N° 11-522-XIF au catalogue

La série des symposiums internationaux de Statistique Canada - Recueil

Symposium 2003 : Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

2003

Statistique Canada Statistics Canada

Canada

Recueil du Symposium 2003 de Statistique Canada
Défis reliés à la réalisation d'enquêtes pour la prochaine décennie

APPLICATIONS DES MÉTHODES DE CONTRÔLE DE LA DIVULGATION DES DONNÉES STATISTIQUES

Eric Schulte Nordholt[1]

RÉSUMÉ

Les instituts nationaux de la statistique (INS) et les bureaux d'études de marché réalisent des enquêtes portant sur de nombreux sujets. À cette fin, ils ont mis au point un processus de production de statistiques entièrement équipé. Le parcours est long de la collecte des données brutes à la publication d'information à l'intention du public. Le présent article porte sur les aspects méthodologiques du processus de production de statistiques. Le but d'un institut national de statistique ou d'un bureau d'études de marché est toujours de publier autant de données que possible. Cependant, il doit respecter la confidentialité des renseignements fournis par les répondants individuels. Par conséquent, on a mis au point des techniques de contrôle de la divulgation des données statistiques pour protéger les renseignements sensibles qui peuvent être attribués à des répondants particuliers. Le contrôle de la divulgation des statistiques (CDS) est un sujet que l'on peut examiner sous divers angles. Le présent article explique comment des résultats de projets de recherche européens sur le contrôle de la divulgation des statistiques peuvent être appliqués à la production de statistiques officielles. Nous décrivons aussi deux progiciels statistiques qui permettent de produire des données ne posant pas de risque de divulgation. Le progiciel τ-ARGUS est utilisé pour traiter les données tabulaires et son jumeau, μ-ARGUS, pour traiter les microdonnées. Les principales techniques utilisées pour protéger l'information sensible sont le recodage global et la suppression locale. Le progiciel τ-ARGUS et son jumeau, μ-ARGUS sont des produits qui ont été développés dans le contexte du projet de CDS financé par le quatrième Programme-cadre de la Communauté européenne. De nouvelles versions des deux progiciels (auxquelles sont intégrés les résultats des travaux de recherche en cours) ont été diffusées dans le cadre du projet concernant les aspects informatiques de la confidentialité des statistiques, ou projet CASC, qui est financé par le cinquième Programme-cadre de la Communauté européenne. Nous discutons aussi d'autres méthodes permettant l'utilisation des données. La plus importante est celle des sites d'accès restreint. Les chercheurs de bonne foi qui ont besoin de plus d'information peuvent se rendre à Statistics Netherlands et travailler sur place dans une aire sécurisée. Le projet AMRADS (mesures complémentaires en recherche et développement en statistique) fait progresser le transfert des CDS dans de nombreux pays.

MOTS CLÉS : μ-ARGUS, τ-ARGUS, contrôle de la divulgation des statistiques, microdonnées, progiciel, tableaux.

1. INTRODUCTION

Les instituts nationaux de la statistique (INS) et les bureaux d'études de marché réalisent des enquêtes portant sur des sujets très variés. À cette fin, ils ont mis en place un processus entièrement équipé de production de données statistiques. Le parcours est long de la collecte des données brutes à la publication d'information à l'intention du public.

L'information tirée des données statistiques est communiquée au public sous forme de données tabulaires et de microdonnées. Dans le passé, seules les données tabulaires étaient diffusées et les INS avaient le monopole des microdonnées. Depuis les années 1980, la révolution causée par les ordinateurs personnels a mis fin à ce monopole. Aujourd'hui, d'autres utilisateurs de statistiques ont aussi la possibilité de se servir de microdonnées. Ces dernières peuvent être transmises sur disquette, disque compact ou d'autres supports. En outre, récemment, d'autres modes de transmission de l'information statistique, comme le téléaccès et la téléexécution, ont gagné en popularité. Grâce à ces techniques, les chercheurs peuvent avoir accès aux données qui demeurent au bureau de la statistique ou peuvent exécuter des formatages sans que les données soient chargées sur leur ordinateur personnel. Quand les données sont

1 Eric Schulte Nordholt, chercheur principal, Statistics Netherlands, Division Social and Spatial Statistics, Department Support and Development, Section Research and Development, P.O. Box 4000, 2270 JM, Voorburg, The Netherlands, ESLE@CBS.NL.

de nature très délicate, certains INS ont la possibilité d'autoriser les chercheurs de bonne foi à travailler sur place dans leurs locaux.

