

L'approche de l'ISQ pour assurer la confidentialité des fichiers de microdonnées et des tableaux de résultats

Jimmy Baulne^{*}, Éric Gagnon^{*} et Lyne Des Groseilliers^{*}

^{*} Institut de la statistique du Québec, Direction de la méthodologie, de la démographie et des enquêtes spéciales, 200, chemin Sainte-Foy, 3^e étage, Québec (Québec), Canada G1R 5T4

Résumé : En vertu de sa loi constitutive, l'Institut de la statistique du Québec (ISQ) doit s'assurer de préserver la confidentialité des renseignements qu'il recueille. À cette fin, l'Institut s'est doté d'énoncés de politique qui lui permettent de respecter ses engagements en matière de confidentialité. Deux de ces énoncés concernent la confidentialité des produits statistiques diffusés par l'Institut : les fichiers de microdonnées et les tableaux de résultats. Dans la première partie de cet article, il est question plus particulièrement des fichiers de microdonnées ainsi que du contrôle statistique de la divulgation (CSD) qu'on leur applique. On verra ainsi que les mesures de CSD peuvent être plus ou moins sévères selon les conditions d'utilisation des fichiers de microdonnées. La deuxième partie de l'article portera sur les tableaux de résultats. On procédera au survol des différentes composantes en matière de diffusion de tableaux à l'Institut. Par la suite, l'approche intégrée de l'Institut sera discutée afin de mettre en évidence les divers volets étudiés : les mesures de CSD en ce qui concerne les statistiques démographiques, les enquêtes auprès des individus et les enquêtes auprès des entreprises.

1 Introduction

L'Institut de la statistique du Québec (ISQ) est l'organisme statistique officiel du gouvernement du Québec. Sa mission est de fournir des informations statistiques fiables et objectives sur tous les aspects de la société québécoise. À cette fin, l'Institut mène chaque année plusieurs enquêtes auprès d'individus et d'entreprises. De plus, il établit et tient à jour le bilan démographique du Québec, et assume la responsabilité du Registre des événements démographiques (RÉD). En vertu de sa mission, l'Institut doit exploiter tout le potentiel d'information statistique que recèlent les renseignements recueillis dans ses différents champs d'activité. Cependant, l'Institut ne dispose pas de ressources internes suffisantes pour y parvenir. En conséquence, l'Institut a choisi, comme orientation stratégique, de maximiser l'exploitation de ses produits statistiques par des tiers. Il doit toutefois s'assurer que cet usage respecte sa loi constitutive en vertu de laquelle l'Institut doit préserver la confidentialité des renseignements qu'il recueille. Il a donc adopté une approche pour la diffusion de ses fichiers de microdonnées et de ses tableaux, qui vise à offrir une souplesse quant à l'accessibilité aux produits tout en conservant la rigueur essentielle en matière de respect de la confidentialité.

Dans un premier temps, la section 2 de cet article traite de l'approche adoptée pour la diffusion des fichiers de microdonnées produits à partir d'enquêtes auprès des individus. Dans un deuxième temps, la section 3 présente l'approche retenue pour la

diffusion des tableaux produits à la suite d'enquêtes auprès d'individus et auprès d'entreprises, ainsi que pour la diffusion de tableaux produits à partir du RED.

2 Fichier de microdonnées

2.1 Diffusion de fichiers de microdonnées

Afin de favoriser la production maximale d'information à partir de ses enquêtes, l'Institut met à la disposition de chercheurs externes différents types de fichiers de microdonnées qui comportent un potentiel analytique variable. L'accessibilité à ces fichiers est permise tout en assurant la confidentialité des données des répondants. Dans cette section, il sera question des différents types de fichiers de microdonnées produits à partir d'enquêtes auprès d'individus, ainsi que des mesures de contrôle de la divulgation appliquées à ceux-ci. L'accès aux fichiers produits à partir d'une enquête auprès des entreprises ne sera donc pas abordé dans cet article. La nature de l'information disponible sur de tels fichiers requiert des mesures de contrôle de la divulgation plus sévères que celles que nous présentons ici.

