5 Analyse du risque de divulgation

Conformément à sa mission, l'ISQ doit veiller à ce que l'information statistique de ses fichiers de données soit exploitée à son plein potentiel. Le CADRISQ joue un rôle primordial dans l'atteinte de cet objectif. Cependant, en vertu de la Loi sur l'Institut de la statistique du Québec, quiconque dont les services sont retenus par le statisticien en chef ne peut révéler ni faire révéler, par quelque moyen que ce soit, des renseignements recueillis en vertu de cette loi si ces révélations permettent de rattacher un renseignement à une personne, à une entreprise, à un organisme ou à une association en particulier. Pour ce faire, l'ISQ s'est doté de règles de confidentialité qui permettent de diminuer le risque de divulgation de renseignements confidentiels lors de la diffusion de résultats.

Tant pour les enquêtes auprès des individus et des ménages que pour les données administratives, tant en accès CADRISQ qu'en accès à distance, les membres autorisés d'une équipe de recherche doivent suivre les procédures en vigueur afin de faire approuver la diffusion de leurs résultats de recherche par un ou une analyste de l'ISQ.

Il est nécessaire d'analyser le risque de divulgation de chaque figure ou tableau destiné à être diffusé, mais également des fichiers de syntaxe, des documents texte et de tout autre document afin d'éviter toute fuite de données confidentielles.

Cette section présente les types de risque de divulgation et les différents concepts associés aux règles de contrôle. Cette vérification se fait par le biais de neuf thèmes[1].

1. Ces thèmes sont tirés de la *Politique relative à la confidentialité des tableaux de résultats pour diffusion – Procédure pour les tableaux de résultats produits aux CADRISQ à partir d'un fichier de microdonnées non masquées d'une enquête menée auprès des individus et des ménages*, diffusée en septembre 2021 par l'ISQ.

Définitions et exemples de divulgation

La divulgation de renseignements confidentiels signifie qu'un lien peut être établi entre des données diffusées et unités répondantes (personnes, ménages, organisations, entreprises, etc.).

Les résultats de recherche peuvent présenter un risque de divulgation. Vous trouverez dans les tableaux suivants quatre types de divulgation, d'autres concepts et quelques exemples.

Types de divulgation[1]
1. Divulgation par identification
La divulgation de l'identité se produit lorsqu'un individu peut être identifié à partir de données diffusées. L'identification ou l'auto-identification peut conduire à la découverte d'une rareté ou d'une unicité dans la population. L'élément ainsi découvert révélerait une information jusque-là inconnue sur un individu.
2. Divulgation par attributs
La divulgation par attributs survient quand de l'information confidentielle est révélée et peut être attribuée à un individu. Il n'est pas nécessaire qu'un individu précis soit identifié ou qu'une valeur précise soit révélée pour que cela se produise. Par exemple, le fait de diffuser une fourchette de salaires étroite concernant une profession particulière dans une région donnée peut constituer un cas de divulgation. **Attribut individuel** : Déduire de nouvelles informations sur une personne grâce à l'utilisation de données diffusées. On observe ce type de divulgation particulièrement lorsque de nombreux tableaux sont diffusés à partir d'un seul ensemble de données. **Attribut pour un groupe** : Déduire de nouvelles informations sur un groupe identifiable d'individus ou savoir que ce groupe n'a pas un attribut particulier.
3. Divulgation par recoupement
La divulgation par recoupement survient lorsque de l'information diffusée peut être combinée pour obtenir des données confidentielles. Il y a lieu de s'assurer d'examiner toutes les données qui doivent être diffusées. Alors qu'un tableau en soi peut ne pas révéler d'information confidentielle, la divulgation peut survenir si l'on combine des informations provenant de plusieurs sources, dont des sources externes (les données supprimées dans un tableau peuvent être dérivées à partir d'autres tableaux, par exemple). Déduire de nouvelles informations en extrayant les informations résiduelles entre plusieurs tableaux qui se chevauchent. Ce type de divulgation s'observe particulièrement lorsque les comparaisons entre deux tableaux ou plus permettent d'obtenir des petites cellules résiduelles (0,1 ou 2). Par exemple : • Recoupement de régions (p. ex. : tableau région A - définition 2000 et tableau région A – définition 1991). • Recoupement de lien entre deux tableaux ou plus (p. ex. : tableau du statut d'emploi et tableau des résidents dans la région A). • Recoupement de tableaux de sous-population (p. ex. : catégorie d'âges « < 25 ans » et deux catégories « < 20 ans » et « 20-24 ans »).
4. Perception du risque de divulgation
Le public peut avoir une compréhension différente des risques liés au contrôle des risques et sa perception peut être influencée par ce qu'il voit dans les tableaux. Gérer les perceptions est important pour maintenir la crédibilité de l'ISQ et des ministères et organismes. Même l'**apparence de divulgation** peut ternir la réputation d'une institution relativement à la confidentialité.