La tâche des bureaux de la statistique est de produire et de publier des renseignements statistiques sur la société. Les données recueillies sont, en dernière analyse, diffusées sous une forme adaptée à l'usage statistique qu'en font les décideurs, les chercheurs et le grand public. La diffusion de ce genre d'information peut avoir l'effet indésirable de divulguer des renseignements sur des entités individuelles au lieu d'un groupe suffisamment grand d'individus. La question qui se pose alors est de savoir comment modifier l'information disponible de façon à ce que les données diffusées soient utiles du point de vue statistique, sans toutefois compromettre la protection des renseignements confidentiels fournis par les entités concernées. On fait appel à la théorie du contrôle de la divulgation des données statistiques pour déterminer comment publier et diffuser des données aussi détaillées que possible sans divulguer de renseignements confidentiels (Willenborg et De Waal, 1996 et 2001).

Le présent article se fonde en partie sur des travaux antérieurs sur le contrôle de la divulgation des statistiques (p. ex. Schulte Nordholt, 2001). Y sont discutées les méthodes disponibles pour protéger les données sensibles. Les tableaux produits par les bureaux de la statistique d'après les microdonnées provenant des enquêtes doivent être protégés contre le risque de divulgation. Pour cela, on peut se servir du progiciel τ-ARGUS (Hundepool et coll., 2002a) pour traiter les tableaux produits. Au chapitre 2, nous décrivons plus en détail τ-ARGUS et la façon dont il peut être appliqué. Au chapitre 3, nous expliquons comment on peut produire des microdonnées pour la recherche et des fichiers de microdonnées à grande diffusion au moyen du progiciel μ-ARGUS (Hundepool et coll., 2002b). Au chapitre 4, nous discutons de l'option qu'ont les chercheurs de bonne foi de travailler sur place à Statistics Netherlands sur des fichiers de microdonnées plus riches, ainsi que d'autres méthodes permettant d'utiliser les données. Enfin, au chapitre 5, nous discutons de la situation courante et de certaines extensions possibles des progiciels ARGUS. Nous tirons également certaines conclusions. Nombre d'idées exposées dans le présent article sont extraites de Citteur et Willenborg (1993), Groot et Citteur (1997), Willenborg (1993) et Willenborg et De Waal (1996 et 2001).

Les progiciels τ-ARGUS et μ-ARGUS sont le fruit du projet relatif au contrôle de la divulgation des statistiques (CDS) réalisé aux termes du quatrième Programme-cadre de la Communauté européenne. Le projet sur les aspects informatiques de la confidentialité des statistiques, ou projet CASC, peut être considéré comme le prolongement du projet CDS. Le projet CASC est financé par le cinquième Programme-cadre de la Communauté européenne pour la recherche, le développement technologique et la démonstration (RTD). Ce projet s'appuie sur les réalisations du projet CDS et les étoffe, mais a également de nouveaux objectifs. Il met davantage l'accent sur les outils pratiques et la recherche nécessaire pour les développer. Le projet CASC regroupe 14 partenaires provenant de 5 pays européens (Allemagne, Italie, Pays-Bas, Espagne et Royaume-Uni) qui travaillent en étroite collaboration. L'une des tâches principales de ce groupe est de pousser plus loin le développement des logiciels ARGUS qui ont été mis dans le domaine public par le groupe du projet CDS. Le projet CASC comprend à la fois des travaux de recherche et de développement de logiciel. En ce qui concerne la recherche, le projet se concentre sur des domaines dans lesquels on peut s'attendre à obtenir des solutions pratiques qui seront alors intégrées dans les logiciels. Le projet CASC s'articule sur les logiciels jumeaux ARGUS. Il permettra d'appliquer les résultats de la recherche aux activités quotidiennes des instituts nationaux de la statistique et des bureaux d'études de marché. Des renseignements plus détaillés sur le projet CASC figurent dans Hundepool (2001).

2. LA DIFFUSION DE DONNÉES TABULAIRES

Nombre de tableaux sont produits d'après les données d'enquête. Le progiciel τ-ARGUS peut être utilisé pour protéger ces tableaux contre le risque de divulgation statistique (Hundepool et coll., 2002a). Deux stratégies courantes adoptées pour éliminer le risque de divulgation sont le remaniement des tableaux et la suppression de valeurs individuelles. La suppression des valeurs contenues dans certaines cellules des tableaux est nécessaire, car la publication de (bonnes approximations de) ces valeurs peut entraîner la divulgation de renseignements confidentiels. Ces suppressions sont appelées suppressions primaires.

