Journées de méthodologie statistique de l'Insee 2025

COLLECTER ET DIFFUSER LES DONNÉES SENSIBLES : ENJEUX ET QUESTIONNEMENTS

Olivier Lefebvre (*), Eric Rancourt (**), Albane Gourdol (***), Samuel Givois (****), Kévin Milin (*****)

(*) Insee, Direction des statistiques démographiques et sociales
(**) Statistique Canada
(***) INED
(****) Service statistique ministériel de la Sécurité Intérieure
(*****) Insee, Département de l'Action Régionale

Mots-clés *:* Données sensibles, éthique, collecte, protection des données, diffusion
Domaine concerné : Éthique

Résumé

L'article vise à partager des éléments de réflexion et des exemples de pratiques au sujet des données sensibles. Le sujet n'est pas étranger aux considérations de méthodologie statistique car, les méthodologies de test, de collecte, de traitement et de valorisation des données doivent être adaptées à ces données particulières. Par ailleurs, même si les progrès techniques offrent des possibilités accrues en matière de collecte et de traitement il convient de les utiliser avec discernement, de manière adaptée à la sensibilité des données. Sur le plan technique comme sur le plan juridique, pouvoir n'est pas devoir.

Pourquoi considère-t-on une donnée comme sensible ? Il peut s'agir de critères juridiques, culturels, ou liés au recueil de l'information (il peut être difficile pour les enquêtés d'évoquer le sujet) ; une donnée peut également être sensible car elle touche à des sujets potentiellement polémiques ou clivants ; cela jouera sur la manière de la collecter, mais aussi de la diffuser et de la commenter.

Le fait que des données ou des sujets soient sensibles n'interdit pas de les collecter, mais impose des conditions de recueil et de traitements spécifiques, fixées par la loi ou résultant de considérations éthiques.

La décision de collecter ces données, et éventuellement de les combiner à d'autres, doit relever d'une réflexion approfondie où sont pris en compte plusieurs facteurs : le bénéfice pour la collectivité de l'information produite, (via l'éclairage du débat public, l'aide à la décision ou la recherche publique), leur sensibilité et l'acceptabilité sociale d'un tel traitement, la capacité à protéger les données collectées tout au long de leur cycle de vie.

Parfois, une donnée est sensible parce qu'elle porte sur des faits ou situations douloureuses, ou fait référence à des pratiques illicites. Les protocoles de collecte doivent être adaptés à des situations, qu'il s'agisse de l'échantillonnage, du questionnaire en lui-même ou du mode d'administration. Les phases de conception et de test sont cruciales, les partenaires de l'opération et les enquêteurs, notamment lors des tests y jouent un rôle déterminant. Les enquêteurs doivent être formés à la collecte et accompagnés pour cela.

Une fois les données collectées, leur protection doit être à la mesure des enjeux, compte tenu de leur sensibilité et des conséquences possibles d'une rupture de confidentialité.

Lors de la diffusion et de l'analyse, une attention toute particulière doit être portée à la documentation et à la précision des termes, à la significativité des données et de leurs évolutions, mais aussi à leur contextualisation. L'enjeu est que les usages soient fondés sur des bases claires, solides et partagées, chacun étant, bien entendu, libre des conclusions qu'il en tire.

Sur tout le cycle de vie de la donnée, la transparence doit être une exigence forte et partagée, c'est une condition essentielle à la fois pour des motifs d'acceptabilité (acceptabilité sociale, mais aussi acceptabilité par les personnes directement concernées), de qualité du recueil et des traitements, et d'utilisation.

Le fait que des données ou des sujets soient sensibles n'interdit pas de les collecter, mais impose des conditions de recueil et de traitements spécifiques, fixées par la loi ou résultant de considérations éthiques.
La décision de collecter ces données, et éventuellement de les combiner à d'autres, doit relever d'une réflexion approfondie où sont pris en compte plusieurs facteurs : le bénéfice pour la collectivité de l'information produite, (via l'éclairage du débat public, l'aide à la décision ou la recherche publique), leur sensibilité et l'acceptabilité sociale d'un tel traitement, la capacité à protéger les données collectées tout au long de leur cycle de vie.
Parfois, une donnée est sensible parce qu'elle porte sur des faits ou situations douloureuses, ou fait référence à des pratiques illicites. Les protocoles de collecte doivent être adaptés à des situations, qu'il s'agisse de l'échantillonnage, du questionnaire en lui-même ou du mode d'administration. Les phases de conception et de test sont cruciales, les partenaires de l'opération et les enquêteurs, notamment lors des tests y jouent un rôle déterminant. Les enquêteurs doivent être formés à la collecte et accompagnés pour cela.
Une fois les données collectées, leur protection doit être à la mesure des enjeux, compte tenu de leur sensibilité et des conséquences possibles d'une rupture de confidentialité.
Lors de la diffusion et de l'analyse, une attention toute particulière doit être portée à la documentation et à la précision des termes, à la significativité des données et de leurs évolutions, mais aussi à leur contextualisation. L'enjeu est que les usages soient fondés sur des bases claires, solides et partagées, chacun étant, bien entendu, libre des conclusions qu'il en tire.
Sur tout le cycle de vie de la donnée, la transparence doit être une exigence forte et partagée, c'est une condition essentielle à la fois pour des motifs d'acceptabilité (acceptabilité sociale, mais aussi acceptabilité par les personnes directement concernées), de qualité du recueil et des traitements, et d'utilisation.

Abstract

The article aims to share insights and examples of practices relating to sensitive data, from several institutions : two NSIs (Insee and Statistic Canada), one institution of research (INED), a Ministerial Statistical Service (Interior Security).
The subject is not unrelated to considerations of statistical methodology, as the methodologies used to test, collect, process, and exploit data must be adapted to these particular data. Furthermore, even though technical advances offer increased possibilities for collection and processing, they must be used judiciously, in a manner appropriate to the sensitivity of the data. From both a technical and legal standpoint, power does not equal duty.

1. Pourquoi une session sur les données sensibles au cours des journées de méthodologie statistique ?

Parce que ces données méritent une attention particulière, du fait justement de leur caractère sensible. Ce n'est pas parce qu'une donnée est sensible qu'il faut s'interdire de la collecter, mais cela appelle des dispositions spécifiques.