Par ailleurs, il existe deux méthodes pour rendre disponibles des fichiers de microdonnées à des chercheurs. La première consiste à demander le consentement préalable des répondants pour transmettre les données qui les concernent à des chercheurs. La seconde méthode accorde plutôt aux chercheurs un accès aux fichiers de microdonnées lorsqu'il n'y a pas de consentement préalable. Dans cet article, il sera seulement question des fichiers diffusés grâce à cette deuxième méthode.

2.2 Accès aux fichiers de microdonnées sans consentement préalable

L'Institut peut donner accès, pour différentes raisons, à des fichiers de microdonnées sans consentement préalable. Par exemple, le consentement peut ne pas avoir été demandé parce qu'un chercheur désire avoir accès à l'ensemble des répondants du fichier, de manière à ce que ses résultats soient cohérents avec ceux de l'Institut. Il se peut également que le consentement préalable ait été demandé pour les chercheurs d'un organisme et que des chercheurs d'un autre organisme manifestent, en cours de projet, le désir d'avoir eux aussi accès au fichier. Évidemment, le consentement préalable ne s'applique pas dans ce cas et un autre type d'accès doit être proposé.

Différentes approches sont proposées par l'Institut en l'absence de consentement préalable. Ces approches permettent de rendre accessibles les microdonnées tout en préservant la confidentialité des répondants. Pour ce faire, différentes mesures de contrôle du risque de divulgation sont mises de l'avant :

- les mesures statistiques : contrôle statistique de la divulgation (CSD);
- les exigences légales et administratives;
- les mesures de sécurité physique et informatique.

L'application combinée de ces mesures permet de contrôler adéquatement le risque. Il est possible de faire varier la sévérité de chacune de ces mesures tout en assurant un contrôle du risque adéquat. Par exemple, si l'on décide d'appliquer des mesures statistiques moins sévères, alors les exigences légales et administratives ainsi que les mesures de sécurité physique et informatique devront être plus strictes.

2.2.1 Fichier de microdonnées à grande diffusion

Un premier type d'accès proposé, lorsqu'il n'y a pas de consentement préalable, consiste à rendre disponible à un chercheur un fichier de microdonnées à grande diffusion (FMGD) sur son lieu de travail. Avant de produire un tel fichier, il faut classer les variables du fichier en trois catégories : les identifiants directs, les identifiants indirects et les variables non identificatrices. Les identifiants directs sont des variables grâce auxquelles on peut identifier directement une personne; par exemple, le nom, l'adresse et le numéro de téléphone sont des identifiants directs. Quant aux identifiants indirects, ils permettent d'identifier une personne lorsqu'ils sont croisés entre eux, notamment le sexe, l'âge et la profession. Enfin, toutes les autres variables du fichier sont non identificatrices et ne font donc pas partie de l'analyse du CSD. Pour créer le FMGD, les mesures de CSD sont appliquées aux identifiants directs et indirects du fichier. Tout d'abord, les identifiants directs sont supprimés, puis des mesures de CSD très sévères sont appliquées aux identifiants indirects. Les mesures de CSD sont appliquées en deux étapes. La première est l'identification du risque et la seconde est le masquage qui permet de minimiser le risque. Pour identifier le risque, les critères suivants sont utilisés:

- Une région distinguable dans le fichier doit abriter au moins 80 000 habitants.
- Chaque cellule provenant de la combinaison des modalités de trois identifiants indirects doit compter au moins 800 personnes dans la population.
- L'un des identifiants indirects de la combinaison doit être la région distinguable dont il a été question précédemment.

Le critère minimal de 80 000 habitants dans une région a déjà été utilisé par Statistique Canada pour la création de FMGD (Béland, 1999). Dans certaines circonstances, ce critère et celui du nombre minimal de personnes dans une cellule peuvent être assouplis ou raffermis. Pour utiliser des seuils différents, on doit tenir compte de la nature plus ou moins sensible de l'information disponible dans le fichier et du type de population visée par l'enquête. De plus, il faut s'assurer que le recours à des seuils inférieurs n'entraîne pas une importante augmentation du risque.