On applique souvent une règle de dominance pour décider quelles cellules doivent être supprimées. Cette règle dit que la publication d'une cellule pose un risque de divulgation si les *n* principaux contributeurs à cette cellule sont responsables d'au moins *p* pour cent de la valeur totale de la cellule. L'idée sous-jacente est que, dans les cellules posant un risque de divulgation, les principaux contributeurs peuvent déterminer avec grande précision la contribution de leurs compétiteurs. Dans τ-ARGUS, la valeur par défaut de *n* est 3 et celle de *p* est 70 %, mais l'utilisateur peut les modifier facilement s'il le souhaite. En utilisant la règle de dominance choisie, τ-ARGUS indique à l'utilisateur quelles cellules présentent un risque. Dans les publications, des croix (×) remplacent normalement les valeurs des cellules dangereuses. La règle du pourcentage p et la règle pq peuvent aussi être utilisées pour décider quelles cellules il faut supprimer. La règle du pourcentage p dit que la divulgation approximative de données quantitatives (données fournies par des entreprises qui déclarent des quantités non négatives au sujet de certains établissements ou d'entités comparables) a lieu si l'utilisateur peut estimer de façon trop exacte la valeur déclarée par certains répondants. Ce genre de divulgation a lieu si les estimations supérieure et inférieure de la valeur fournie par le répondant sont plus proches de la valeur déclarée qu'un pourcentage spécifié a priori, p, et la cellule correspondante du tableau est donc déclarée sensible. Pour établir la règle du pourcentage p, on suppose qu'il existe une connaissance a priori limitée des valeurs propres à un répondant. Certaines personnes estiment que les bureaux de la statistique ne devraient pas émettre ce genre d'hypothèse. Dans le cas de la règle pq, les bureaux de la statistique peuvent préciser la quantité de connaissances a priori en attribuant une valeur q, qui représente le degré d'exactitude avec lequel les répondants peuvent estimer la valeur déclarée par un autre répondant avant que toute donnée soit publiée ($p<q<100$).

La technique la plus répandue pour repérer les cellules sensibles est l'application de la règle de dominance. On peut considérer la règle du pourcentage p comme un cas particulier de la règle pq. Cette dernière est intuitivement plus claire et plus facile à étendre à des situations particulières que la règle de dominance. La règle pq peut aussi être utilisée dans le cas de contributions ou de valeurs de cellule négatives dans le tableau. Si certains contributeurs savent approximativement quelle est la contribution de certains concurrents à la valeur d'une cellule, il est possible de tenir compte de cette information a priori en appliquant la règle pq. Par contre il n'en est pas ainsi de la règle de dominance. Un exemple de cette situation est celui où un répondant à l'intérieur d'une cellule jugée sensible permet de publier la cellule. Ce genre de renonciation à la confidentialité peut être utile aux fins de publication et ne pas être trop contraignante pour une grande entreprise publique pour laquelle des renseignements comparables font déjà partie du domaine public. La règle pq permet de traiter les cas de renonciation; par contre, si l'on utilise la règle de dominance, la façon de poursuivre n'est pas certaine, puisqu'il ne devrait pas être permis de divulguer approximativement la valeur d'un autre contributeur à la cellule. Enfin, la règle pq a l'avantage de tenir compte à la fois des limites inférieure et supérieure, alors que la règle de dominance ne permet d'utiliser qu'une limite supérieure. Ce dernier inconvénient de la règle de dominance s'applique aussi à la règle du pourcentage p. Malgré ces inconvénients, rares sont les pays qui ont acquis de l'expérience dans l'utilisation d'autres règles que celle de dominance pour l'identification des cellules sensibles dans les tableaux. Toutefois, quand la règle du pourcentage p et la règle pq seront offertes dans des progiciels standards de contrôle de la divulgation statistique, on peut s'attendre à ce que ces règles deviennent plus populaires.