Parce que le mouvement général d'expansion des données concerne aussi les données sensibles : de nouvelles données sensibles peuvent donc émerger, qui appellent des questionnements spécifiques.

Parce que certaines d'entre elles correspondent à des questions sur lesquels des éclairages sont nécessaires : des thématiques plus précises et des questions plus ciblées amènent de nouvelles

questions dans des domaines non traditionnels pour les INS.

Parce qu'elles conduisent à des défis méthodologiques liés aux enjeux déontologiques ou éthiques, à différents niveaux de la chaîne de traitement. Il s'agit de développer et tester de nouveaux protocoles de collecte, de nouveaux traitements, des accompagnements de résultats adaptés à la sensibilité des données traitées ou des questions qu'elles éclairent.

Parce qu'il convient aussi d'avoir conscience que, si les progrès techniques et méthodologiques permettent, en théorie, de collecter davantage de données et de mieux les protéger, cela nous confère une capacité mais pas une légitimité. En d'autres termes, cela ne nous exonère pas de nous poser la question de la nécessité et de la légitimité de ce traitement. **Sur le plan technique comme sur le plan juridique, pouvoir n'est pas devoir.**

Parce que la notion est multi-dimensionnelle, à géométrie variable, dépendante du contexte (national, historique, culturel, actualité…) et mérite d'être précisée pour d'une part, bien savoir de quoi on parle et faciliter les échanges entre les différents acteurs, d'autre part construire les solutions adaptées aux différents enjeux.

Enfin, parce que sur de tels sujets les partages de points de vue, d'expériences, de questionnements issus d'horizons divers permettent de progresser.

2. Comment définir les données sensibles ?

Définir les données sensibles n'est pas chose aisée, tant la sensibilité peut s'exercer dans des dimensions différentes. Il n'existe pas de définition universelle et plusieurs approches sont possibles pour appréhender cette notion.

Une approche juridique, centrée sur la protection des données

Pour les données à caractère personnel, il existe une approche juridique. En France, elle est portée par le Règlement général pour la protection des données (RGPD) et la Loi Informatique et Libertés. Elle distingue certaines catégories de données, en fonction des thèmes dont elles relèvent : santé, religion, opinions politiques, orientation sexuelle …

Quelle que soit la façon dont on en apprécie la sensibilité, le traitement des données personnelles doit s'inscrire dans le cadre du RGPD.

Celui-ci inscrit dans son article 5 (cf extrait ci-dessous) quelques règles fondamentales qui s'appliquent au traitement de données à caractère personnel, notamment en matière de transparence, de légitimité au regard de finalités déterminées, de pertinence et de minimisation (ne pas collecter ni traiter plus que nécessaire au regard de la finalité poursuivie) :

Les données à caractère personnel doivent être:
a) ***traitées de manière licite, loyale et transparente*** *au regard de la personne concernée (licéité, loyauté, transparence);*
b) collectées pour des ***finalités déterminées, explicites et légitimes****, et ne pas être traitées ultérieurement d'une manière incompatible avec ces finalités ; le traitement ultérieur à des fins archivistiques dans l'intérêt public, à des fins de recherche scientifique ou historique ou à des fins statistiques n'est pas considéré, conformément à l'article 89, paragraphe 1, comme incompatible avec les finalités initiales (limitation des finalités);*
c) ***adéquates, pertinentes et limitées*** *à ce qui est nécessaire au regard des finalités pour lesquelles elles sont traitées (minimisation des données) ;*

Le RGPD distingue plusieurs catégories de données, qu'il soumet à des restrictions particulières :

L'article 9 est relatif aux traitements portant sur des catégories particulières de données à caractère personnel. Ces données sont qualifiées de « données sensibles » par la CNIL (Commission nationale de l'informatique et des libertés).
Son paragraphe 1 liste les thématiques concernées : origine raciale ou ethnique, opinions politiques ou appartenance syndicale, convictions religieuses ou philosophiques, vie sexuelle ou orientation sexuelle, traitement des données génétiques, des données biométriques, des données de santé.
Il indique que les traitements portant sur ces données sont interdits.
Son paragraphe 2 mentionne les exceptions à cette interdiction, dont font partie les traitements statistiques et les traitements de recherche scientifique ou historique.

L'article 10 est relatif aux traitements portant sur les condamnations pénales et infractions
Ceux-ci ne peuvent être effectués que sous le contrôle de l'autorité publique, ou si le traitement est autorisé par le droit de l'Union ou par le droit d'un État membre qui prévoit des garanties appropriées pour les droits et libertés des personnes concernées .

Contrairement à ce que l'on entend parfois, le traitement de ces données à des fins statistiques ou de recherche est donc autorisé, avec un encadrement strict, et possiblement, obligation d'une étude d'impact sur la protection des données. Pour autant, il convient de s'interroger sur le périmètre de la collecte et du traitement de telles données, au regard d'autres critères : acceptation par les personnes concernées ou par la société, qualité de l'information recueillie, tout en s'inscrivant dans les principes du RGPD, notamment la nécessité et la proportionnalité du traitement.

Au-delà des textes …

Mais au-delà de cette liste de données sensibles, le responsable du traitement, en lien avec les autres parties prenantes (utilisateurs, Cnis - Conseil national de l'information statistique -, représentants des personnes concernées) peut considérer que des données non sensibles au sens du RGPD (données de revenus par exemple) méritent également une attention particulière en matière de protection.

Une donnée sensible peut également l'être parce qu'elle touche à l'intime (avec qui on vit, par exemple) ou à des choses difficiles à évoquer (ruptures, événements difficiles ou illicites, mauvais traitements …) ; il faut alors se poser la question de ce que l'on peut collecter, comment on peut le collecter, et intégrer cette dimension dès le protocole de collecte.

Il y a aussi des sujets sensibles, car se prêtant à polémique, parfois uniquement en fonction du contexte économique, social ou politique du moment. Cela peut rendre les choses difficiles en collecte mais aussi au stade de la diffusion : comment limiter le risque de mauvaise interprétation ?