Pour la seconde étape des mesures de CSD, les techniques de masquage suivantes peuvent être appliquées :

- recatégorisation des variables régionales et des identifiants indirects à risque;
- suppression d'un identifiant indirect à risque du fichier;
- suppression de l'identifiant indirect à risque pour certains répondants;
- regroupement de valeurs extrêmes (*top-coding* et *bottom-coding*);

- arrondissement ou ajout d'un bruit aléatoire.

L'application de ces mesures de CSD permet de minimiser le risque de divulgation d'information et, ainsi, d'assouplir les autres mesures de contrôle du risque. Par exemple, il n'est pas nécessaire, pour le chercheur, d'appliquer des mesures de CSD aux tableaux de résultats produits à partir de ce type de fichier. Lorsqu'il utilise un tel fichier, le chercheur doit tout de même s'engager à :

- utiliser le fichier à des fins d'analyse et de recherche;
- ne pas coupler le fichier à un autre fichier, ni tenter de réidentification;
- ne pas faire de copie de sécurité du fichier.

S'il ne respecte pas ses obligations, le chercheur peut se voir retirer le FMGD.

2.2.2 Fichier aux fins d'analyse ou de recherche externe (FARE) disponible sur les lieux de travail d'un chercheur

Un deuxième type d'accès proposé consiste à rendre disponible à un chercheur un fichier FARE sur son lieu de travail. Cet accès permet au chercheur de travailler sur un fichier comportant un potentiel d'analyse supérieur à ce qu'offre le FMGD. Pour obtenir un tel accès, le chercheur doit cependant signer une entente avec l'Institut, en vertu de laquelle il s'engage à protéger la confidentialité des données qui lui sont transmises. Pour créer un fichier FARE, il faut classer les variables du fichier selon les mêmes catégories que pour le FMGD; les identifiants directs sont supprimés et les mesures de CSD appliquées aux identifiants indirects sont moins sévères que pour la création d'un FMGD. Comme dans le cas des FMGD, les mesures de CSD s'appliquent en deux étapes : l'identification du risque et le masquage. Quant aux fichiers FARE, l'identification du risque est faite à l'aide des critères suivants :

- Une région distinguable dans le fichier doit abriter au moins 10 000 habitants.
- Chaque cellule provenant de la combinaison des modalités de trois identifiants indirects doit compter au moins 100 personnes dans la population.
- L'un des identifiants indirects intervenant dans la combinaison doit être la région distinguable dont il a été question précédemment.

Ces critères d'identification du risque, inspirés des méthodes élaborées par Statistics Netherlands (Schulte Nordholt, 2001), sont moins sévères que ceux qu'on présente pour les FMGD. De plus, ces critères peuvent être assouplis ou renforcés dans certaines circonstances. Les conditions à respecter pour leur modification sont les mêmes que celles qu'on a indiquées pour les FMGD.

En ce qui concerne la seconde étape des mesures de CSD, le masquage appliqué au fichier FARE utilise les mêmes techniques que celles qu'on emploie pour les FMGD. Cependant, l'application de ces techniques résulte en un masquage moins sévère que celui du FMGD, puisqu'un risque inférieur a été identifié à la première étape.

L'application des mesures de CSD pour créer les fichiers FARE permet de réduire le risque de divulgation, mais ne l'élimine pas. Afin de contrôler adéquatement le risque, il faut donc appliquer des mesures d'ordre légal, physique et informatique plus sévères que celles qu'on emploie pour les FMGD :

- Le transport par l'utilisateur des microdonnées est interdit.
- Les copies papier doivent être conservées dans un endroit sécurisé.
- L'accès à la copie du fichier original de microdonnées ou à ses sous-produits doit être contrôlé et limité aux seules personnes autorisées.
- Le fichier doit être conservé dans un lieu sécuritaire sous forme encryptée.
- À la fin du projet, la copie du fichier original de microdonnées doit être détruite et une note confirmant cette action doit être envoyée à l'Institut.
- Le chercheur doit appliquer des mesures de CSD aux tableaux produits à partir de ce fichier. Des détails sur ces mesures sont donnés à la section 3.3.1.
- Etc.