Lorsqu'on donne des totaux de marge ainsi que des valeurs de cellule, il est nécessaire de supprimer des cellules supplémentaires pour s'assurer que les valeurs des cellules supprimées au départ ne puissent être recalculées d'après les totaux de marge. Même s'il n'est pas possible de recalculer exactement les valeurs des cellules supprimées, il est souvent possible de les reproduire dans un intervalle suffisamment petit. En pratique, la valeur de chaque cellule est souvent non négative et ne peut donc excéder les totaux de marge de la ligne ni de la colonne. Si la taille de l'intervalle est faible, il est possible d'estimer la valeur de la cellule supprimée avec grande précision, ce qui est naturellement indésirable. Par conséquent, il faut supprimer d'autres cellules pour s'assurer que les intervalles soient suffisamment grands. L'utilisateur doit indiquer la grandeur de l'intervalle qu'il considère comme suffisant. Cet intervalle porte le nom de fourchette de sécurité et cette dernière pourrait, par exemple, avoir une borne inférieure égale à 70 % et une borne supérieure égale à 130 % de la valeur de la cellule. L'utilisateur d'un tableau n'est pas capable de déterminer si une suppression est primaire ou secondaire : normalement, toutes les cellules supprimées sont marquées d'une croix (×). Ne pas révéler la raison pour laquelle une cellule a été supprimée permet d'empêcher la divulgation d'information.

De préférence, on exécute les suppressions secondaires de façon optimale, mais la définition de l'optimalité est un problème intéressant. On peut définir plusieurs mesures de la perte d'information, puis minimiser cette dernière en fonction de la mesure choisie. Quatre possibilités sont :

- la minimisation du nombre du suppressions secondaires;
- la minimisation du total des valeurs supprimées;
- la minimisation du nombre total de contributions individuelles aux cellules supprimées;
- la minimisation d'une fonction pondérée de scores attribués aux cellules qui symbolisent l'information, où les cellules vides obtiennent un poids 0 et les cellules voisines des cellules faisant l'objet d'une suppression primaire obtiennent un poids plus faible que celles qui en sont plus éloignées.

Souvent, la minimisation du nombre de suppressions secondaires est considérée comme étant optimale. De temps à autre on recourt aussi à la minimisation du total des valeurs supprimées ou du nombre total de contributions individuelles aux cellules supprimées. La minimisation du total des valeurs supprimées n'est naturellement pertinente que si toutes les valeurs des cellules sont non négatives. Dans le cas de la dernière minimisation mentionnée, on peut tenir compte de la hiérarchie du tableau, puis adapter le logiciel aux besoins particuliers, s'il y a lieu. Dans τ-ARGUS, l'option de minimisation du total des valeurs supprimées a été définie comme option par défaut. Dans la version 2.1 de τ-ARGUS, il est également possible de minimiser le nombre total de contributions individuelles aux cellules supprimées. Si l'on utilise ce critère, une variable dite de coût, dont la valeur est égale à 1 pour chaque enregistrement, est utilisée pour exécuter les suppressions secondaires. L'option de la minimisation du nombre de suppressions secondaires a également été implémentée dans la version 2.1 de τ-ARGUS. Par conséquent, les trois options qui peuvent donner lieu à des groupes différents de suppressions secondaires, peuvent être comparées avec la version 2.1 de τ-ARGUS.

Souvent, si l'on exécute les suppressions secondaires directement sur les tableaux les plus détaillés disponibles, on obtient un grand nombre de suppressions locales. Par conséquent, il est préférable d'essayer de combiner les catégories des variables explicatives. Le remaniement d'un tableau par regroupement de strates produira un plus petit nombre de lignes et de colonnes. La combinaison de deux cellules sûres produira une nouvelle cellule sûre. Par contre, si l'on combine deux cellules dont au moins une est dangereuse, il est impossible de savoir d'avance si la cellule résultante sera sûre ou dangereuse, mais il est facile de le vérifier par après à l'aide de τ-ARGUS. Cependant, les cellules retenues, qui contiennent un plus grand nombre d'entreprises, ont tendance à mieux protéger l'information individuelle, ce qui sous-entend que le pourcentage de cellules dangereuses a tendance à diminuer lorsque l'on regroupe les strates. Donc, une stratégie pratique pour protéger un tableau consiste à commencer par regrouper les lignes ou les colonnes, tâche qui peut être exécutée facilement dans τ-ARGUS. La façon la plus facile d'apporter de petits changements aux variables explicatives consiste à procéder à une révision manuelle dans la boîte de recodage de τ-ARGUS; par contre, il est plus efficace de traiter les changements importants dans un fichier de recodage produit extérieurement que l'on peut importer dans τ-ARGUS sans problème. Quand ce processus de remaniement est terminé, on peut exécuter les suppressions locales au moyen de τ-ARGUS sachant les valeurs des paramètres n, p et des bornes inférieure et supérieure de la fourchette de sécurité.

