous permet alors de proposer une valeur pour le préjudice sans plainte pour chaque année en supposant ce ratio constant sur la période : nous multiplions simplement ce ratio par le montant du préjudice sans plainte pour les personnes physiques pour la même année obtenu à partir des données administratives comme décrit ci-dessus. Du fait du caractère ad hoc de cette méthode, nous ne donnons pas d'intervalle de confiance pour le montant de ce préjudice sans dépôt de plainte.

4.3 Données des plateformes en ligne Perceval et THESEE

Un dernier ajustement doit être fait pour tenir compte des données correspondant aux signalements Perceval. Rappelons qu'à partir de juin 2018, les victimes de certains débits frauduleux avaient la possibilité de faire un signalement en ligne sur cette plateforme, ce qui ouvre dans de nombreux cas la possibilité d'un remboursement par l'établissement bancaire sans avoir à déposer une plainte. Le questionnaire de l'enquête CVS 2018 portait sur des faits antérieurs à l'ouverture de cette plateforme, mais le questionnaire de l'enquête CVS 2019 prenait en compte cette nouveauté : dans cette enquête, 42 répondants déclaraient avoir effectué en 2018 un signalement Perceval sans dépôt de plainte suite à un débit frauduleux. En reprenant les estimateurs ci-dessus, cela correspond à un préjudice estimé à 37 millions d'euros, ce qui est relativement proche du montant total de 43 millions d'euros effectivement enregistré sur la plateforme en 2018.

Pour tirer parti de l'exhaustivité des données de Perceval, nous les intégrons à notre estimation du préjudice total pour les années entre 2019 et 2023. Nous faisons l'hypothèse que la majorité des transactions frauduleuses signalées sur cette plateforme pendant cette période n'ont pas fait l'objet d'un dépôt de plainte. Il nous faut alors ajuster le calcul du ratio (1) utilisé pour les années 2019 à 2023 : nous soustrayons au numérateur dans (1) quatre fois le montant des 43 millions d'euros de préjudice déclarés sur Perceval au deuxième semestre de l'année 2018 afin d'avoir une somme équivalente en ordre de grandeur à ce qui aurait pu être déclaré pendant l'ensemble des deux années 2017 et 2018. Par la suite, lorsque nous multiplions ce ratio par le préjudice des personnes physiques avec dépôt de plainte, nous estimons ainsi un préjudice de personnes physiques sans dépôt de plainte ni signalement Perceval entre 2019 et 2023. Nous ajoutons enfin à cette estimation le montant effectivement déclaré sur Perceval. Nous n'appliquons pas cette démarche ni cet ajustement du ratio pour les années 2016 à 2018.

À partir de mars 2022, les victimes d'une escroquerie en ligne ont la possibilité de déposer leur plainte en ligne sur la plateforme dédiée THESEE. Ces plaintes en ligne sont bien distinguées des plaintes usuelles dans les bases statistiques. Nous choisissons de ne pas prendre en compte les chiffres issus de THESEE dans les calculs décrits ci-dessus, ni pour l'évaluation du préjudice qui a donné lieu à un dépôt de plainte classique, ni pour celle du préjudice des atteintes qui n'ont pas donné lieu à un dépôt de plainte : nous rajoutons les chiffres de THESEE postérieurement aux autres calculs que nous effectuons pour évaluer le préjudice total subi. Comme pour l'introduction de Perceval, nous faisons donc l'hypothèse que celle de THESEE n'a pas radicalement modifié les comportements des victimes en matière de dépôt de plainte ni les pratiques d'enregistrement des forces de sécurité.

Notons que les victimes qui portent plainte sur THESEE comme celles qui font un signalement sur Perceval doivent saisir le montant du préjudice qu'elles ont subi, montant qui nous est donc accessible : nous n'avons pas à imputer de valeurs manquantes en ce qui concerne ces données.

5 Résultats

La table 9 et la figure 1 regroupent l'ensemble de nos résultats. Selon nos estimations, le préjudice total subi par les personnes physiques est en augmentation constante au cours de la période considérée. Il augmente d'un peu plus de deux milliards d'euros en 2016 à plus de quatre milliards d'euros en 2023. Si on lui ajoute le préjudice des personnes morales qui a donné lieu à un dépôt de plainte, on passe d'un peu plus de trois milliards d'euros à un total qui excède cinq milliards d'euros en 2023.

	2016	2017	2018	2019	2020	2021	2022	2023
Pers. morales avec plainte	0,82	0,73	0,69	0,68	0,58	0,60	0,64	0,67
Pers. phys. avec plainte	0,71	0,74	0,81	0,96	1,06	1,36	1,33	1,38
Pers. phys. sans plainte, préjudice inféré	1,57	1,64	1,80	2,02	2,23	2,86	2,80	2,90
Pers. phys. sans plainte, réponses CVS		2,04	1,41					
Prej. Perceval	0,00	0,00	0,04	0,06	0,14	0,14	0,16	0,15
Prej. THESEE	0,00	0,00	0,00	0,00	0,00	0,00	0,07	0,08
Préj. total estimé (personnes physiques)	2,27	2,38	2,61	3,04	3,42	4,36	4,36	4,52

TAB. 9 – *Préjudice total (en milliards d'euros)*

Champ : France métropolitaine.

Sources : Insee-ONDRP-SSMSI, enquêtes Cadre de vie et sécurité 2018 et 2019, et SSMSI, bases statistiques des infractions et des victimes de crimes et délits enregistrés par la police et la gendarmerie entre 2016 et 2023.