1. Ces concepts sont inspirés des documents de Statistique Canada : CENTRES DE DONNÉES DE RECHERCHE DE STATISTIQUE CANADA (2005), *Guide à l'intention des chercheurs ayant conclu une entente avec Statistique Canada*, p. 23-24.

Autres concepts

Variable de croisement : Un tableau est, en général, formé par le croisement d'une variable d'analyse (variable dépendante ou d'intérêt) et d'une variable de croisement (variable indépendante ou explicative). La variable de croisement est celle considérée comme exerçant une influence sur la variable d'analyse.

Par exemple, lors de l'analyse de la détresse psychologique selon l'âge, on peut comparer la proportion estimée d'individus se situant à un niveau élevé de l'échelle de détresse psychologique (variable d'analyse) entre les différents groupes d'âge (variable de croisement).

Sous-population identifiable : Une sous-population est considérée comme identifiable lorsque l'appartenance d'un individu ou d'un ménage à celle-ci est soit observable (p. ex. : handicap physique), soit caractérisée par un environnement particulier (p. ex. : classe de première année du secondaire, type d'emploi occupé, ménage dont la taille est supérieure à sept individus). La sous-population identifiable peut être de taille variable.

Notons que les notions de variable de croisement et de sous-population identifiable sont étroitement liées. En effet, les modalités d'une variable de croisement peuvent, dans certains cas, constituer des sous-populations identifiables.

Exemples de divulgation

Quelques exemples fictifs de divulgation

- Un joueur de hockey professionnel est sélectionné dans une enquête et l'information diffusée à propos de sa localité a presque assurément divulgué des renseignements confidentiels sur ce joueur, par exemple il est facile de savoir que le revenu le plus élevé dans cette localité est le sien (**divulgation de l'identité**).
- Les résultats d'une enquête longitudinale mettent en évidence un ménage qui a un profil migratoire inusité, ce qui mène à son identification (**divulgation de l'identité**).
- Les parents d'un jeune de 16 ans sélectionné dans une enquête voient un tableau montrant que toutes les personnes de 16 ans de l'échantillon dans leur région ont consommé des drogues (**divulgation d'attributs**).
- Un article de journal fait état d'une plainte déposée par un veuf de 37 ans à propos d'une enquête à laquelle il participait, alors qu'un tableau croisé montre qu'il y a seulement deux veufs dans la trentaine qui font partie de l'échantillon
- (cela mène éventuellement à la **divulgation de l'identité ou d'attributs**).
- En combinant plusieurs résultats, il est possible d'obtenir une information volontairement exclue d'un fichier de microdonnées à grande diffusion parce qu'elle présentait un risque de divulgation trop élevé (le pays de naissance d'un immigrant récent, par exemple).

Règles générales

Qu'il s'agisse de produits statistiques créés à partir de données administratives ou de données d'enquêtes, tant en **accès CADRISQ** qu'en **accès à distance**, vous devez en tout temps respecter les règles énoncées au tableau suivant. S'il existe un risque de divulgation, il faut alors appliquer une technique de masquage pour le réduire. Pour chacun des neuf thèmes, une technique de masquage est suggérée. Les techniques de masquage préconisées sont soit le regroupement de modalités, soit un recalcul selon un domaine contenant plus d'unités[2]. Une vigilance accrue est de mise lorsque des résultats sont produits à partir de données provenant d'une enquête longitudinale ou de cohortes extraites des données administratives, ou que l'échelle géographique est fine et les effectifs faibles, car le risque de divulgation est plus élevé.