De la donnée sensible au jeu de données sensible

Un jeu de données porte sur plusieurs entités du champ d'observation (toutes ou un échantillon) et rassemble plusieurs variables (identifiants, caractéristiques). La question de la sensibilité des jeux de données se pose, notamment au regard de sa sécurisation informatique (gestion des accès, niveau de protection) et de son ouverture (plus un jeu de données est jugé sensible, plus il convient d'en contrôler l'ouverture).
La sensibilité du jeu de données va dépendre de son contenu en matière de variables (identifiants, variables ou informations sensibles…), de son caractère exhaustif ou non, mais aussi des possibilités de ré-identification des personnes, en croisant plusieurs variables.
Le rapprochement de deux données « non sensibles » au sein d'un même jeu de données peut apporter une information sensible. Par exemple, pris isolément la taille et le poids d'un individu ne sont pas sensibles, alors que rapprochés, ils apportent une information sur l'état de santé : obésité ou extrême maigreur).

À l'inverse, on peut « désensibiliser un jeu de données » en supprimant des variables correspondant à des données sensibles, ou en réduisant les risques de ré-identification : aller de la pseudonymisation (suppression des identifiants directs) à l'anonymisation (suppression de toute possibilité de ré-identification, même indirecte).

Quelques exemples de pratique pour l'Insee

Une attention particulière est apportée au respect des principes de nécessité, minimisation et proportionnalité lors de la conception de nos opérations, notamment lorsque cela implique des données sensibles ou des sujets sensibles.

Une concertation approfondie, en s'appuyant notamment sur le Cnis, en amont de la collecte de données sensibles (par exemple pour l'enquête TeO - Trajectoires et Origines) ou de la construction de dispositifs potentiellement sensibles (le Répertoire Statistique des Individus et des Logements et les services d'appariement de données à caractère personnel - Resil).

Une grille d'aide à la caractérisation de la sensibilité des jeux de données, permettant aux responsables de ces jeux de données de prendre les bonnes décisions en matière de protection informatique et d'ouverture a été proposée en interne. Les critères de classification sont les suivants : présence ou non d'identifiants directs, risques de ré-identification, respect du secret statistique pour les données agrégées

La vigilance sur la diffusion des résultats (prévenir les risques de mauvaise interprétation ou détournement de message initial), réagir publiquement si nécessaire en cas de manquement, répondre aux demandes de « fact checking »..

3. Comment le sujet est-il abordé au Canada ?

Les éléments d'enjeux et de contexte rappelés au point 1 sont également pertinents pour le Canada.
L'activité de Statistique Canada, comme celle des autres INS s'inscrit dans un cadre législatif relatif à la production de données statistiques et à la protection des données personnelles, ce cadre ne fait pas tout : pouvoir n'est pas devoir.
Par ailleurs, la révolution des données a augmenté la conscience collective des données et de leur valeur, ainsi que la masse des données produites et potentiellement utilisables. La pandémie a agi comme catalyseur d'utilisation des données, à la fois parce qu'elle apportait des nouvelles questions et parce qu'elle bouleversait les modes de recueil des données.
Au-delà de la pandémie, des thématiques plus précises et des questions plus ciblées amènent de nouvelles questions, pour certaines sensibles. Il convient également de mentionner que les brèches dans certaines organisations non-statistiques ont agi comme frein à l'utilisation des données, en suscitant un débat dans des conditions non sereines.

Il n'y a pas de définition claire et acceptée des données sensibles. Cependant, la sensibilité des données s'apprécie en lien avec :

- le domaine en question,
- les lois (en gardant en tête le fait que pouvoir n'est pas devoir),
- la protection de la vie privée et la confidentialité, en croisant le risque et l'impact d'une rupture de protection de la vie privée ou de la confidentialité,
- l'éthique et l'acceptabilité sociale du recueil et du traitement de telles données.

Exemples de données sensibles dont le traitement a été envisagé à Statistique Canada

- Enquête sur les tests sanguins de la COVID 19

- Impact économique des gagnants de loterie
- Composante sur la maltraitance de l'enquête sur les enfants
- Utilisation potentielle de micro-données bancaires
- Positionnement selon les téléphones cellulaires
- Moissonnage des réseaux sociaux

Note: La plupart de ces cas n'ont pas été implantés, suite à des polémiques survenues très en amont de la démarche (micro-données bancaires) ou à l'analyse en opportunité réalisée à Statistique Canada.

La question de la sensibilité s'inscrit dans la gouvernance de la donnée, aux différents stades du cycle de vie, de la donnée collectée à l'information diffusée, avec comme étapes-clés les 4 P : Procurer, Protéger, Préparer, Partager (ou 4 G en Anglais : Gather, Guard, Grow, Give) tels que décrits dans (Rancourt,2019).

A l'étape « partager », les décisions d'ouverture ou de restriction d'accès sont prises en fonction de la cotation du jeu de données selon un score OCC (Outil de Classification de la Confidentialité), croisant le risque de rupture de confidentialité (ou de stigmatisation d'un groupe social) et la sensibilité. Le score va de 0 à 9, plus le niveau est élevé, plus l'accès est restreint.
NB : l'Insee a suivi la même logique pour sa grille d'aide à caractérisation de la sensibilité des données.

En amont de l'étape « procurer », l'opportunité de l'opération, au regard de la sensibilité des données traitées, est appréciée à l'aune de différents critères, par des experts en éthique et en protection de la vie privée :

- Nécessité et bénéfices pour les Canadiens
- Niveau de détail et fréquence du recueil de données
- Niveau de qualité des données
- Sources et approches alternatives à ce traitement
- Objectifs d'appariements de données
- Plans de diffusion

La démarche s'appuie sur le cadre de nécessité et de proportionnalité, développé à Statistique Canada et fondé sur une démarche scientifique permettant d'analyser rigoureusement la question, et d'en rendre compte de manière transparente.

Le traitement des données sensibles s'inscrit dans la démarche générale d'éthique des données de Statistique Canada tirés du cadre développé par Marcovitch et Rancourt (2022), avec 5 axes majeurs : une culture commune supportée par la gestion supérieure, l'inscription dans les processus de traitement incluant une revue de la sensibilité, une gouvernance claire et adaptée avec rôles et responsabilités bien définies, des normes établies et respectées, intégrant les dimensions d'éthique des données, et enfin une transparence proactive non seulement de description mais également de plans et activités à venir.