En cas de non-respect de ses obligations, le chercheur peut se voir retirer l'accès au FARE. L'Institut peut même le poursuivre en justice.

2.2.3 Fichier de microdonnées dénominalisé mais non masqué disponible au CADRISQ

Un troisième type d'accès proposé consiste à rendre disponible à un chercheur un fichier de microdonnées, dénominalisé mais non masqué, dans les locaux du Centre d'accès aux données de recherche de l'Institut de la statistique du Québec (CADRISQ). Cette approche peut être privilégiée par un chercheur qui ne serait pas satisfait du potentiel analytique du fichier FARE.

Dans ces fichiers, seuls les identifiants directs sont supprimés. Aucune mesure de CSD n'est donc appliquée aux identifiants indirects. En conséquence, les risques de divulgation à partir de ce fichier sont grands. Cependant, pour compenser l'absence de CSD, les mesures d'ordre légal, physique et informatique relatives à l'utilisation de ce fichier sont plus sévères que celles qu'on utilise pour les fichiers FARE :

- Le fichier de microdonnées demeure dans les locaux du CADRISQ.
- Les analyses sont menées sous supervision du responsable du CADRISQ.
- Le chercheur est assermenté et soumis aux mêmes obligations de confidentialité qu'un employé de l'Institut.
- Des mesures de CSD doivent être appliquées par le chercheur aux tableaux de résultats qu'il désire sortir du CADRISQ. La section 3.3.1 donne plus de détails au sujet de ces mesures de CSD. Le responsable du CADRISQ s'assure que les mesures de CSD ont été appliquées adéquatement par le chercheur, ce qui permet de rendre les tableaux sécuritaires.

2.2.4 Fichier FARE consultable par accès à distance

Un quatrième type d'accès proposé consiste à rendre disponible à un chercheur un fichier FARE par accès à distance. Il permet au chercheur de travailler chez lui en mode terminal sur le fichier FARE. Il est important de mentionner que, même si le chercheur effectue ses analyses sur son lieu de travail, le fichier FARE demeure physiquement dans les locaux de l'Institut. Les mesures de CSD appliquées pour créer ce fichier sont les mêmes que pour le fichier FARE remis chez le chercheur. Cependant, les autres mesures relatives à l'utilisation du fichier sont plus strictes. En effet, toutes les opérations effectuées par le chercheur sont visualisées à distance par un employé de l'Institut. Cette supervision s'apparente à celle en vigueur au CADRISQ. De plus, le chercheur ne peut télécharger des parties du fichier sur son ordinateur, ni en imprimer des extraits sur papier. Notons que tous les tableaux de résultats que le chercheur désire extraire de cet environnement sécuritaire doivent être vérifiés par un employé de l'Institut qui s'assure que les tableaux ne comportent pas de risque de divulgation.

Par ailleurs, l'Institut envisage d'assouplir les mesures de CSD appliquées aux fichiers de microdonnées accessibles à distance, puisque les mesures de sécurité informatique utilisées sont plus sévères que lorsqu'on rend disponible un fichier FARE chez le chercheur. Dans le futur, l'Institut aimerait donc rendre disponible par accès à distance un fichier qui serait moins masqué qu'un fichier FARE.

3 Tableaux de résultats

3.1 Diffusion de tableaux de résultats

L'approche présentée dans les prochains paragraphes porte sur la diffusion de tableaux de résultats produits à partir d'un fichier de microdonnées appartenant à l'Institut. Cette diffusion de tableaux peut être faite par un chercheur externe qui exploite un fichier de l'Institut, mais également par un employé de l'Institut au moment de la publication des résultats de son enquête.

Contrairement à l'approche ayant trait à la diffusion des microdonnées qui porte seulement sur les données individus, l'approche qui se préoccupe de la diffusion des tableaux porte autant sur les données individus que sur les données entreprises.