Définitions et règles selon le type de statistique
A. Cellule à faible fréquence
Définition : Une cellule de tableau doit être basée sur un certain nombre d'observations pour éviter le risque de divulgation (5, 10, 15)[1].
Règle : Toutes les statistiques des cellules d'un tableau doivent être basées sur le minimum d'observations requis par l'ensemble de données (effectifs non pondérés). Dans le cas d'une estimation de proportion, il doit y avoir au moins le minimum d'observations requis au numérateur. C'est dire que toutes les cellules du tableau de fréquences non pondéré correspondant doivent contenir au moins le minimum d'observations requis. Si ce n'est pas le cas, un masquage doit être appliqué (p. ex. : regroupement des modalités problématiques). Notons que cette règle ne s'applique pas dans le cas des cellules vides. Lorsque ces dernières sont non structurelles, il faut appliquer la règle du thème C.
B. Cellule complète
Définition : Une cellule est complète si elle contient la totalité des observations, c'est-à-dire 100 % de la somme d'une colonne ou d'une rangée d'un tableau. Elle peut être structurelle ou non.
Règle : La diffusion d'un résultat issu d'une cellule complète **ne peut se faire**, sauf s'il s'agit d'une cellule complète structurelle. Il faut penser à un regroupement des modalités de façon à ce qu'il n'y ait plus de cellule complète non structurelle dans le tableau.
C. Cellule vide
Définition : Une cellule vide est dite non structurelle si elle peut techniquement comporter des individus, mais qu'elle n'en comporte pas. Il ne faut pas confondre ce type de cellule avec la cellule vide structurelle, qui représente une combinaison impossible. Par exemple, « avoir neuf ans » et « avoir trois enfants » sont deux caractéristiques qui ne pourraient être combinées. La cellule vide structurelle ne pose pas de problème de confidentialité.
Règle : La diffusion d'un résultat issu d'une cellule vide **ne peut se faire**, sauf s'il s'agit d'une cellule vide structurelle. Il faut penser à un regroupement des modalités de façon à ce qu'il n'y ait plus de cellule vide non structurelle dans le tableau
D. L'étendue et la valeur minimale ou maximale
Définition : Certaines statistiques peuvent représenter une valeur associée à des observations en particulier. C'est parfois le cas de l'étendue de la distribution, notamment par la valeur minimale et maximale.
Règle : L'étendue et la valeur minimale ou maximale de certaines variables comme l'âge, le poids, le revenu ou la taille du ménage ne doivent pas être diffusées. Afin d'illustrer la dispersion de ces valeurs, il faut plutôt utiliser une statistique comme écart-type.

Suite à la page 25

2. Une unité répondante peut représenter un individu ou un regroupement d'individus.

E. Statistique individuelle

Définition : Moyenne ou total d'une variable continue.

Règle : Toute statistique individuelle produite doit se fonder sur un nombre minimal d'observations[1] (effectifs non pondérés). Une statistique produite à partir d'effectifs trop faibles doit être recalculée selon un domaine contenant plus d'unités.

F. Statistique du ratio

Règle : Un ratio ne peut être diffusé si l'une de ses composantes (numérateur ou dénominateur) ne peut être diffusée. Le cas échéant, le ratio doit être recalculé selon un domaine contenant plus d'unités.

G. Statistique d'ordre

Définition : Médiane, centiles, etc.

Règle : On doit trouver un nombre minimal d'observations[1] au-dessus et au-dessous de ces statistiques d'ordre. Si ce n'est pas le cas, d'autres statistiques d'ordre doivent être calculées.

H. Modèle saturé ou presque saturé

Définition : Un modèle est saturé ou presque saturé s'il comporte de nombreux coefficients, c'est-à-dire presque autant qu'il y a de combinaisons possibles de valeurs de covariables. Ces modèles peuvent être obtenus lors d'une analyse de variance ou d'une régression.

Règle : Les résultats provenant d'un modèle saturé ou presque saturé ne doivent pas être diffusés.