Nous sommes ainsi en mesure de proposer une évaluation pour l'ensemble du préjudice subi par les personnes physiques entre 2016 et 2023, que celui-ci ait ou non fait l'objet d'une plainte auprès de la police ou de la gendarmerie. On peut remarquer à cet égard que chez les personnes physiques, le préjudice global est plus de deux fois plus élevé pour les atteintes qui n'ont pas donné lieu à un dépôt de plainte que pour celles enregistrées par la police ou la gendarmerie. Cela contraste avec les conclusions de [18] (cf. table 8 de cette référence) qui obtient un montant supérieur d'environ 25% pour les atteintes qui n'ont pas fait l'objet d'une plainte aux États-Unis en 2017, cependant sur un périmètre d'atteintes nettement plus restrictif que le nôtre puisqu'il n'inclut notamment pas les fraudes aux moyens de paiement. Notons de plus que les taux de dépôt de plainte ou les mécanismes de remboursement des assurances peuvent différer sensiblement entre la France et les États-Unis.

On constate une évolution distincte pour les préjudices des personnes physiques et morales qui ont donné lieu à un dépôt de plainte : forte hausse pour les premières, relative stabilité, voire baisse, pour les secondes. Cela peut sembler étonnant au vu de la table 2 qui indique au contraire une forte hausse en ce qui concerne les personnes morales à la fois du montant des préjudices individuels et du nombre de victimes. Il s'agit cependant des données où le préjudice est renseigné, donc très majoritairement issues de la gendarmerie ; or il s'avère que sur la période, le nombre de personnes morales enregistrées comme victimes par la police pour des infractions d'escroqueries a connu une baisse sensible, de 29 000 en 2016 à 18 000 en 2023, alors que ce même nombre était en hausse pour la gendarmerie (phénomène qui n'est en fait pas limité aux seules escroqueries et s'étend à d'autres familles d'infractions). La conjonction de ces deux évolutions explique la relative stabilité du montant total estimé voire sa baisse légère pour le préjudice des personnes morales. Il faut cependant noter que la catégorie des personnes morales dans les bases du SSMSI n'a pas encore fait l'objet d'un examen approfondi (appariement avec les données de l'Insee notamment), et qu'il n'est pas impossible qu'elle présente certains artefacts statistiques.

Comme nous l'avons indiqué, il manque dans nos résultats le montant du préjudice subi par les personnes morales qui n'a pas donné lieu à un dépôt de plainte, pour lequel nous ne disposons pas de données fiables. Une hypothèse qu'on peut émettre serait que ce préjudice des personnes morales sans plainte est environ deux fois plus élevé que celui avec plainte, comme ce que l'on

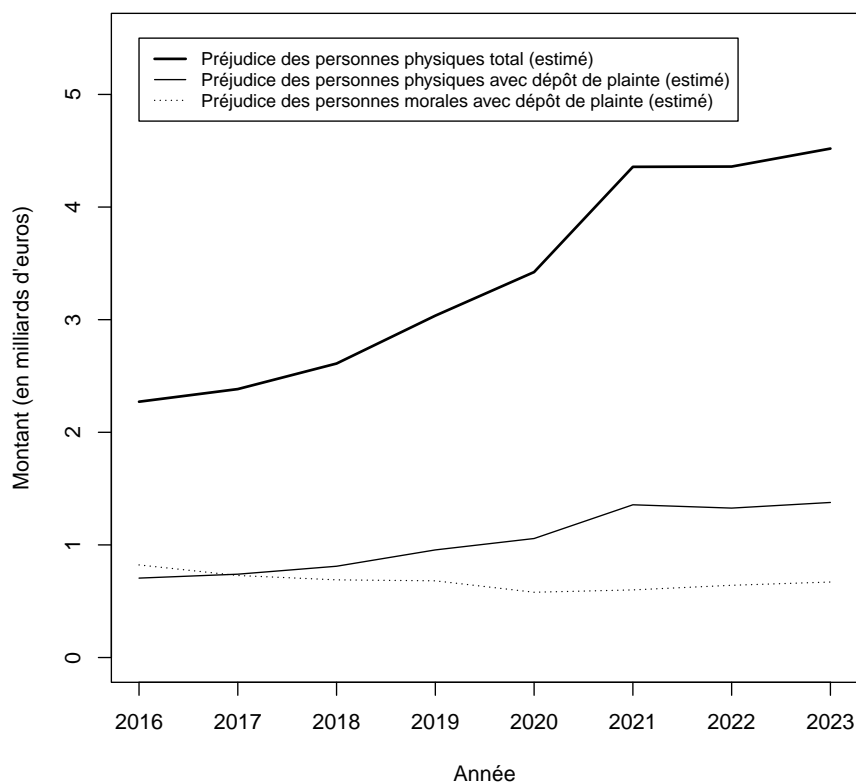


FIG. 1 – *Préjudice total (estimé)*

Champ : France métropolitaine.

Sources : Insee-ONDRP-SSMSI, enquêtes Cadre de vie et sécurité 2018 et 2019, et SSMSI, bases statistiques des infractions et des victimes de crimes et délits enregistrés par la police et la gendarmerie entre 2016 et 2023.

trouve pour les personnes physiques. Cette hypothèse est confortée par une étude de l'agence fédérale canadienne de statistiques. Cette dernière indique [3] que seulement 10% des entreprises canadiennes victimes d'atteintes cybercriminelles en 2021 avaient signalé l'atteinte à la police, pour un préjudice moyen de 53 500 dollars canadiens en cas de plainte et de 14 000 dollars canadiens sans plainte. Cela donne un ratio entre le montant du préjudice total des atteintes sans plainte et celui des atteintes avec dépôt de plainte d'environ : $(14\,000/53\,500) \times (90\%/10\%) \approx 2,36$. Il n'est cependant pas évident que de tels chiffres soient transposables à d'autres pays et d'autres champs contentieux. Ainsi, une autre étude trouve un autre ordre de grandeur. Il s'agit de l'enquête britannique Cyber Security Breaches Survey de 2023 dont les données sont disponibles sur le site UK Data service. Des questions de cette enquête portaient spécifiquement sur les escroqueries en ligne (catégorie qui, dans l'enquête, inclut les fraudes aux moyens de paiement telles que les usurpations d'identifiants bancaires) dont ont pu être victimes les entreprises et les associations (*charities*) britanniques répondantes. Les données de cette enquête indiquent un préjudice total neuf fois plus grand pour les escroqueries en ligne qui n'ont pas donné lieu à un signalement aux forces de l'ordre que pour celles qui ont été signalées. Notons cependant que le nombre de répondants qui déclarent avoir été victimes et avoir fait un signalement était très faible, et que de plus les auteurs de l'enquête mentionnent explicitement que les résultats des questions concernant les montants de préjudice sont à manier précautionneusement. Quoi qu'il en soit, en l'absence d'une enquête de victimation dédiée à cette population spécifique en France, il semble à l'heure actuelle difficile de

formuler une estimation fiable concernant le préjudice total des personnes morales.