Par ailleurs, dans la partie qui présente l'approche relative à la diffusion des fichiers de microdonnées, une distinction a été effectuée au sujet des différents types de fichiers qui peuvent être mis à la disposition des utilisateurs. Or, cette distinction a une incidence directe sur les mesures de CSD appliquées aux tableaux de résultats. C'est pourquoi l'approche relative à la diffusion des tableaux prend en considération le type de fichier à partir duquel le tableau est produit, non masqué ou FARE. Mais encore, les mesures de CSD appliquées aux tableaux sont également dépendantes du

type d'utilisateur qui diffuse le tableau, soit un employé de l'Institut ou un utilisateur externe.

En tenant compte de toutes ces distinctions, voyons plus en détail l'approche de l'Institut concernant les mesures de CSD applicables lors de la diffusion de tableaux.

3.2 Élaboration d'une politique

Que ce soit pour ses propres publications ou pour une publication faite par un chercheur qui exploite l'un de ses fichiers de microdonnées, l'Institut doit fournir à l'exploitant de ses fichiers une procédure dans laquelle sont énoncées des règles auxquelles il doit se conformer. Ces règles ont pour objectif d'assurer la confidentialité des renseignements diffusés. En outre, dans le cas d'un chercheur qui exploite un fichier de l'Institut, le non-respect de cette procédure pourrait entraîner des poursuites judiciaires contre celui-ci et l'organisme qui l'emploie.

Étant donné que l'Institut a l'obligation d'assurer la confidentialité des renseignements publiés, celui-ci s'est doté d'une politique qui énonce des lignes directrices en matière de confidentialité des tableaux de résultats pour diffusion.

Cette politique couvre différents types de tableaux : de fréquence, de quantité – moyenne, total ou ratio –, percentiles et résultats d'analyse par modèle (régression). De plus, les tableaux peuvent être produits à partir d'un fichier d'enquête portant sur des individus, sur des entreprises, ou du RED, dont l'Institut est copropriétaire avec le ministère de la Santé et des Services sociaux du Québec (MSSS).

Il existe donc différentes circonstances dont l'Institut doit tenir compte dans l'élaboration de sa politique relative à la diffusion de tableaux. Par exemple, qui souhaite diffuser le tableau, un employé de l'Institut ou un chercheur externe ? Le tableau a été produit à partir de quel fichier (non masqué ou FARE) ? Et le tableau porte sur quel type de données (individus, entreprises ou d'ordre démographique) ? Tous ces éléments ont une incidence sur le choix des mesures de CSD à appliquer aux tableaux. L'Institut a donc dû se doter d'énoncés de politique, assortis chacun d'une procédure distincte, qui lui permettent de respecter ses engagements en matière de confidentialité dans toutes les situations menant à la diffusion de tableaux.

3.3 Organisation des lignes directrices

Les lignes directrices sur la confidentialité des tableaux pour diffusion ont été scindées en trois volets : le volet des enquêtes individus, celui des enquêtes auprès des entreprises et celui des statistiques démographiques. Chaque volet comporte une procédure pour chaque situation différente menant à la diffusion de tableaux.

3.3.1 Volet des enquêtes auprès des individus

Le volet des enquêtes individus comporte trois procédures. La première porte sur les tableaux produits par un utilisateur qui est externe à l'Institut à partir d'un fichier non

masqué. Un tel fichier peut être mis à la disposition d'un chercheur soit au CADRISQ, soit dans les locaux de son organisme public lorsqu'il y a eu consentement préalable des répondants. Étant donné qu'aucune mesure de CSD n'a été appliquée aux identifiants indirects des fichiers de microdonnées de ce type (voir section 2.2.3), le risque de divulgation est donc très grand et les mesures de CSD appliquées aux tableaux sont sévères.

Pour cette procédure, l'identification du risque de divulgation d'un tableau s'effectue en vérifiant notamment l'absence d'un nombre minimum de répondants dans chacune des cellules du tableau ou la présence de cellules vides ou complètes. Une cellule est complète lorsqu'elle contient toutes les unités répondantes. À l'opposé, une cellule vide ne renferme aucune unité répondante. Les techniques de masquage appliquées aux tableaux jugés à risque sont fonction des variables les composant. En effet, cette procédure utilise deux concepts importants, soit la présence d'une variable liée à l'ethnie et la taille de la classification géographique. La distinction des tableaux en fonction de l'ethnicité est justifiée par le fait que, au Québec, cette notion est très sensible, et que les sous-populations formées par les différentes communautés culturelles sont relativement petites, ce qui augmente le risque d'identification. Le constat est le même pour les sous-populations définies par certains territoires géographiques, d'où la spécificité de tels tableaux.