La transparence est indispensable, plus encore lorsqu'il s'agit de données sensibles, pour lesquelles il s'agit de rassurer les citoyens, de partager les enjeux et d'expliquer les décisions prises en matière de collecte, traitement, protection des données, mais aussi d'assurer la meilleure utilisation possible de l'information produite, d'éclairer le débat public et de prévenir les polémiques.
Elle s'inscrit dans toutes les étapes du processus et doit viser toutes les parties prenantes.

« Un institut national de statistique est transparent lorsque les approches, méthodes, décisions, et informations (de toutes les étapes du cycle de vie des données) sont rendues disponibles aux utilisateurs, chercheurs, collaborateurs et citoyens. » (National Academies, 2022).

En conclusion, la gestion des données sensibles recoupe plusieurs dimensions et champs d'expertise. À Statistique Canada, plusieurs outils ont été développés et mis en place, et le souci de la sensibilité et

l'éthique des données font maintenant partie de la culture. C'est bien l'affaire de tous et non d'un seul groupe ou comité.

4. Collecter des données sensibles : les pratiques de l'INED

De par ses missions et ses champs d'intérêt, l'INED est très souvent conduit à mener des opérations statistiques spécifiques. Ses enquêtes portent souvent sur des thématiques sensibles et jugées difficiles à mesurer scientifiquement telles que la fin de vie, les violences, la sexualité, les inégalités (en particulier de genre), etc. Elles sont généralement menées auprès de populations en situation vulnérable, difficiles à contacter ou à interroger, ou très spécifiques comme les sans-abri, des personnes illettrées, des jeunes adultes suivis par l'aide sociale à l'enfance, des détenus, des usagers de drogue, des migrants. L'INED suit par ailleurs des cohortes d'enfants, qui représentent un public particulier.

Les projets de recherche exigent donc à chaque fois des dispositifs d'enquêtes innovants, qui se révèlent souvent pionniers et qui tiennent compte de la sensibilité des sujets traités et des populations enquêtées.

Si le RGPD pose un principe général d'interdiction de recueil ou de traitement des données sensibles, il prévoit des dérogations à cette interdiction (alinéa 2 de l'article 9), et la CNIL clarifie les conditions dans lesquelles il est possible de les collecter et de les utiliser pour la recherche scientifique. .
La Déléguée à la Protection des Données de l'INED joue un rôle essentiel dans l'appréciation et la mise en œuvre de ces traitements. Elle veille à leur conformité, en lien avec les concepteurs, et autorise in fine les traitements mobilisant des données sensibles. Ces traitements sont inscrits au registre des activités de traitement de l'INED et font l'objet d'une Analyse d'Impact relative à la Protection des Données (AIPD).

Données sensibles et recherche – un indispensable regard éthique complétant le regard juridique

Lorsque l'on collecte des données sensibles, la capacité juridique ne suffit pas. Il convient de s'interroger sur d'autres dimensions, relevant de considérations éthiques, qu'il s'agisse de la décision de faire ou ne pas faire l'enquête, que des conditions dans lesquelles on va la mener (information des enquêtés, protection des personnes intervenant dans le processus…).

C'est notamment le cas quand l'enquête suppose la révélation de situations ou activités illicites des enquêtés (statut migratoire, consommation ou trafic de drogues, violences, actes médicaux interdits, …) susceptible de mettre en danger les enquêtés ou leur entourage, d'exposer les enquêteurs à des situations difficiles, ou de poser des problèmes juridiques spécifiques (obligation de signalement, secret professionnel et confidentialité).

Le comité d'éthique de l'INED peut alors être saisi, pour un examen de toute question relative aux aspects éthiques de la recherche, tel que le degré d'information à donner aux enquêtés dans les enquêtes sensibles ou bien la responsabilité des enquêteurs en cas de recueil de faits susceptibles d'être délictueux, notamment au regard des obligations de signalement.

Ces enjeux éthiques conduisent à l'élaboration de protocoles de recueil et de traitement adaptés. Par exemple, une enquête sur la fin de vie a été réalisée par l'INED en métropole en 2010 et dans les Dom en 2020-2022. Elle interrogeait des médecins sur des pratiques punies par la loi. Un protocole très sécurisé a été mis en place pour qu'il soit impossible de rattacher un questionnaire à un médecin, garantissant l'anonymat du médecin, de la personne décédée et le secret médical. Par ce protocole spécifique, les parties prenantes ont été protégées tout en garantissant la sincérité et la pertinence/'exactitude des réponses obtenues.

L'INED mène de nombreux projets de recherches recueillant des données dites sensibles. Certains points méritent une attention particulière lors de la conception de ces projets : le questionnement, l'échantillonnage, les conditions de passation du questionnaire, le rôle des enquêteurs pour n'en citer que quelques-uns.
Il n'existe pas, bien évidemment, de solution universelle, chaque enquête, chaque public, appelant une réponse adaptée. On peut néanmoins, à travers l'exemple de quatre enquêtes, portant sur des sujets délicats, voire douloureux, ou susceptibles de révéler des comportements ou situations illicites, présenter des points d'attention et les réponses apportées.

Les enquêtes prises en exemple

ENVIE – enquête sur la vie affective des jeunes adultes
→Mieux comprendre les évolutions de la vie affective et sexuelles des jeunes adultes : diversité des relations intimes, modes de rencontre, expériences du célibat, manières de faire couple, expériences propres aux personnes LGBT

AMP sf - assistance médicale à la procréation sans frontière
Analyser les recours à l'aide à la procréation de personnes résidant en France qui se font en dehors du cadre médical ou légal français (y.c. recours à une AMP à l'étranger ou réalisés en France de façon « artisanale »)

ChiPRe - Les Chinois et les Chinoises de Paris et d'Île-de-France
→1ère enquête quantitative spécifiquement consacrée à la population née en Chine et vivant en Île-de-France
→Comprendre leurs conditions de vie et les difficultés qu'ils rencontrent dans leur vie quotidienne (obtention d'un titre de séjour, conditions de logement, recherche d'un emploi, victimes de discriminations)
→Décrire les réseaux de relations des personnes enquêtées selon leur région d'origine en Chine

SENOVIE – Mobilités thérapeutiques et cancers du sein
Etudier l'impact du cancer du sein et de sa prise en charge dans les parcours de femmes originaires d'Afrique Subsaharienne (parcours résidentiels, professionnels, sociaux, familiaux, administratifs, migratoires)

Ces points d'attention et les réponses apportées s'inscrivent à différents stades de la conception de l'opération. Une attention particulière doit être portée sur des phases en amont telles que les études exploratoires (le défrichage) et la conception du questionnaire. Prendre du temps pour ces phases va permettre de placer le terrain réel dans un contexte qui facilite l'échange et « désensibilise » le sujet.