Références

- [1] J.-F. BEAUMONT et L.-P. RIVEST : Dealing with outliers in survey data. *In Handbook of statistics*, vol. 29, p. 247–279. Elsevier, 2009.
- [2] P. BERTAIL, E. CHAUTRU et S. CLÉMENÇON : Tail index estimation based on survey data. *ESAIM: Probability and Statistics*, 19:28–59, 2015.
- [3] S. CANADA : Impact of cybercrime on canadian businesses. *The Daily*, October 18, 2022.
- [4] CONSUMER SENTINEL NETWORK DATA : *Consumer sentinel network data book*. Federal Trade Commission, Washington, DC., 2023.
- [5] S. G. CORREIA : Making the most of cybercrime and fraud crime report data: a case study of UK Action Fraud. *International Journal of Population Data Science*, 7(1), 2022.
- [6] COUR DES COMPTES : *La lutte contre les fraudes aux prestations sociales*. Communication à la Commission des affaires sociales du Sénat, 2021.
- [7] M. DEEVY et M. BEALS : *The scope of the problem - An overview of fraud prevalence measurement*. Stanford, California: Financial Fraud Research Center, 2013.
- [8] L. DUVERNET : Les escroqueries enregistrées par les services de sécurité entre 2016 et 2023. *Interstats Analyse n°68*, SSMSI, 2024.
- [9] W. FELLER : *An introduction to probability theory and its applications, Volume 2*, vol. 81. John Wiley & Sons, 1991.
- [10] I. GRAMA et V. SPOKOINY : Statistics of extremes by oracle estimation. *The Annals of Statistics*, 36(4):1619 – 1648, 2008.
- [11] P. HALL : Asymptotic properties of the bootstrap for heavy-tailed distributions. *The Annals of Probability*, p. 1342–1360, 1990.
- [12] E. HARRELL : *Victims of identity theft, 2018*. US Department of Justice, Office of Justice Programs, Bureau of Justice, 2019.
- [13] M. HEEKS, S. REED, M. TAFSIRI et S. PRINCE : The economic and social costs of crime second edition. *Home Off Res Report99*. <https://www.gov.uk/government/publications/the-economic-and-social-costs-of-crime>, 2018.
- [14] A. IMBERT et N. VIALANEIX : Décrire, prendre en compte, imputer et évaluer les valeurs manquantes dans les études statistiques: une revue des approches existantes. *Journal de la société française de statistique*, 159(2):1–55, 2018.
- [15] S. KEMP, F. MIRÓ-LLINARES et A. MONEVA : The dark figure and the cyber fraud rise in Europe: Evidence from Spain. *European Journal on Criminal Policy and Research*, 26(3):293–312, 2020.
- [16] C. B. LAKHDAR, N. LALAM et D. WEINBERGER : *L'argent de la drogue en France. Estimation des marchés des drogues illicites en France*. Rapport synthétique de la recherche Argent de la drogue à destination de la Mission Interministérielle de Lutte contre les Drogues et les Conduites Addictives (MILDECA), 2016.
- [17] M. LEVI et J. BURROWS : Measuring the impact of fraud in the UK: A conceptual and empirical journey. *The British Journal of Criminology*, 48(3):293–318, 2008.
- [18] R. E. MORGAN : *Financial fraud in the United States, 2017*. US Department of Justice, Office of Justice Programs, Bureau of Justice, 2021.
- [19] OBSERVATOIRE DE LA SÉCURITÉ DES MOYENS DE PAIEMENT : *Rapport annuel*. <https://www.banque-france.fr/stabilite-financiere/observatoire-de-la-securite-des-moyens-de-paiement>, 2023.
- [20] OFFICE FOR NATIONAL STATISTICS : *Nature of fraud and computer misuse in England and Wales: year ending March 2022*. <https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice>, 2022.

- [21] S. QUANTIN et C. WELTER-MÉDÉE : *Estimation des montants manquants de versements de TVA: exploitation des données du contrôle fiscal*. Insee, Institut national de la statistique et des études économiques, 2022.
- [22] S. I. RESNICK : *Heavy-tail phenomena: probabilistic and statistical modeling*. Springer Science & Business Media, 2007.
- [23] L. SALEMBIER : Les bases statistiques du SSMSI sur la délinquance enregistrée. *Interstats Méthode 26*, SSMSI, 2024.
- [24] C.-E. SÄRNDAL, B. SWENSSON et J. WRETMAN : *Model assisted survey sampling*. Springer Science & Business Media, 2003.
- [25] SSMSI : *Insécurité et victimation : les enseignements de l'enquête Cadre de vie et sécurité édition 2021*. 2022.

A Queues de distribution épaisses et indices de queue

Nous présentons ici quelques points saillants de la théorie mathématique des variables aléatoires à queue de distribution épaisse et de son application au présent document.

A.1 Définition

Soit X une variable aléatoire à valeurs strictement positives et définie sur un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$. De manière informelle, on dit que X a une distribution à queue épaisse s'il existe un nombre $\alpha > 0$ tel que

$$\mathbb{P}(X > x) \approx x^{-\alpha} \quad \text{quand } x \text{ prend des valeurs grandes.} \quad (2)$$

Le nombre $\alpha > 0$ est alors appelé l'indice de queue : plus il est proche de 0, plus la variable X peut prendre des valeurs extrêmement élevées avec une probabilité non négligeable.