Comme on produit normalement un grand nombre de tableaux d'après les données d'une même enquête et que le progiciel utilisé pour protéger les données est fondé sur des tableaux individuels, il existe un risque, même si chaque tableau est sûr, que la combinaison des données qu'ils contiennent entraîne la divulgation d'information individuelle. Il pourrait en être ainsi quand les tableaux ont des variables explicatives et des variables de réponse en commun. La version 2.1 de τ-ARGUS peut traiter des tableaux couplés. Une version antérieure contenait une option pour protéger ce genre de tableau, mais le résultat n'était pas garanti. Aujourd'hui, nous avons atteint le but consistant à étendre τ-ARGUS de façon à ce qu'il puisse traiter une sous-classe importante de tableaux couplés, à savoir les tableaux hiérarchiques. Un tableau hiérarchique est un tableau ordinaire avec totaux de marge, mais aussi des totaux partiels supplémentaires. Les tableaux hiérarchiques demandent la résolution de problèmes d'optimisation nettement plus complexes que ceux posés par les tableaux simples. On dispose de certaines méthodes d'approximation pour trouver les solutions optimales de ces problèmes. La version étendue 2.1 de τ-ARGUS a été diffusée dans le cadre du projet sur les aspects informatiques de la confidentialité des statistiques (projet CASC).

3. LA DIFFUSION DE MICRODONNÉES POUR LES CHERCHEURS ET DE FICHIERS PUBLICS DE MICRODONNÉES

De nombreux utilisateurs des données d'enquête se satisfont des tableaux de données sûrs diffusés par les bureaux de la statistique. Cependant, certains d'entre eux ont besoin de plus d'information. Dans le cas de nombreuses enquêtes, les bureaux de la statistique produisent aussi des microdonnées à l'intention des chercheurs. Le progiciel µ-ARGUS (Hundepool et coll., 2002b) facilite la production de ces microdonnées. Statistics Netherlands utilise l'ensemble de règles suivant pour produire les microdonnées pour les chercheurs :

1. Les identificateurs directs ne doivent pas être diffusés.
2. Les identificateurs indirects sont subdivisés en variables extrêmement identifiantes, variables très identifiantes et variables identifiantes. Seules les variables régionales directes sont considérées comme extrêmement identifiantes. Chaque combinaison de valeurs d'une variable extrêmement identifiante, d'une variable très identifiante et d'une variable identifiante devrait survenir au moins 100 fois dans la population.
3. Le niveau maximal de détail pour la profession, l'entreprise et le niveau de scolarité est déterminé d'après la variable régionale directe la plus détaillée. Cette règle ne remplace pas la règle 2, mais en est plutôt une extension.
4. Une région que l'on peut discerner dans les microdonnées doit compter au moins 10 000 habitants.
5. Si les microdonnées sont tirées de données de panel, on ne doit pas diffuser de données régionales directes. Cette règle empêche la divulgation d'information individuelle en utilisant la longitudinalité des microdonnées.

Dans le cas de la plupart des statistiques sur les entreprises produites par Statistics Netherlands, les entreprises répondantes sont tenues, aux termes d'une loi sur la statistique officielle, de fournir leurs données au Bureau. Cette loi, dont l'adoption remonte à 1936, a été renouvelée en 1996 sans que l'obligation qu'ont les entreprises de répondre soit modifiée. Aucune information individuelle ne peut être divulguée lors de la publication des résultats de ces enquêtes auprès des entreprises. La loi stipule que l'on ne peut publier aucune microdonnée pour la recherche tirée de ces enquêtes. Par conséquent, Statistics Netherlands peut fournir deux types d'information provenant de ces enquêtes, à savoir des tableaux et des fichiers publics de microdonnées. Ces derniers contiennent des renseignements nettement moins détaillés que les microdonnées pour la recherche. Le progiciel µ-ARGUS (Hundepool et coll., 2002b) facilite aussi la production de fichiers publics de microdonnées. Pour la production de ces fichiers, Statistics Netherlands applique l'ensemble de règles suivant :