L'étude exploratoire

Dans une étude exploratoire on peut s'interroger sur comment questionner les futurs enquêtés sur un sujet délicat – l'intime et la construction de la vie affective. L'enquête ENVIE distingue quatre types de relations : les couples, les histoires d'un soir, les relations suivies et les relations envisagée. Le terrain de l'enquête été précédé d'enquêtes qualitatives préparatoire pour saisir les termes utilisés par les jeunes pour nommer leurs relations, et ainsi utiliser les bons mots et le bon niveau de langage dans le questionnaire. Cette phase a identifié des questions jugées indiscrètes de leur point de vue et de proposer une bonne formulation pour qu'ils se sentent à l'aise pour y répondre.
Dans un autre registre, les phases exploratoires permettent de définir la méthodologie la plus appropriée pour capter des enquêtés difficiles à joindre telles que des personnes ayant recours à l'aide à la procréation médicalement assistée. La réflexion a conduit à s'appuyer sur une méthodologie de recherche participative en proposant un questionnaire en ligne diffusé via des réseaux identifiés comme étant largement utilisés par les personnes qui recourent à une AMP à l'étranger ou en dehors du cadre légal et médical français.

Le questionnaire

La structuration du questionnaire joue un rôle essentiel dans la confiance des répondants et la sincérité des réponses. L'ordre des questions a toute son importance, en plaçant au début les généralités, éléments biographiques, thématiques de l'enquête. Il s'agit de replacer les questions les plus sensibles dans un contexte plus large, sans pour autant les édulcorer, ni prendre les répondants par surprise. La collecte doit rester loyale.
Une fois la structure du questionnaire décidée, et en s'appuyant sur les travaux exploratoires, s'inscrit alors un travail sur la formulation des questions pour recueillir des informations précises sans brusquer, heurter, en posant des questions les moins invasives possibles. En d'autres termes, garantir une collecte loyale mais pas brutale.
Enfin, les questionnaires de certaines enquêtes font l'objet de traductions en plusieurs langues. Au-delà de lever la barrière de la langue, cela permet d'établir un climat de confiance avec les enquêtés, mais aussi d'assurer la perception des nuances, donc la qualité des réponses.
Quand on construit une enquête sur une thématique sensible, une attention particulière doit également être portée sur le protocole de contact et l'échantillonnage. En effet, **comment capter les populations difficiles à atteindre, en l'absence de base de sondage adaptée ?**
L'enquête CHIPRE s'appuie sur une méthode innovante, dite Network Sampling with Memory, s'appuyant sur les réseaux de contacts des enquêtés. Cela consiste dans un premier temps à sélectionner des « graines » aux profils démographiques, caractéristiques socio-économiques, parcours migratoires et aux origines géographiques en Chine aussi différents que possible, pour accéder à des réseaux distincts, tout en intégrant des personnes en situation régulière ou irrégulière. Puis, dans un second temps, on mobilise le réseau des enquêtés en leur demandant des contacts (plus d'information sur cette méthode est disponible dans un article présenté aux JMS 2022 - JMS 2022 – ChIPRe.

L'enquête SENOVIE, quant à elle, se place dans un cadre de recherches participatives, elle s'est appuyée sur un partenariat avec des structures de soins pour identifier les personnes à interroger. Le recrutement des répondantes s'est fait parmi les personnes ayant eu des consultations dans l'une des 4 structures partenaires de l'étude. Une telle intermédiation aide à établir un climat de confiance et à assurer la motivation des répondantes, en expliquant l'intérêt scientifique de l'enquête, donc de leur participation, même si le sujet est difficile ; elle permet aussi de sélectionner des femmes en capacité émotionnelle de répondre.

Dans les deux cas, il s'agit d'échantillons non probabilistes, ce qui pose des défis spécifiques dans la phase de pondération de l'enquête.

Le rôle crucial des enquêteurs

Leurs retours d'expérience, ou l'analyse de leurs pratiques, sont très précieux lors de la phase de test, pour rendre compte de la perception de la thématique par les enquêtés, de la compréhension des questions, de leur acceptabilité. Ils peuvent mettre en évidence le besoin de faire évoluer l'argumentaire, le lexique des termes utilisés, ou la formulation des questions.

Par ailleurs, la phase de formation est encore plus importante que pour une enquête « classique » : il s'agit pour l'enquêteur de recueillir des faits ou des opinions sur des situations souvent très interpellantes » sans porter de jugement, de prêter une attention encore plus marquée à l'assertivité, de faire preuve d'empathie sans laisser les émotions prendre le pas sur l'objectivité.

Face à des situations ou des événements difficiles, les enquêteurs doivent être particulièrement soutenus. Par exemple, il leur est possible de recourir à la psychologue de l'INED et, pour certaines enquêtes, un débriefing collectif est organisé.

En conclusion

Ces exemples, pris parmi d'autres, montrent toute l'importance des précautions à prendre quand on mène des enquêtes sur des terrains sensibles, mais aussi la capacité à construire des protocoles adaptés à ces contextes, tout en assurant à la fois le respect du droit et la protection de la vie privée, les principes éthiques, la prise en compte de situations difficiles, et la qualité des résultats.
C'est un terrain d'innovation permanente.

5. Faire parler des données sensibles: l'exemple de données sur la délinquance

Les données sur la délinquance et la victimation sont sensibles à plusieurs titres : elles relèvent de l'article 10 du RGPD (*supra*), relatif aux données pénales et infractions ; elles peuvent dans certains cas relever également de l'article 9, notamment parce que certaines infractions sont directement liées à des caractéristiques faisant l'objet de cet article, par exemple l'orientation sexuelle, l'origine réelle ou supposée… ; leur recueil auprès des personnes concernées peut être délicat, notamment en raison de la douleur associée ou de la difficulté à parler des faits correspondants. La diffusion de ces données peut alors nécessiter une vigilance accrue pour protéger la confidentialité de ces victimes, tout comme celle des personnes mises en cause qui sont également susceptibles d'être stigmatisées. Plus globalement les résultats statistiques associés peuvent prêter à polémique.