De manière plus précise, la définition standard est la suivante : pour dire que X est à queue épaisse d'indice de queue $\alpha > 0$, on demande que pour tout $x > 0$,

$$\mathbb{P}(X > x) = x^{-\alpha} L(x),$$

où L est une fonction définie sur \mathbb{R}_+^* dite à variation lente, c'est-à-dire telle que pour tout $x > 0$,

$$\lim_{t \rightarrow +\infty} \frac{L(tx)}{L(t)} = 1.$$

Des exemples de telles fonctions à variation lente sont donnés par les fonctions qui ont une limite finie non nulle en $+\infty$, ou encore par la fonction $x \mapsto \log(x)$. Nous nous limiterons cependant ici à la définition intuitive (2), et nous référerons entre autres à [22] pour une présentation de la théorie des distributions à queue épaisse plus rigoureuse que le résumé que nous en proposons dans les paragraphes qui suivent.

Un des exemples le plus simples de telles distributions est donné par la famille des lois de Pareto pour lesquelles (2) devient

$$\mathbb{P}(X > x) = (x/x_m)^{-\alpha}, \quad \text{pour tout } x \geq x_m > 0.$$

D'autres exemples de lois classiques qui suivent (2) sont donnés par les familles des lois de Fréchet, de Cauchy, de Student, ou les lois stables dont le paramètre de stabilité est strictement inférieur à 2.

La table 10 présente l'espérance, la variance (lorsqu'elles existent) et les principaux quantiles théoriques de lois de Pareto selon différents indices de queue α . La valeur du paramètre x_m est choisie en fonction de α de sorte que la médiane de la loi soit égale à 500. On vérifie que plus l'indice de queue α est petit, plus la progression des quantiles les plus élevés s'accélère.

A.2 Moyenne, somme et variance

En partant de l'approximation (2), et en utilisant la formule classique pour les variables aléatoires positives

$$\mathbb{E}(X^p) = \int_0^{+\infty} x^{p-1} \mathbb{P}(X > x) dx,$$

pour laquelle on s'attend à ce que l'intégrale diverge en $+\infty$ dès que $p \geq \alpha$, on en déduit que les moments de X d'ordre $p \geq \alpha$ ne sont pas définis. En particulier, si α est plus petit que 2, X n'admet pas de variance, et si α est plus petit que 1, X n'admet pas d'espérance.

Indice de queue	0.5	0.8	1	1.5	2	3
E(X)				945	707	595
Var(X)						236 235
q0.25	222	301	333	382	408	437
q0.5	500	500	500	500	500	500
q0.75	2 000	1 189	1 000	794	707	630
q0.9	12 500	3 738	2 500	1 462	1 118	855
q0.99	1 250 000	66 479	25 000	6 786	3 536	1 842
q0.999	125 000 000	1 182 177	250 000	31 498	11 180	3 969

TAB. 10 – Valeurs théoriques d'une loi de Pareto de médiane 500

Note : La première ligne correspond à l'espérance de la loi, la deuxième à sa variance et les lignes suivantes aux quantiles. Les cases non renseignées correspondent à des valeurs infinies.

Lecture : une variable aléatoire qui suit une loi de Pareto de médiane 500 et d'indice de queue 0,5 a une probabilité 0,75 de prendre une valeur inférieure ou égale à 2 000.

Comment interpréter en pratique le fait que la loi d'une variable aléatoire n'admette pas d'espérance ? Cela peut correspondre notamment au fait que la moyenne empirique $\frac{1}{N} \sum_{i=1}^N X_i$ d'une suite de variables X_1, \dots, X_N devienne de plus en plus grande quand N croît au lieu de s'approcher d'une valeur limite finie : plus le nombre de variables augmente, plus la somme est influencée par des valeurs extrêmes de plus en plus élevée.

Rappelons de plus que selon le théorème central-limite, si une suite de variables aléatoires indépendantes et de même loi admettent une variance, alors leur somme se comporte comme une loi gaussienne. Si en revanche cette loi n'admet pas de variance, il ne peut y avoir de convergence gaussienne pour la somme (mais si cette loi vérifie (2) avec $0 < \alpha < 2$, alors la somme correctement normalisée converge en loi vers une loi à queue épaisse d'indice de queue α qui est appelée loi stable, cf. [9] section XVII.5). L'explication intuitive de ce comportement demeure la même : le régime de convergence se modifie en fonction de la prépondérance dans la somme des éléments les plus extrêmes, selon que ceux-ci sont négligeables au regard de la somme dans les cas à variance finie ($\alpha \geq 2$), ou au contraire selon que la somme devient de même ordre de grandeur que le ou les plus grands éléments quand α se rapproche de 0.

Lorsqu'on est confronté à des observations qui présentent des distributions à queue épaisse, il est donc crucial d'estimer le paramètre α pour savoir quel ensemble de procédures statistiques sont susceptibles de converger sur les données dont on dispose. Notamment, s'il semble crédible au vu des données qu'elles aient été générées selon des lois à queue épaisse d'indice de queue $\alpha < 2$, le champ des procédures applicables se trouve grandement restreint, et encore plus si $\alpha < 1$. Même des approches souvent appréciées pour leur généralité et leur caractère non-paramétrique comme le bootstrap rencontrent des difficultés spécifiques dans ce contexte [11].

A.3 Estimateurs de l'indice de queue

Bien estimer l'indice de queue est malheureusement un problème assez difficile, du fait notamment de ce que la relation (2) n'est supposée vraie qu'au voisinage de $+\infty$, ce qui suppose donc de se limiter à une petite part inconnue des observations les plus élevées pour en extraire l'information sur α .