1. Les microdonnées doivent avoir au moins un an avant d'être publiées.
2. Les identificateurs directs ne doivent pas être diffusés. Les données sur les variables régionales directes, la nationalité, le pays de naissance et le groupe ethnique ne doivent pas être diffusées non plus.
3. On ne peut diffuser qu'une seule sorte de variables régionales indirectes (p. ex. la catégorie de taille du lieu de résidence). Les combinaisons des valeurs des variables régionales indirectes devraient être suffisamment dispersées, autrement dit chaque région que l'on peut discerner devrait avoir une population cible comptant au moins 200 000 personnes et, de surcroît, devrait être formée de municipalités appartenant à au moins 6 des 12 provinces des Pays-Bas. Le nombre d'habitants d'une municipalité comprise dans une région que l'on peut discerner devrait être inférieur à 50% du nombre total d'habitants dans la région en question.
4. Le nombre de variables identifiantes contenues dans les microdonnées ne doit pas excéder 15.
5. Les données sur les variables sensibles ne doivent pas être diffusées.
6. Il doit être impossible d'obtenir des renseignements identifiants supplémentaires d'après les poids d'échantillonnage.
7. Au moins 200 000 membres de la population devraient fournir une réponse pour chaque valeur d'une variable identifiante.
8. Au moins 1 000 membres de la population devraient fournir une réponse pour chaque valeur du recoupement de deux variables identifiantes.
9. Pour chaque ménage dans lequel plus d'une personne a participé à l'enquête, il faut que le nombre total de ménages qui correspond à une combinaison particulière de valeurs des variables ménage soit au moins 5 dans les microdonnées.
10. Les enregistrements de microdonnées doivent être diffusés dans un ordre aléatoire.

Selon cet ensemble de règles, les données comprises dans les fichiers publics sont protégées de façon nettement plus rigoureuses que les microdonnées pour la recherche. Notons que, pour ces dernières, il est nécessaire de vérifier certaines combinaisons trivariées de valeurs des variables identifiables, alors que pour les fichiers publics de microdonnées, il suffit de vérifier les combinaisons bivariées. Cependant, il n'est pas permis de diffuser les variables régionales directes dans les fichiers publics. Si aucune donnée sur des variables régionales directes n'est diffusée dans un ensemble de microdonnées pour la recherche, seules certaines combinaisons bivariées de valeurs de variables identifiables doivent être vérifiées conformément aux règles de contrôle de la divulgation statistique. Pour les fichiers publics de microdonnées correspondants, toutes les combinaisons bivariées de valeurs de variables identifiables doivent être vérifiées.

Le progiciel µ-ARGUS aide à repérer et à protéger les combinaisons posant un risque de divulgation dans le fichier de microdonnées souhaité. Donc, le progiciel permet de vérifier l'application de la règle 2 aux microdonnées pour la recherche et des règles 7 et 8 aux fichiers publics de microdonnées. Le recodage global et la suppression locale sont deux techniques de protection des données utilisées pour produire des fichiers de microdonnées ne posant pas de risque de divulgation. Dans le cas du recodage global, on regroupe plusieurs catégories d'une variable identifiante en une seule. On applique cette technique à l'ensemble complet de données, plutôt qu'à la partie posant un risque de divulgation uniquement, afin d'obtenir une catégorisation uniforme de chaque variable identifiante.

Dans le domaine des microdonnées, plusieurs nouvelles techniques sont étudiées à l'heure actuelle dans le cadre du projet sur les aspects informatiques de la confidentialité des statistiques, ou projet CASC. De nouvelles méthodes, comme la randomisation a posteriori (PRAM), la microagrégation et l'ajout d'un bruit seront mises en œuvre dans les nouvelles versions de µ-ARGUS qui seront diffusées dans un avenir proche. La randomisation a posteriori est une méthode de perturbation permettant de protéger la divulgation de données sur des variables nominales (voir, p. ex., Gouweleeuw, Kooiman, Willenborg et De Wolf, 1998). L'application de la méthode de randomisation a posteriori signifie que, pour chaque enregistrement d'un fichier de microdonnées, le score sur une ou plusieurs variables nominales identifiantes peut être classifié erronément dans divers scores selon un mécanisme probabiliste prédéterminé (Van den Hout et Van der Heijden, 2002). Puisque le fichier de données original est perturbé, il est difficile pour un intrus de déterminer avec certitude si certains enregistrements correspondent à des membres particuliers de la population. Autrement dit, le caractère aléatoire de la procédure sous-entend que l'appariement d'un enregistrement du fichier perturbé à un enregistrement d'un individu connu dans la population a une forte probabilité d'être incorrect. Par conséquent, les enregistrements du fichier original sont protégés, ce qui est l'objectif principal de l'application de la randomisation a posteriori. Par contre, puisqu'on connaît le mécanisme probabiliste utilisé pour appliquer la randomisation, on peut estimer les caractéristiques des données réelles d'après le fichier de données perturbé. Donc, il est encore possible de faire toutes sortes d'analyses statistiques après la randomisation a posteriori. Cependant, l'utilisation de la matrice de transition contenant les probabilités de classification erronée afin de tenir compte de la perturbation due à la randomisation a posteriori demande un effort supplémentaire et devient plus compliquée à mesure qu'augmente la complexité des questions de recherche. Deux questions importantes que soulève la randomisation a posteriori sont les suivantes :

- Comment faut-il choisir les probabilités de classification erronée afin que le fichier de microdonnées ne présente pas de risque de divulgation?
- Comment faut-il rajuster les analyses statistiques pour tenir compte des probabilités de classification erronée?