Données administratives et données d'enquête

Les données sur la délinquance s'appuient sur deux grands types de recueil : les données administratives et les enquêtes. Les données administratives proviennent des logiciels de rédaction de procédures, des dispositifs de déclaration en ligne et des systèmes de traitement des amendes. Elles ne correspondent donc qu'à ce qui est connu des forces de sécurité (police et gendarmerie nationales) et dûment enregistré par ces dernières. Leur contenu dépend de la propension des personnes à porter plainte pour les faits dont elles ont été victimes (Encadré, exemple 1), aux outils et pratiques de recueil de l'information, mais peut aussi dépendre du taux d'élucidation des procédures (différent selon le type de délinquance et le territoire), des orientations politiques qui induisent une plus ou moins grande attention au recueil de tel ou tel type d'infractions (Encadré, exemple 2). Les données administratives sont néanmoins incontournables puisqu'une infraction doit être qualifiée par un officier de police judiciaire pour être dénommée comme telle. Sur le champ des infractions enregistrées, elles sont exhaustives, ce qui permet des déclinaisons fines, au niveau géographique comme par catégorie d'infractions, mais nécessite l'hypothèse d'un taux de dépôt de plainte homogène pour toutes les populations pour permettre les analyses.
Les données d'enquêtes - Cadre de Vie et Sécurité jusqu'en 2021, puis Vécu et Ressenti en matière de Sécurité - portent sur le vécu ou le ressenti des personnes, que les faits relevant potentiellement d'une infraction aient ou non été portés à la connaissance des forces de l'ordre. Toutefois, étant réalisées sur échantillons, les déclinaisons sont plus limitées. Elles ne portent que sur les faits dont les personnes sont victimes puisque les individus interrogés ne sont pas questionnés en tant qu'auteur éventuel. Les protocoles de recueil de ces enquêtes sont adaptés à la sensibilité des données. En particulier, les questionnaires sont testés, avec des questions formulées de manière neutre et factuelle. Le protocole de collecte (une partie des questions sous casque pour l'enquête CVS, collecte par internet pour VRS) permet la distanciation nécessaire, les regards extérieurs sont apportés par le Comité du Label et le comité scientifique de l'enquête.

S'agissant des données administratives, l'enjeu est de pouvoir les traiter pour en tirer des statistiques interprétables et permettant d'alimenter le débat public. Outre le fait qu'elles ne portent que sur les infractions enregistrées, elles posent des défis liés à leur niveau de détail et à leur diversité (107 index statistiques suivis historiquement par le ministère de l'intérieur pour les seuls crimes et délits dit de « l'État 4001 », et près de 20 000 natures d'infractions différentes, incluant les contraventions, dans les nomenclatures du ministère de la justice[1]). Il s'agit entre autres d'extraire les bonnes informations, y compris dans des champs de texte libre, en s'appuyant pour cela sur les progrès récents en analyse

[1] Ce chiffre peut paraître très élevé, mais il s'explique par la très grande diversité des types d'infractions, et au nombre de règles qui encadrent la société et nos activités (une infraction pouvant être considérée comme un manquement à une règle)

textuelle[2], afin de trouver les niveaux de classification et d'agrégation pertinents pour constituer des groupes cohérents d'infractions enregistrées de même nature, à l'exemple des escroqueries et fraudes aux moyens de paiement.

En matière de diffusion, il est essentiel de toujours rappeler très précisément ce qui est effectivement mesuré et comment on le mesure, d'autant que la perception d'un même vocable peut être très différente d'une personne à l'autre (Encadré, exemple 3). Une attention particulière doit être portée aux unités de comptes retenues (un nombre de victimes, de véhicules volés, ...etc), pour ne pas agréger ce qui est sans rapport. Par ailleurs, les données diffusées doivent être présentées de manière à prévenir autant que possible les risques de mauvaise interprétation : replacer la donnée dans son contexte (taille des mailles de diffusion, saisonnalités éventuelles, événements particuliers…), attirer l'attention sur la significativité (ou non) des écarts ou des évolutions… Cela passe par exemple par des techniques de désaisonnalisation, des processus de Markov visant à distinguer les évolutions non prévisibles, ou plus simplement par la présentation de valeurs relatives plutôt qu'absolues.

Ces deux dispositifs de recueil sont donc comme souvent incomplets, mais complémentaires. Ils apportent une information sur des faits précis, là où le débat public porte le plus souvent sur des notions vagues ou très agrégées, comme la violence, la délinquance ou les incivilités.

Exemples de difficultés.

Violences sexuelles ou conjugales

La libération de la parole impulsé par le mouvement *#Me too* à compter de 2017-2018, et plus globalement la prise de conscience sociale récente vis-à-vis des violences sexuelles ou conjugales peuvent jouer sur la perception du phénomène, la propension à le signaler ou à porter plainte, et la manière dont on en rend compte dans une enquête. Le taux de dépôt de plainte est très faible (<5 % pour les violences sexuelles subies), mais serait à la hausse ces dernières années, étant donné la grande part de faits datant de plus d'une année qui sont déclarés aux forces de sécurité. Depuis quelques années, le dépôt de plainte est facilité avec des plaintiers mieux formés. En outre, le recueil d'informations est difficile, qu'elles soient d'enquête ou administrative en raison de l'intimité et la douleur des victimes, le rappel des situations de danger vécues.

Usage ou trafic de stupéfiants

Il s'agit d'infractions sans victime directe. Leur enregistrement est à l'initiative des forces de sécurité et sensible aux évolutions technologiques et législatives. Par exemple, le nombre de mis en cause pour usage de stupéfiants reflète l'activité des services et le temps accordé par les forces de sécurité à la révélation ce type d'infraction. L'introduction progressive à partir de 2020 des amendes forfaitaires délictuelles sous forme de procès verbal électronique a facilité et a largement contribué à la hausse du nombre de mis en cause pour ce type d'infraction, avec des écarts marqués d'une zone géographique à l'autre. Enfin, mener une enquête est naturellement difficile sur le sujet, en particulier sur le trafic de stupéfiants.