Pour introduire les estimateurs classiques de α , notons $\bar{F} : x \mapsto \mathbb{P}(X > x)$ et \bar{F}^{-1} son inverse (ou son inverse généralisé si \bar{F} n'est pas bijective). Soient x et y deux réels positifs suffisamment grands pour que (2) s'applique, et soient $p = \bar{F}(x)$ et $q = \bar{F}(y)$. On a alors

$$\frac{\bar{F}(y)}{\bar{F}(x)} \approx (y/x)^{-\alpha},$$

ce qui se réécrit, en remplaçant respectivement x et y par $\bar{F}^{-1}(p)$ et $\bar{F}^{-1}(q)$, en

$$\frac{\bar{F}^{-1}(q)}{\bar{F}^{-1}(p)} \approx \left(\frac{q}{p}\right)^{-1/\alpha}$$

ou encore

$$\frac{\log \bar{F}^{-1}(q) - \log \bar{F}^{-1}(p)}{\log(q) - \log(p)} \approx -\frac{1}{\alpha}. \quad (3)$$

Supposons maintenant que nous disposons d'un n -échantillon X_1, \dots, X_n , et notons $X_{(1)}, \dots, X_{(n)}$ les variables ordonnées de manière décroissante :

$$X_{(1)} \geq \dots \geq X_{(n)}.$$

Alors $X_{(i)}$ est un estimateur naturel de $\bar{F}^{-1}(i/n)$, si bien que si (2) s'applique, le nuage des points d'abscisse $\log(i)$ et d'ordonnée $\log(X_{(i)})$ doit être à peu près aligné selon une pente proche de $-1/\alpha$, au moins pour les plus grandes observations, c'est-à-dire pour les petites valeurs de i . Les estimateurs classiques de l'indice de queue appelés estimateurs de Hill et de Pickands peuvent être interprétés comme des manières d'approximer cette pente.

L'estimateur de Hill du paramètre $1/\alpha$ est défini pour $1 \leq k \leq n-1$ par

$$H_{k,n} = \frac{1}{k} \sum_{i=1}^k \log \frac{X_{(i)}}{X_{(k+1)}}.$$

Cette définition correspond à l'estimateur du maximum de vraisemblance pour des modèles dérivés de la loi de Pareto. Elle découle également de (3) avec l'approximation

$$\sum_{i=1}^k \log \frac{k+1}{i} \approx k \quad \text{pour } k \rightarrow +\infty.$$

L'estimateur de Pickands du paramètre $1/\alpha$ est défini pour $1 \leq k \leq n/4$ par

$$P_{k,n} = \frac{1}{\log 2} \log \left(\frac{X_{(k)} - X_{(2k)}}{X_{(2k)} - X_{(4k)}} \right).$$

Cette définition peut elle aussi être vue comme une conséquence de (3).

Nous renvoyons au chapitre 4 de [22] pour une discussion à la fois des propriétés théoriques de ces deux estimateurs, et de leurs performances en pratique sur des données réelles ou simulées. Retenons notamment que le choix du nombre de cutoff k est un problème difficile car on veut à la fois avoir $k \rightarrow \infty$ et $k/n \rightarrow 0$: il s'agit d'utiliser le plus d'information possible pour limiter l'aléa (k grand) tout en s'assurant qu'on ne prend en compte que l'extrémité de la queue de la distribution (k/n petit). En pratique, lorsqu'on est confronté à un jeu de données, plutôt que de choisir d'entrée de jeu k et de calculer les valeurs correspondantes des estimateurs de Hill et de Pickands, une démarche standard est de calculer et de représenter graphiquement ces estimateurs pour toutes les valeurs possibles de k , et de chercher une zone où les estimateurs semblent à peu près constants. Malheureusement, il arrive qu'on soit dans l'incapacité d'observer de tels régimes, ce qui peut être dû notamment au fait que la distribution des observations, bien que satisfaisant (2) de manière asymptotique, ne permet pas de retrouver empiriquement cette relation même avec un nombre n d'observations conséquent [10].

A.4 Estimation des indices de queue sur les données de préjudice

Les figures 2 et 3 représentent les estimateurs de Hill et de Pickands obtenus sur les données de préjudices lisibles dans les données administratives en fonction du cutoff k retenu.