Les mesures de CSD sont donc plus sévères lorsqu'il y a présence d'une variable liée à l'ethnie et lorsque la taille de la classification géographique est petite. Parmi les techniques de masquage utilisées dans cette procédure, on trouve :

- le regroupement de modalités;
- la suppression locale de données (y compris la suppression secondaire);
- la limitation du nombre de variables de croisement utilisées dans un tableau;
- l'interdiction de diffuser des tableaux au niveau régional (dans certains cas).

La deuxième procédure porte sur les tableaux produits à partir d'un fichier FARE exploité par un utilisateur qui est externe à l'Institut. Le risque de divulgation associé à ces tableaux est moindre que celui des tableaux produits à partir de fichiers non masqués, car des mesures de CSD ont été appliquées aux microdonnées. Ainsi, des mesures de CSD moins sévères peuvent être appliquées aux tableaux. Cette procédure utilise les mêmes concepts que la première, à savoir la présence d'une variable liée à l'ethnie et la taille de la classification géographique. La méthode d'identification du risque de divulgation est également la même, et une partie seulement des techniques de masquage est utilisée pour diminuer ce risque. C'est, en quelque sorte, ce qui distingue les deux premières procédures, et cette distinction est motivée par le fait que des mesures de CSD sont appliquées aux microdonnées d'un fichier FARE, mais pas aux microdonnées d'un fichier non masqué. Par exemple, on permettra la diffusion d'un tableau au niveau régional s'il est produit à partir d'un fichier FARE et qu'il respecte les règles de la deuxième procédure, tandis que la

production d'un tel tableau à partir d'un fichier non masqué (première procédure) pourrait ne pas être permise.

La troisième procédure du volet des enquêtes auprès des individus porte sur les tableaux produits par un employé de l'Institut. Bien entendu, le type de fichier de microdonnées utilisé pour produire les tableaux est un fichier non masqué. La présence d'une variable liée à l'ethnie est encore une fois un concept important pour déterminer les mesures de CSD à appliquer aux tableaux. Par contre, le deuxième concept utilisé dans cette procédure est celui de la présence ou non d'une variable délicate dans le tableau. Une variable est délicate si elle contient une information qui se rapporte à la vie privée, que l'on ne connaît pas habituellement et qu'on ne souhaite pas divulguer, tel le comportement sexuel et la cause d'une incapacité. On doit donc classer les tableaux dans l'une des quatre catégories suivantes :

- tableau d'une variable délicate avec croisement d'une variable liée à l'ethnie;
- tableau d'une variable délicate sans croisement d'une variable liée à l'ethnie;
- tableau d'une variable non délicate avec croisement d'une variable liée à l'ethnie;
- tableau d'une variable non délicate sans croisement d'une variable liée à l'ethnie.

Le statut attribué aux variables, c'est-à-dire délicates ou non, relève du chargé de projet de l'enquête, sur approbation de son gestionnaire. Cette stratégie permet d'alléger les mesures de CSD appliquées aux tableaux dans certaines situations précises. Les tableaux de la quatrième catégorie en constituent un exemple. Les méthodes d'identification du risque de divulgation sont les mêmes que pour les deux autres procédures. Encore une fois, la sévérité des mesures varie en fonction du classement du tableau. Les tableaux qui comportent des variables délicates ou liées à l'ethnie se voient appliquer des mesures de CSD plus sévères, tandis que les autres tableaux sont soumis à des mesures moins strictes, notamment par la présence permise de cellules à faible fréquence dans les tableaux. Les techniques de masquage préconisées dans cette procédure sont la combinaison de catégories et la suppression locale des données jugées confidentielles, y compris la suppression secondaire.