Lien entre « délinquance » et nationalité

Ventiler un nombre de victimes ou de mis en cause dans des procédures judiciaires nécessite de questionner la qualité des données. En effet, la saisie de la nationalité dans les logiciels de rédaction des procédures n'est pas certifiée, ce n'est que lors de la phase de conclusion judiciaire que l'identité est

[2] Voir par exemple la présentation aux JMS 2025 https://journees-methodologie-statistique.insee.net/extraction-automatique-des-quantites-de-stupefiants-dans-les-resumes-de-procedure-enregistres-par-les-forces-de-securite-interieure-a-laide-de-modeles-legers/

dûment enregistrée notamment des mis en cause. Étant donné qu'une infraction implique dans de nombreux cas victimes et auteurs, il est utile de s'intéresser à la nationalité des deux pour une analyse complète.

Les liens entre les différents indicateurs de délinquance et les nationalités (victimes et/ou mis en cause) ne sont pas homogènes : on ne relève pas les mêmes corrélations selon l'indicateur de « délinquance enregistrée » considéré. La présence de corrélations ne signifie pas causalité : d'autres variables exogènes peuvent indirectement expliquer la corrélation (*infra*).

Les bases de données locales (commune, département, région) de la délinquance enregistrée.

Les données de l'État 4001, l'outil statistique historique de suivi de la délinquance par le Ministère de l'Intérieur, étaient diffusées mensuellement à l'échelle départementale, et annuellement à l'échelle communale. La première diffusion en open data est ancienne et antérieure à la création du SSMSI (2014). Cependant, les données diffusées étaient brutes et non fiabilisées. En particulier :

- Elles étaient exprimées en lieu d'enregistrement. Par exemple, le vol d'un smartphone d'une personne résidant à Nantes lors d'un déplacement à Paris qui porte plainte près de son domicile, aurait été associé et enregistré en Loire-Atlantique alors que l'infraction aurait été commise à Paris.
- Elles détaillent l'ensemble des 107 index des crimes et délits. Hors certains indicateurs nécessitent des agrégations plus fines. Par exemple les destructions et dégradations volontaires résultent du cumul de délits recensés dans l'État 4001 et de contravention hors 4001, la frontière entre les différents types d'infraction étant très mince.

Cette base n'a pas toujours donné lieu à des réutilisations pertinentes par les utilisateurs. Sa diffusion a pu être arrêtée en 2022, notamment grâce au travail du SSMSI et les nouveaux produits proposés par le service. Au-delà de la fiabilisation d'un plus petit nombre d'indicateurs (niveau d'agrégation et unité de compte bien choisie, dédoublonnage et exploitation des descriptions textuelles des procédures, …), une base communale et départementale de la délinquance enregistrée a été diffusée pour la première fois en 2022.

Ces bases issues des enregistrements de la police et gendarmerie nationales, en particulier la base territoriale, sont documentées et mises à disposition en open data, dans une démarche de transparence et de démultiplication des usages. La sensibilité d'un sujet n'empêche pas l'open data, bien au contraire, c'est un moyen d'alimenter le débat sur des bases partagées. Mais cela implique un effort particulier de documentation et d'accompagnement des utilisateurs.

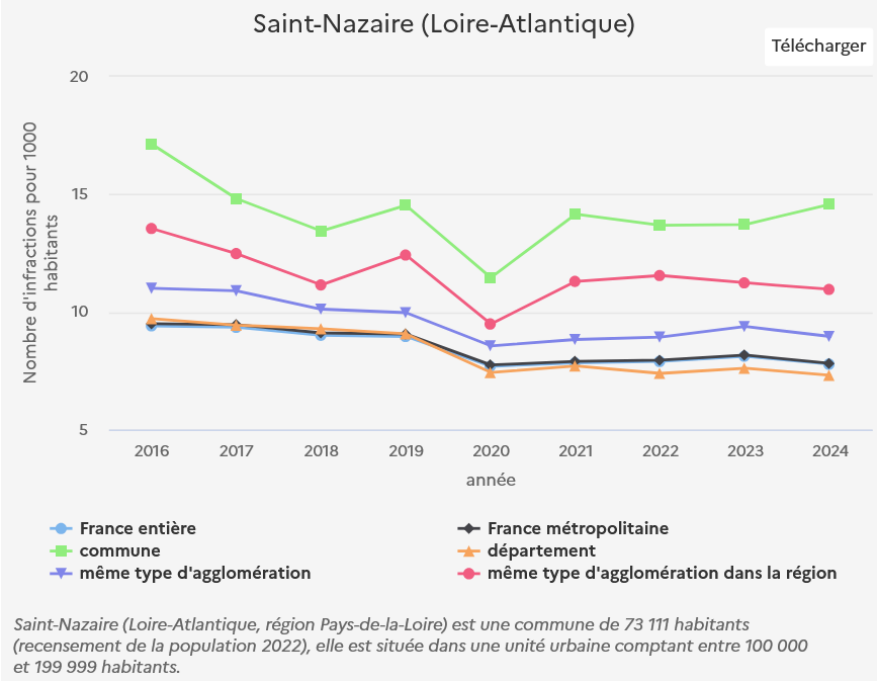
En matière de contenu, la base de données communales est limitée aux communes qui n'enregistrent aucune infraction, ou plus de 5 infractions, trois années de suite. Au-delà du respect du secret statistique, cela permet de se situer au-dessus d'un seuil d'interprétabilité. Il peut en effet exister une imprécision de la donnée, par exemple si la détermination de la commune de commission de l'infraction n'est pas évidente, ou exposée à des comportements déclaratifs variables, en particulier, lorsque le lieu de l'infraction est déterminé à partir des déclarations de la victime. Par exemple, il n'est pas rare qu'une zone d'activité commerciale soit à cheval sur plusieurs communes limitrophes. Un vol sur le parking de cette zone sera alors attribué avec imprécision à l'une des communes. La déclaration du crime ou du délit auprès des services de police ou de gendarmerie n'est pas non plus systématique.

Ainsi, une variation de quelques unités des faits enregistrés dans une commune peut être liée à ces comportements déclaratifs, plus qu'à l'évolution réelle de la délinquance. Les contraintes de diffusion permettent de limiter la valorisation d'évolutions non signifiantes (une évolution de + 100 % liée au passage de deux infractions à quatre infractions par exemple).