I. Nuage de points, courbe de survie, graphique de résidus ou graphique en boîte

Règle : La diffusion de tels graphiques doit être faite avec circonspection, puisque ceux-ci affichent des valeurs qui s'appliquent à des répondants particuliers. Les graphiques comportant des valeurs aberrantes extrêmes ne devraient pas être diffusés.

1. Ce nombre peut varier selon l'ensemble de données utilisées. L'analyste informera les personnes concernées du seuil à respecter.

Règles spécifiques aux données administratives

Pour les données administratives des ministères et organismes, en plus de respecter les règles générales présentées précédemment, vous devez appliquer certaines procédures spécifiques afin de protéger la confidentialité des individus

La présente section énonce la procédure à suivre afin de faire approuver, par le ou la responsable du CADRISQ, la sortie de résultats de recherche produits à l'aide d'un fichier[3] de données administratives au CADRISQ ou en accès à distance.

Lorsque des données administratives sont jumelées avec des données provenant d'une enquête, ce sont les règles de confidentialités de l'enquête qui doivent être appliquées.

Vous pourrez constater que les règles à appliquer présentées dans cette procédure peuvent varier en fonction de la banque de données administratives utilisée. Cela est lié notamment à la provenance de ces fichiers, ainsi qu'à la nature de l'information les composant. Le ou la responsable du CADRISQ s'assurera dès le début du projet que vous avez en main l'ensemble des règles appropriées pour les banques de données utilisées.

Procédure d'analyse du risque de divulgation

La procédure à suivre pour analyser la confidentialité des résultats est la suivante :

1. Sélectionner les résultats à diffuser[4].
2. Vérifier que la taille de la population[5] associée à la zone géographique de diffusion respecte la valeur du critère géographique associé à la banque de données administratives utilisée. Ce critère varie selon la nature de l'information contenue dans la banque de données.
3. Vérifier le risque de divulgation à l'aide des règles générales de confidentialité présentées précédemment, lesquelles dépendent de la banque de données administratives utilisée.
4. S'il existe un risque de divulgation, appliquer une technique de masquage pour diminuer ce risque. Le regroupement des modalités problématiques est la technique de masquage recommandée. Au besoin, consulter le ou la responsable du CADRISQ pour d'autres suggestions de masquage[6].
5. Avant la diffusion des résultats, appliquer l'arrondissement aléatoire aux résultats[7] en utilisant la base d'arrondissement spécifique à votre projet (5, 10,15). Des outils sont disponibles pour vous aider à appliquer un arrondissement aléatoire.

3. Ce fichier peut être constitué des données d'une partie seulement ou de l'entièreté de la population visée. Si le fichier de données administratives ne contient qu'un échantillon aléatoire de la population visée, alors on considère ce fichier comme l'équivalent d'une enquête. Dans ce cas, il faut utiliser les Règles spécifiques aux fichiers de recherche comportant des données d'enquêtes.
4. Notez qu'il est possible d'accéder à certains résultats non masqués en accès à distance pour les opérations de validation. Pour plus de détails, voir la section du guide sur les Résultats intermédiaires.
5. La population associée à une zone géographique de diffusion donnée comprend toutes les personnes vivant dans cette zone, et non pas uniquement celles comprises dans le sous-ensemble visé par le projet de recherche. Cette population peut être estimée à partir des estimations de population basées sur le recensement. Seule exception, les analyses portant sur des travailleurs et des travailleuses ; dans ce cas, la taille de la population doit seulement tenir compte des personnes travaillant dans la zone géographique de diffusion.
6. Notez qu'il arrive parfois que la diffusion ne soit pas possible.
7. Seuls les tableaux de fréquences doivent être arrondis. Les proportions devront être calculées sur les données arrondies (numérateur et dénominateur). Certains graphiques, soit ceux qui sont construits à partir d'un tableau de fréquences (p. ex. un histogramme), devront également être construits à partir des données arrondies. Les résultats de modèles ou de tests n'auront pas à être arrondis. Les autres types de statistiques non plus.