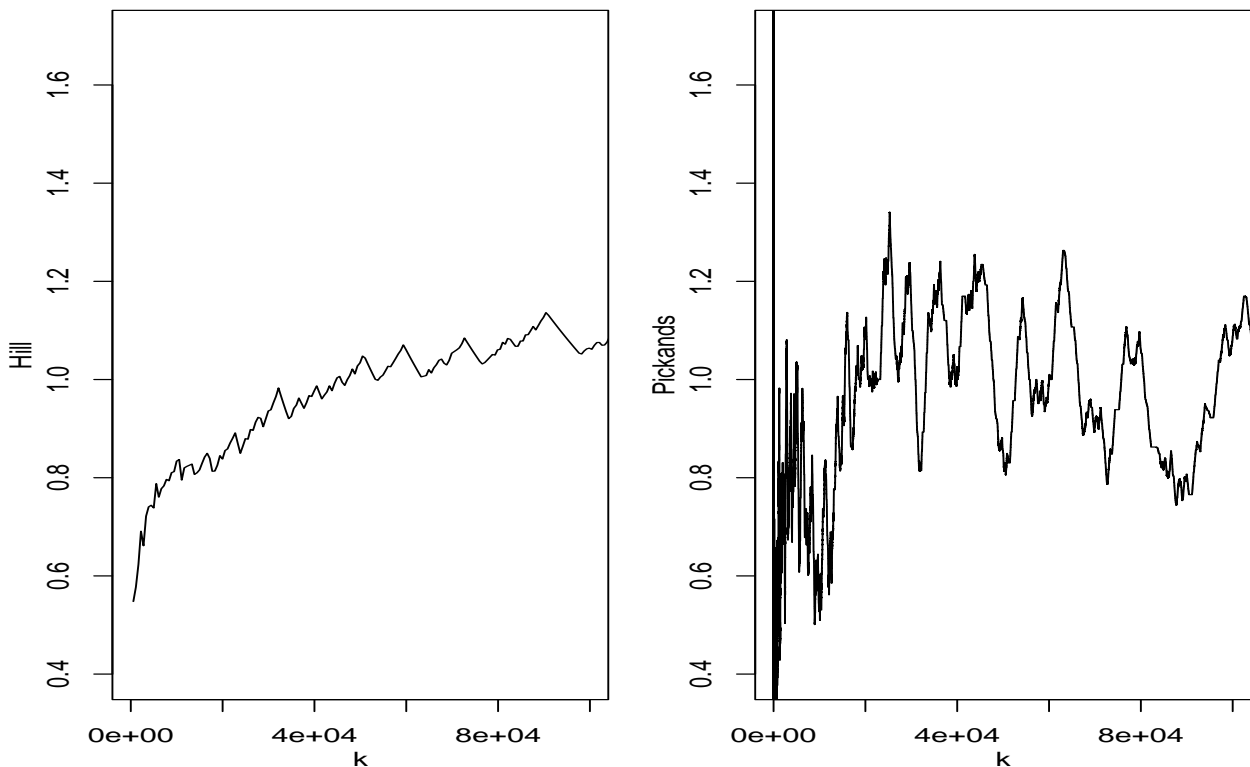


FIG. 2 – Valeur des estimateurs de Hill et de Pickands en fonction du paramètre k , données administratives, préjudice moyen par victime au sein des procédures où toutes les victimes sont des personnes physiques

Champ : France métropolitaine.

Source : SSMSI, bases statistiques des infractions et des victimes de crimes et délits enregistrés par la police et la gendarmerie entre 2016 et 2023.

Pour les préjudices des procédures où toutes les victimes sont des personnes physiques ($n \approx 550\,000$) les estimateurs de Hill et de Pickands prennent des valeurs autour de 0,8 et 1,2 environ, sans qu'il soit vraiment possible de trouver une gamme de valeurs de k entre 1 et 100 000 pour lesquelles les estimateurs soient clairement constants. Cela correspondrait cependant à un indice de queue α proche de 1. Pour les préjudices des procédures avec au moins une victimes personne morale ($n \approx 53\,000$) les estimateurs de Hill et de Pickands prennent des valeurs autour de 0,8 et 1,4 environ. Pour k entre 4 000 et 8 000 environ, les estimateurs prennent des valeurs proches de 1,2, et l'estimateur de Pickands varie peu autour de cette valeur. Cela correspondrait à un indice de queue α proche de $1/1,2 \approx 0,83$. La queue de distribution du préjudice des personnes morales semble ainsi légèrement plus épaisse que celle du préjudice des personnes physiques. Dans les deux cas, les valeurs des estimateurs renvoient à des distributions sans variance, et peut-être sans espérance, particulièrement en ce qui concerne le préjudice subi par les personnes morales.

Nous présentons également dans la figure 4 les valeurs des estimateurs de Hill et de Pickands sur les données CVS en ce qui concerne les préjudices des atteintes qui n'ont pas donné lieu à un dépôt de plainte. Pour l'estimateur de Hill, nous utilisons la version proposée par [2] qui permet de prendre en compte la pondération du plan de sondage. Vu le faible nombre de données ($n \approx 1650$), il est peu crédible que les estimateurs soient dans des zones de convergence. Les graphiques sont cependant compatibles avec des distributions d'indice de queue entre 1 et 2, l'estimateur de $1/\alpha$

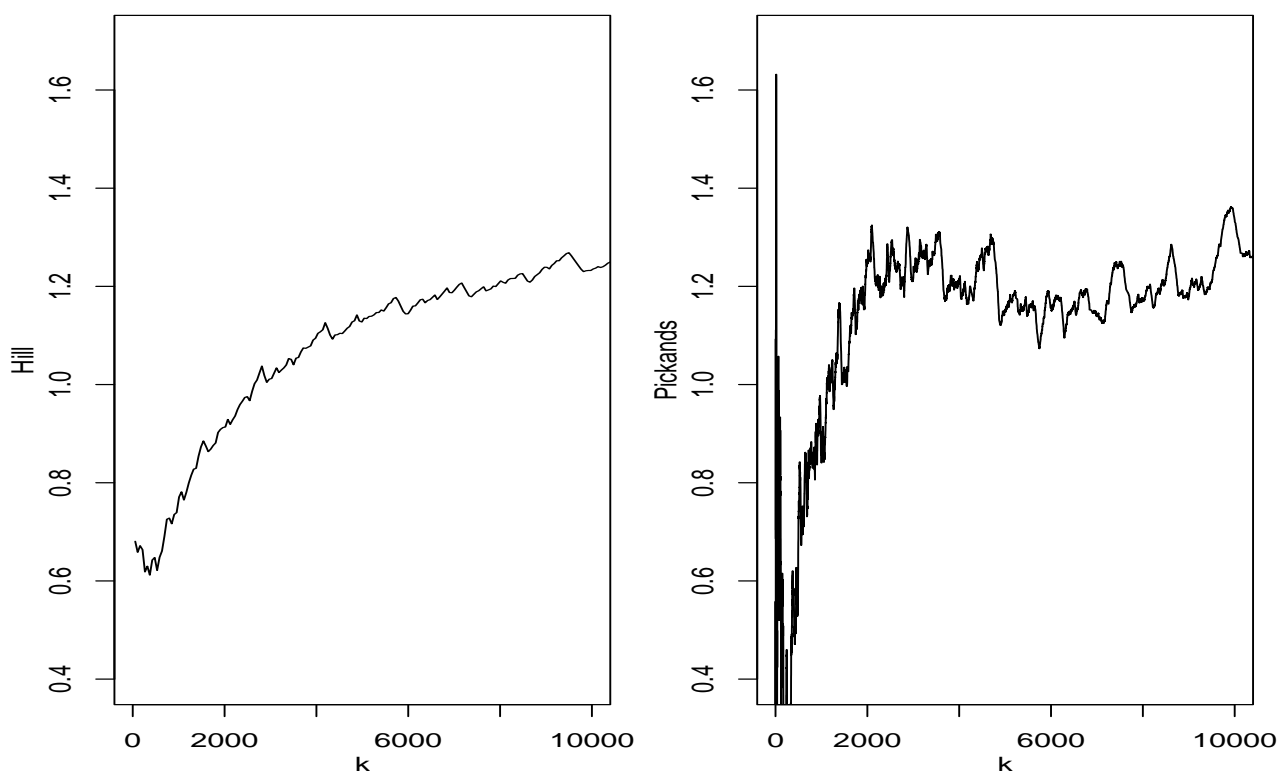


FIG. 3 – Valeur des estimateurs de Hill et de Pickands en fonction de la valeur de k , données administratives, préjudice moyen par victime au sein des procédures où au moins une victime est une personne morale

Champ : France métropolitaine.