Des produits de diffusion adaptés et accompagnés

Le recours à la datavisualisation et à la cartographie permet de guider la lecture de ces données, en les replaçant dans leur contexte, ou en les mettant en perspective. L'effet taille des territoires est quasi systématiquement neutralisé en rapportant un nombre d'infractions, victimes ou mis en cause à une population (habitants ou logements). Mobiliser des taux permet ainsi les comparaisons territoriales : comparer les enregistrements d'une commune à celles de communes de taille comparable (cf. *figure*), ou à ceux du département ou de la région, le tout avec pour objectif d'illustrer l'inscription spatiale d'un phénomène. En outre, les cartes en taux font l'objet d'une attention et méthodologie particulières ce qui permet d'éviter les comparaisons territoriales peu pertinentes. La robustesse de chaque carte est en effet systématiquement testée : la significativité d'appartenance des territoires aux classes qui leur sont associées est vérifiée[3].

Figure – Extraction de l'application de datavisualisation des données territoriales du SSMSI pour la commune de Saint-Nazaire



Lecture : données de la commune de Saint-Nazaire, en taux pour neutraliser l'effet taille, avec différents points de comparaison standardisés, du niveau national ou local.

Faire parler les données communales : contextualiser et accompagner

En matière d'interprétation des données, une analyse des corrélations avec des données exogènes, permettant de caractériser les communes sur des critères socio-économiques (niveau de vie, structure du parc de logements, … etc), démographiques ou géographiques, a été menée par le SSMSI, pour (*i*) contextualiser ses données et (*ii*) prévenir des risques de rapprochements trop rapides. La démarche est notamment décrite par Milin (2025).

Au-delà de s'adapter aux distributions très spécifiques des données de délinquance (modèle à inflation de zéros), la démarche mise en place permet la sélection des variables les plus pertinentes parmi un très grand jeu de données issues de plusieurs sources. Elle permet ainsi de mettre en évidence des indicateurs de contexte qui mettent en perspective les données, en ne retenant que les corrélations les plus significatives « toutes choses égales par ailleurs ». Cela permet d'éviter que certaines variables soient omises, sciemment ou non, lors de la contextualisation.

[3] Voir partie « Sources, méthodes et aides à la lecture » de « Atlas départemental de la délinquance enregistrée en 2024 » (SSMSI, Interstats Références, 2025)

Mais la démarche ne s'arrête pas là : un effort de pédagogie très poussé est réalisé pour indiquer que corrélation n'est pas causalité. Certaines variables peuvent être corrélées car causées chacune par une ou plusieurs variables non observées. Et s'il y a effectivement causalité, la corrélation n'en indique pas le sens.

La démarche a été initialisée sur les données relatives aux cambriolages de logement, et a donné lieu à une opération coordonnée de diffusion (de treize études) entre l'Insee (Direction générale et Directions régionales) et le SSMSI, qui y a consacré une énergie importante : animation du groupe de pilotage, accompagnement des rédacteurs de l'Insee et de leur hiérarchie, relecture des publications. Au-delà des considérations évoquées plus haut en matière de contextualisation, une attention particulière a été apportée à la précision des définitions, à la manière de présenter les concepts retenus, au choix des termes employés, tout en garantissant la lisibilité des analyses publiées.

6. Pour conclure ...

Le fait que des données ou des sujets soient sensibles n'interdit pas de les collecter, mais impose des conditions de recueil et de traitements spécifiques, fixées par la loi ou résultant de considérations éthiques.
La décision de collecter ces données, et éventuellement de les combiner à d'autres, doit relever d'une réflexion approfondie où sont pris en compte plusieurs facteurs : le bénéfice pour la collectivité de l'information produite, (via l'éclairage du débat public, l'aide à la décision ou la recherche publique), leur sensibilité et l'acceptabilité sociale d'un tel traitement, la capacité à protéger les données collectées tout au long de leur cycle de vie.
Parfois, une donnée est sensible parce qu'elle porte sur des faits ou situations douloureuses, ou fait référence à des pratiques illicites. Les protocoles de collecte doivent être adaptés à des situations, qu'il s'agisse de l'échantillonnage, du questionnaire en lui-même ou du mode d'administration. Les phases de conception et de test sont cruciales, les partenaires de l'opération et les enquêteurs, notamment lors des tests y jouent un rôle déterminant. Les enquêteurs doivent être formés à la collecte et accompagnés pour cela.
Une fois les données collectées, leur protection doit être à la mesure des enjeux, compte tenu de leur sensibilité et des conséquences possibles d'une rupture de confidentialité.
Lors de la diffusion et de l'analyse, une attention toute particulière doit être portée à la documentation et à la précision des termes, à la significativité des données et de leurs évolutions, mais aussi à leur contextualisation. L'enjeu est que les usages soient fondés sur des bases claires, solides et partagées, chacun étant, bien entendu, libre des conclusions qu'il en tire.
Sur tout le cycle de vie de la donnée, la transparence doit être une exigence forte et partagée, c'est une condition essentielle à la fois pour des motifs d'acceptabilité (acceptabilité sociale, mais aussi acceptabilité par les personnes directement concernées), de qualité du recueil et des traitements, et d'utilisation.

Bibliographie

- Réglement Général pour la Protection des Données (2016)

- Marcovitch, I. et Rancourt, E. (2022). A Data Ethics Framework for Responsible Responsive Organizations in the Digital World. Statistical Journal of the IAOS 38, 1161-1172. DOI 10.3233/SJI-220067. IOS Press
- National Academies (2022). Panel on Transparency and Reproducibility of Federal Statistics for the National Center for Science and Engineering Statistics. Transparency in Statistical Information, Consensus Study Panel, The National Academies of Science, Engineering, Medicine; The National Academies Press, Washington D.C.

- Rancourt, E. (2019). The scientific approach as a transparency enabler throughout the data life-cycle. Statistical journal of the IAOS 35, 549-558. DOI 10.3233/SJI-190581. IOS Press.

- Milin K. (2025). Modélisation de la délinquance enregistrée à l'échelle communale, *Journées de méthodologie statistique de l'Insee*