3.3.2 Volet des enquêtes auprès des entreprises

Le volet des enquêtes auprès des entreprises ne comporte qu'une seule procédure, qui a trait aux tableaux produits par un employé de l'Institut. Tout comme son pendant du volet individus, le type de fichier utilisé pour produire les tableaux est un fichier non masqué, donc sans aucune mesure de CSD appliquée aux identifiants indirects.

La procédure de ce volet utilise un concept équivalent à la troisième procédure du volet des enquêtes individus, à savoir les variables délicates, mais adapté à la notion d'entreprise. Ainsi, les tableaux de ce volet sont catégorisés selon qu'ils sont constitués d'une variable stratégique ou d'une variable non stratégique. Toute information susceptible de donner à une entreprise un avantage sur ses concurrents, peut être considérée comme une variable stratégique.

Les mesures de CSD appliquées aux tableaux faisant intervenir une variable stratégique sont plus sévères que les mesures appliquées aux autres tableaux. Pour l'unique procédure de ce volet, l'identification du risque de divulgation dépend de l'absence d'un nombre minimum de répondants dans chacune des cellules ou de la présence de cellules vides ou complètes et, dans le cas des tableaux de quantité, d'une mesure de sensibilité, telles la règle de dominance (n,k) et la règle p-pourcent (Willenborg, 2001). Pour réduire ce risque, les techniques de masquage suivantes sont utilisées :

- la suppression locale de données (y compris la suppression secondaire);
- le regroupement de modalités;
- l'ajout d'un bruit aléatoire;
- l'arrondissement contrôlé ou aléatoire.

Tout comme pour le volet individus, le choix des variables stratégiques et non stratégiques relève du chargé de projet de l'enquête, sur approbation de son gestionnaire. Cependant, dans le cas du volet des enquêtes auprès des entreprises, un comité, constitué d'employés de l'Institut, a été formé spécialement pour établir une liste de variables, regroupées sous forme de thèmes, devant obligatoirement être considérées comme stratégiques. Ainsi, un employé qui désire diffuser des tableaux doit, conformément à l'application de la procédure, s'aider de cette liste pour déterminer le statut des variables.

3.3.3 Volet des statistiques démographiques

L'Institut diffuse des données à partir des fichiers des naissances, des mariages, des décès et des mortinaissances, en son nom et en tant que mandataire du MSSS.

Le volet des statistiques démographiques est particulier du fait que la publication qui émane du RED est constituée de tableaux statutaires auxquels s'ajoutent des demandes à la pièce en nombre restreint. Contrairement aux fichiers d'enquête qui peuvent porter sur un éventail de sujets, donc de variables, le RED se concentre sur un nombre fixe d'indicateurs qui servent à la production annuelle d'un ensemble récurrent de tableaux.

Contrairement aux procédures des autres volets, basées davantage sur l'obtention d'un nombre minimal d'unités dans chaque cellule du tableau, celle du volet des statistiques démographiques utilise plutôt les caractéristiques des variables qui constituent le tableau. En effet, étant donné que les variables sont les mêmes, année après année, un poids leur est attribué et l'identification du risque est fonction de ce poids. Bien entendu, le poids des variables, ainsi que la valeur des seuils utilisés pour la prise de décision, sont exposés dans des documents confidentiels.

4 Conclusion

En terminant, l'Institut, en tant qu'organisme statistique officiel du gouvernement du Québec, a l'obligation d'assurer la confidentialité des informations qu'il diffuse. Les approches présentées dans ce document lui permettent de répondre à ses obligations, tout en rendant accessibles des données avec un potentiel analytique satisfaisant.

Références

- Béland, Y. (1999). « Release of Public Use Microdata Files for NPHS? Mission... Partially Accomplished! », *Proceedings of the Survey Research Methods Section*, American Statistical Association, p. 404-409.
- Schulte Nordholt, E. (2001). « Statistical Disclosure Control (SDC) in Practice. Some Examples in Official Statistics of Statistics Netherlands », article présenté à la Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Skopje, The former Yugoslav Republic of Macedonia.
- Willenborg, L. et T. De Waal (2001). *Elements of Statistical Disclosure Control. Lecture Notes in Statistics 155*, New York, Springer-Verlag.