Source : SSMSI, bases statistiques des infractions et des victimes de crimes et délits enregistrés par la police et la gendarmerie entre 2016 et 2023.

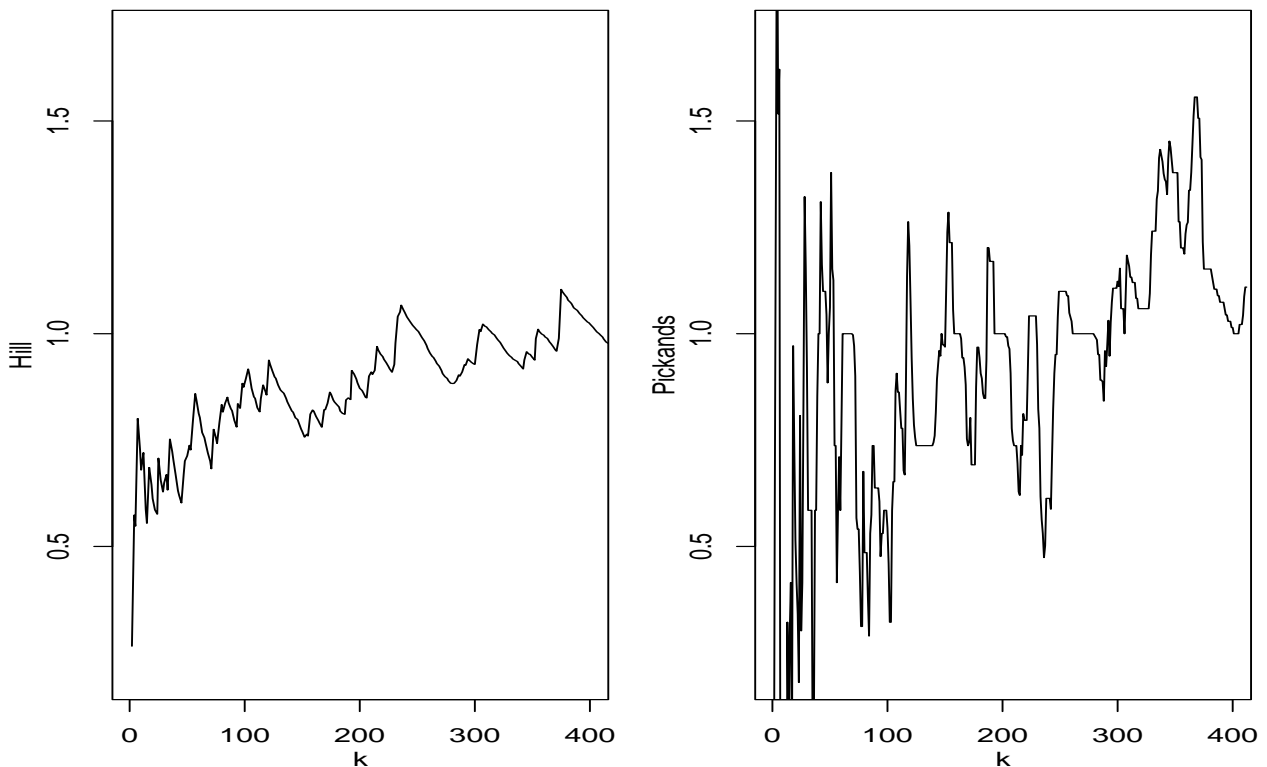


FIG. 4 – Valeur des estimateurs de Hill et de Pickands en fonction du paramètre k , données CVS, montant des atteintes qui n'ont pas donné lieu à un dépôt de plainte

Champ : France métropolitaine.

Source : Insee-ONDRP-SSMSI, enquêtes Cadre de vie et sécurité 2018 et 2019

restant à peu près compris entre 0,5 et 1 pour les petites valeurs de k . Cela correspondrait ainsi à des queues de distribution moins épaisses que celles des préjudices présents dans les données administratives, et donc ayant donné lieu à un dépôt de plainte.

A.5 Sondages et distributions à queue épaisse

Dans le cadre de la théorie des sondages, on cherche généralement à reconstruire la moyenne ou le total d'une grandeur sur toute une population à partir de l'observation d'un petit échantillon. Cependant, si la distribution de la grandeur dans la population correspond à une distribution à queue épaisse (par exemple si dans le cadre d'une modélisation de super-population, on fait l'hypothèse que cette grandeur a été générée selon une telle distribution), alors les approximations de la théorie des sondages sont à manier avec prudence, ce que nous illustrons par des simulations dans la figure 5.

Pour chacune des lois de Pareto de la table 10, nous avons simulé $N = 250\,000$ variables X_1, \dots, X_N . Nous nous sommes placés dans le cadre de l'estimation de la somme $S = \sum_{i=1}^N X_i$ par un sondage aléatoire simple. C'est-à-dire que pour $n = 62\,500 = N/4$ ou $n = 1\,500 \approx N/167$, nous avons tiré 1 000 échantillons Z_1, \dots, Z_n sans remise de taille n au sein des valeurs X_1, \dots, X_N , et nous avons calculé l'estimation classique \hat{S} de S donnée par $\frac{N}{n} \sum_{i=1}^n Z_i$. Le cas $n = 62\,500$ correspond en ordre de grandeur à celui des données administratives, pour lesquelles sur les quelques centaines de milliers de procédures annuelles, environ un quart des montants sont lisibles. Le cas $n = 1\,500$

renvoie aux données d'enquête pour lesquelles nous disposons d'un nombre comparable de données sur lesquelles estimer la distribution du montant des préjudices qui n'ont pas donné lieu à un dépôt de plainte.