La première question est abordée dans Willenborg et De Waal (2001) et la seconde dans Van den Hout et Van der Heijden (2002).

L'implémentation des nouvelles méthodes dans µ-ARGUS permettront d'expérimenter ces techniques. Dans l'avenir, on pourra appliquer une combinaison de plusieurs méthodes de protection contre la divulgation, par exemple une combinaison de randomisation a posteriori, de recodage global et de suppression locale. Généralement parlant, à l'heure actuelle, les répercussions en ce qui concerne les règles de contrôle de la divulgation statistique appliquées dans le domaine des statistiques officielles ne sont pas claires.

Pour évaluer la qualité des méthodes appliquées, on intégrera également des modèles de risque de divulgation et de perte d'information dans les nouvelles versions de µ-ARGUS. On spécifie un modèle de risque de divulgation pour

faire la distinction entre les microdonnées sûres de celles qui posent un risque de divulgation. Le degré de complexité des modèles de divulgation peut être fort variable. Dans un modèle de divulgation assez simple, une combinaison de valeurs est sûre uniquement si la fréquence estimée de son occurrence dans la population est supérieure à une valeur-seuil donnée. Les combinaisons qu'il convient de considérer font également partie du modèle de risque de divulgation appliqué et doivent être précisées par la personne qui protège les données. Fienberg et Makov (1998) et Skinner et Holmes (1998) ont proposé des modèles de risque de divulgation plus avancés. Ils utilisent des modèles log-linéaires pour estimer le risque individuel. Quel que soit le modèle de risque de divulgation utilisé, on doit systématiquement émettre des hypothèses quant à la nature des attaques visant les renseignements sur une entité particulière que pourraient perpétrer des intrus (voir, p. ex., Keller et Bethlehem, 1992, Mokken et coll., 1992, Elliot et Dale, 1999 et Elliot, 2001).

Si des microdonnées posant un risque de divulgation doivent être transformées en microdonnées sûres, il est nécessaire de disposer d'une mesure de la perte d'information. On utilise cette mesure pour limiter l'altération des microdonnées due aux mesures prises pour les protéger. Dans le cas de suppressions locales, μ-ARGUS utilise le nombre de ces suppressions comme mesure de la perte d'information. Plus le nombre de suppressions est élevé, plus la perte d'information est grande. Le problème d'optimisation qu'il faut résoudre consiste à sélectionner les suppressions locales de telle façon que les microdonnées résultantes ne posent pas de risque de divulgation et que la perte d'information connexe soit réduite au minimum. Comme μ-ARGUS utilise le nombre de suppressions locales comme mesure de la perte d'information, la question de savoir comment faire les suppressions de façon optimale peut être résolue du point de vue des enregistrements (De Waal et Willenborg, 1998). Une autre possibilité consiste à choisir le nombre de catégories distinctes visées par la suppression comme mesure de la perte d'information. Le problème de minimisation a alors tendance à être plus difficile à résoudre en pratique. Dans certains cas, on peut le décomposer en un nombre de problèmes plus petits et, par conséquent, plus faciles à résoudre. Dans le cas du recodage global d'une variable identifiante, la perte d'information dépend de l'évaluation de l'importance de la variable et de l'évaluation de chaque codage prédéfini possible pour la variable. Hurkens et Tiourine (1998) ont élaboré des modèles d'optimisation pour une forme spéciale de recodage global. Ce dernier et la suppression locale entraînent l'un et l'autre une perte d'information, parce que dans le premier cas moins d'information est fournie et dans le second, une partie de l'information n'est pas fournie du tout. Il convient donc de trouver systématiquement le juste équilibre entre le recodage global et la suppression locale afin de réduire autant que possible la perte d'information due aux mesures de contrôle de la divulgation statistique. Il est recommandé de commencer par recoder globalement certaines variables jusqu'à ce que le nombre de combinaisons dangereuses qu'il convient de protéger soit suffisamment faible. Puis, les combinaisons dangereuses qui persistent doivent être protégées par suppressions locales.