Remarques importantes

- Il faut limiter autant que possible les sorties de résultats n'étant pas destinés à être diffusés. Pour toute analyse exploratoire, il convient plutôt d'utiliser l'outil d'**accès aux résultats intermédiaires**, qui permet aux membres autorisés d'une équipe de consulter à distance des résultats temporaires n'étant pas destinés à la diffusion.
- Il est possible que le ou la responsable du CADRISQ doive refuser la diffusion de certains tableaux lorsque ceux-ci sont produits à partir d'une population ou d'une sous-population identifiable qui serait jugée très petite ou très visible ou que le nombre de variables de croisement est trop élevé, ce qui pourrait contrevenir aux obligations de protection des renseignements de l'ISQ.
- Une vigilance accrue est de rigueur lorsqu'une étude ou une analyse est faite sur plusieurs années, car le risque de divulgation est plus élevé lorsque l'on croise des variables provenant de périodes différentes, par exemple, lorsqu'on effectue le suivi géographique d'un individu (déménagement d'une région à une autre).
- L'ISQ a à cœur de suivre l'évolution des préoccupations de protection des renseignements personnels, des méthodologies de recherche et des technologies. Par conséquent, les règles de confidentialité sont sujettes à changement. Lors d'une révision de la politique ou des règles, le ou la responsable du CADRISQ vous fera savoir si des règles relatives à l'un des ensembles de données viennent à changer.

Critère géographique

Vérifiez que la taille de la population de la zone géographique de diffusion pour laquelle des résultats sont produits est d'au moins **1 000** individus[8]. Si ce n'est pas le cas, il faut produire des résultats à un niveau géographique qui respecte ce critère.

La population associée à une zone géographique de diffusion donnée comprend toutes les personnes vivant dans cette zone, et non pas uniquement celles comprises dans le sous-ensemble visé par le projet de recherche. La taille de cette population peut être estimée à partir des estimations de population basées sur le recensement[9].

Seule exception, les analyses portant sur des travailleurs et des travailleuses ; dans ce cas, la taille de la population doit tenir compte uniquement des personnes travaillant dans la zone géographique de diffusion. Cette distinction est faite car la distribution géographique des travailleurs et des travailleuses n'est pas la même que celle de la population en général. Par ailleurs, s'il s'avère possible de localiser les travailleurs et travailleuses de manière précise à l'intérieur de la zone géographique de diffusion, alors l'estimation de la taille de la population devrait être basée sur les travailleurs et travailleuses visés par le projet de recherche. Le ou la responsable du CADRISQ peut, au besoin, vous accompagner dans la détermination de la population estimée.

Arrondissement

Comme les données administratives ne bénéficient pas de l'incertitude créée par l'échantillonnage que l'on retrouve dans les enquêtes et de la pondération en découlant, on doit arrondir les valeurs. En effet, l'arrondissement permet de réduire les risques de divulgation en créant un léger bruit. De nombreuses institutions appliquent cette procédure, dont Statistique Canada[10] qui le fait depuis des décennies pour le Recensement.

L'arrondissement est une méthode efficace pour protéger les tableaux de fréquences, surtout quand plusieurs tableaux sont produits à partir du même ensemble de données, tout en conservant le plus possible l'utilité des résultats (c'est-à-dire en entraînant une perturbation minimale). L'arrondissement permet donc de sortir un éventail de données sans trop affecter la précision des résultats, surtout lorsque l'effectif sur lequel on travaille est de taille convenable, ce qui est généralement le cas lorsque l'on veut produire des résultats statistiquement fiables.

8. Dans certaines situations, il se peut que le seuil du critère géographique à appliquer soit plus sévère. Ce seuil dépend, entre autres, des informations contenues dans la banque de données administratives.
9. Consulter le site Web de Statistique Canada ou de l'ISQ afin d'accéder aux plus récentes données, selon la ventilation souhaitée (âge, sexe, découpage géographique, etc.).
10. Aussi US Census Bureau, Office for National Statistics (Royaume-Uni), Stats Nz (Nouvelle-Zélande).

La règle est la même pour toutes les bases de données provenant des différents ministères et organismes (RAMQ, MED-ECHO, etc.). De façon générale[11] :

- tout tableau d'effectifs doit être arrondi en base 5[12] ;
- les proportions et certains graphiques doivent être créés à partir d'effectifs arrondis.