Nous représentons dans la figure 5 la densité de \hat{S} estimée à partir des 1 000 échantillons que nous avons obtenus. On peut voir que \hat{S} a une distribution asymétrique avec une médiane et un mode en dessous de la vraie valeur S , ce qui est d'autant plus marqué que l'indice de queue α se rapproche de 0 et que la taille de l'échantillon n est petite devant la taille N de départ. La raison de cela est intuitive : si l'échantillon Z_1, \dots, Z_n ne contient qu'une petite fraction des valeurs X_1, \dots, X_N , il est fréquent qu'il ne prenne pas en compte les valeurs les plus élevées, lesquelles sont d'autant plus prédominantes dans la somme S que α est petit.

Ainsi, quand α vaut 1 (cas qui semble proche de la distribution des montants de préjudice ayant donné lieu à dépôt de plainte, cf. section précédente), un sondage avec $n = 1\,500$ a souvent tendance à fortement sous-estimer le total S de la population : notamment le mode de \hat{S} est presque deux fois inférieur à la vraie valeur. Ce phénomène est nettement atténué lorsqu'une proportion plus élevée de la population est sondée. Lorsque α augmente (cas qui semble proche de la distribution des montants de préjudice n'ayant pas donné lieu à dépôt de plainte), les performances des sondages s'améliorent et la distribution de l'estimateur \hat{S} se rapproche d'une gaussienne. En revanche, si α est trop faible, d'ordre 0,5, il devient impossible de reconstruire S même en sondant un quart de la population.

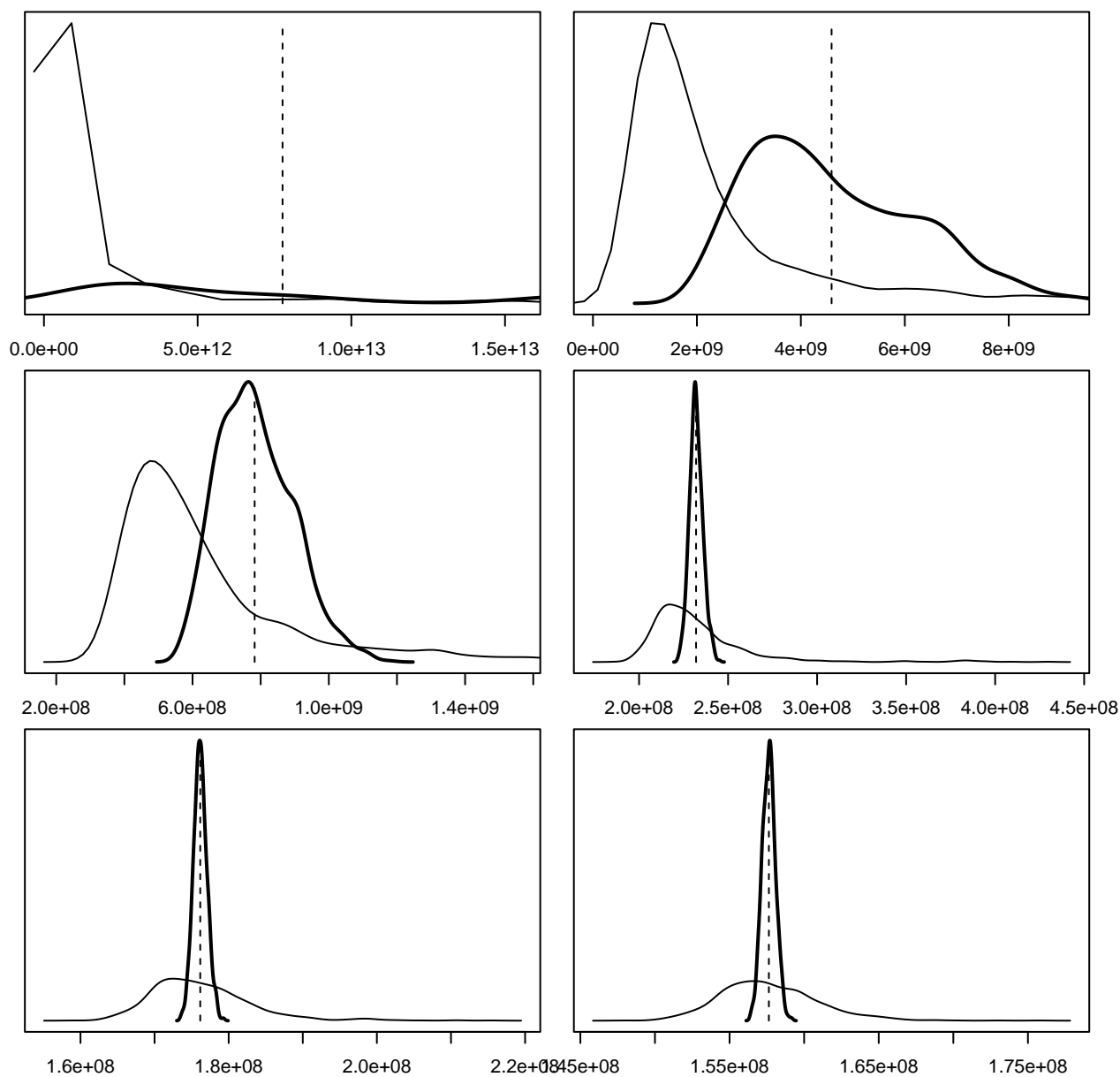


FIG. 5 – Densité empirique (sur 1 000 échantillons sans remise) de l'estimateur par sondage d'une somme S de $N = 250\,000$ variables de Pareto.

Note : de gauche à droite et de haut en bas, α vaut 0,5, 0,8, 1, 1,5, 2 et 3. Le trait pointillé correspond à la vraie valeur de S , le trait gras correspond à l'estimation de S à partir d'échantillons de taille 62 500, le trait fin correspond à l'estimation de S à partir d'échantillons de taille 1 500.