Il existe différentes méthodes pour arrondir une valeur : l'arrondissement peut être normal, aléatoire ou contrôlé.

L'arrondissement contrôlé est la méthode recommandée par Eurostat[13] pour la diffusion des tableaux de fréquences. Il a l'avantage d'offrir des tableaux additifs ainsi qu'une protection accrue.

L'ISQ offre des outils afin de faciliter la tâche aux chercheurs et aux chercheuses. Des syntaxes de programmation pour appliquer la procédure d'arrondissement automatiquement aux résultats sont fournies avec les fichiers de recherche.

Règles propres aux fichiers de recherche comportant des données d'enquêtes

Pour les données d'enquêtes, en plus de respecter les règles générales présentées précédemment, vous devez vous assurer que vos tableaux respectent les règles de confidentialité, dont certaines sont propres à chaque enquête. **Vous devez notamment** :

1. produire pour une même analyse, dans des fichiers séparés et bien identifiés, les résultats pondérés (en fonction des poids d'enquête) et non pondérés (produits issus des données brutes) pour chaque type de statistique produite ;
2. produire les tableaux comportant des variables liées à l'ethnie[14] séparément et les identifier clairement. Ceux-ci sont soumis à des vérifications supplémentaires en raison de leur nature sensible.

Il est possible que le ou la responsable du CADRISQ doive refuser la diffusion de certains tableaux si :

- les tableaux sont produits à partir d'une enquête portant sur l'ethnie ;
- les tableaux contiennent une variable de croisement liée à l'ethnie ;
- les tableaux sont produits à partir d'une sous-population identifiable jugée très petite ou très visible ;
- les tableaux sont produits à partir de plusieurs variables de croisement.

Utilisation des poids

Quand le fichier de recherche disponible comprend des données d'enquêtes, il faut recourir à la pondération de l'enquête pour éviter que les estimations ne comportent des biais. Même dans le cas où les estimations de population (inférences des résultats à la population) ne sont pas le premier intérêt de la recherche et même si l'équipe de recherche travaille sur une partie du fichier qui ne comporterait que des données administratives, la pondération permet de corriger le biais découlant du plan d'échantillonnage. Ce biais peut se produire à la suite d'un suréchantillonnage ou d'un sous-échantillonnage de sous-groupes de la population, de la non-réponse, de la collecte des données et des opérations de traitement. Ainsi, seuls les résultats pondérés sont autorisés à sortir du CADRISQ.

Toutefois, la sortie peut être autorisée dans certaines circonstances. Par exemple, si vous souhaitez décrire l'échantillon pour un article scientifique ou si vous menez des analyses longitudinales ou multiniveaux pour lesquelles certains logiciels ne permettent pas l'utilisation des pondérations. Vous devrez alors justifier par écrit la nécessité de sortir des résultats non pondérés. Des règles supplémentaires pourraient cependant s'appliquer.

11. Ces règles sont sujettes à changement. Lors de tout changement, les guides d'utilisation sont mis à jour, et les utilisateurs et utilisatrices sont avisés dès que les nouveaux guides sont disponibles.
13. Pour certaines banques de données administratives, il se peut que la base d'arrondissement à appliquer soit plus élevée. Le cas échéant, l'analyste du CADRISQ vous indiquera la base spécifique à appliquer.
13. CENEX-projects Handbook (2009).
14. L'ethnie comprend la langue, la culture, le statut de minorité visible et l'identité autochtone.

Institutions scolaires

Afin de protéger la confidentialité des renseignements des enfants, du personnel et des institutions, certaines vérifications supplémentaires doivent être faites pour les enquêtes dans lesquelles les participants et participantes ont été sélectionnés à partir d'une institution, telles que l'*Enquête québécoise sur la santé des jeunes au secondaire* (EQSJS) et l'*Enquête québécoise sur le développement des enfants à la maternelle* (EQDEM). Une vérification de la confidentialité par rapport à l'institution doit être effectuée. Veuillez consulter le *Guide de l'utilisateur – Procédures pour la confidentialité des tableaux d'enquêtes touchant les écoles* se trouvant dans la documentation de l'enquête.