4. AUTRES MÉTHODES PERMETTANT D'UTILISER LES DONNÉES

Toutes les techniques décrites au chapitre 3 comprennent nécessairement la manipulation ou la suppression de données et, par conséquent, réduisent vraisemblablement la qualité des estimations produites d'après les données ainsi traitées. Par conséquent, les instituts nationaux de la statistique (INS) ont entrepris d'étudier d'autres méthodes permettant d'utiliser les données tout en protégeant la confidentialité des renseignements sensibles fournis par les répondants. Ces méthodes permettent d'utiliser les données dans un environnement contrôlé par l'INS et imposent aux utilisateurs les mêmes obligations légales et éthiques de protection des données que celles imposées à l'INS proprement dit.

Certains INS (p. ex. aux États-Unis) ont adopté le processus d'octroi de licence en vertu duquel les établissements et les chercheurs ne faisant pas partie de l'INS obtiennent provisoirement l'accès à une partie des données dans leurs locaux, à condition qu'ils signent une entente indiquant qu'ils se conformeront aux exigences légales concernant la protection des données imposée à l'INS. L'octroi d'une licence d'utilisation des données est donc un moyen de donner accès à ces dernières quand elles ne peuvent être diffusées publiquement pour des raisons de confidentialité. Les sites auxquels sont octroyées les licences doivent faire l'objet d'inspections périodiques. Pour que les accords d'octroi de licence portent leurs fruits, il est aussi nécessaire que les fichiers de l'INS visés par le contrat de licence soient bien organisés.

La modalité d'accès aux données la plus importante mise au point au cours de la dernière décennie est sans doute la création de sites d'accès restreint qui permettent aux INS de répondre aux besoins de microdonnées des chercheurs. En effet, certains chercheurs ont besoin de plus d'information que celle disponible dans les ensembles de microdonnées diffusées pour les chercheurs ou dans les fichiers publics de microdonnées. Comme il n'est pas permis de diffuser des données plus riches, certains chercheurs reçoivent l'autorisation d'effectuer leur recherche sur de telles microdonnées dans les locaux de l'INS. Statistics Netherlands est l'un des INS qui offrent ce genre d'option. Les chercheurs de bonne foi ont la possibilité de travailler sur place, dans une aire sécurisée installée dans les locaux de l'organisme. Ils ont le choix entre deux emplacements, soit Voorburg dans l'ouest du pays et Heerlen dans le sud. Ils ne peuvent toutefois exporter des données sans la permission de l'agent statistique responsable. Ils peuvent appliquer les progiciels statistiques standards, ainsi que leurs propres programmes. À l'instar de tous les employés de Statistics Netherlands, les personnes qui travaillent sur place doivent prêter serment de ne pas divulguer les renseignements fournis par les répondants individuels (Kooiman, Nobel et Willenborg, 1999).

Les chercheurs qui utilisent sur place les données économiques de Statistics Netherlands doivent tenir compte des règles établies par le centre de recherche sur les microdonnées économiques. Les plus importantes de ces règles sont les suivantes :

- le chercheur doit être affilié à un établissement de recherche reconnu (p. ex. une université);
- le chercheur doit rédiger une proposition de recherche conforme aux normes scientifiques courantes;
- le chercheur et son supérieur doivent signer une déclaration de discrétion;
- le chercheur obtient uniquement l'accès aux données nécessaires pour la réalisation de son projet;
- les données ne contiennent pas d'information sur la raison sociale et l'adresse des entreprises;
- les données ayant trait aux deux années les plus récentes ne sont pas fournies;
- il est strictement interdit de sortir les données ou des résultats intermédiaires non protégés des locaux de Statistics Netherlands;
- toutes les publications éventuelles seront soumises à un examen visant à déterminer le risque de divulgation;
- toutes les publications doivent appartenir au domaine public;
- un registre public contient le nom du chercheur, le titre du projet de recherche, le titre de la ou des publications et le nom des bases de données fournies.

Statistics Netherlands n'offrent pas les installations susmentionnées gratuitement. Il est de règle que le chercheur paye les frais qu'occasionne la fourniture des données demandées. De plus, il y a un tarif pour l'utilisation des installations sur place.

Enfin, une dernière option consiste à accorder le téléaccès. Cette modalité d'accès aux données combine les avantages de l'octroi d'une licence d'utilisation des données qui permet aux chercheurs de travailler dans leur propre établissement aux avantages du travail sur place du fait que les données ne quittent pas les locaux de l'INS. Normalement, les chercheurs obtiennent l'accès aux données par l'entremise d'un intermédiaire contrôlé par l'INS qui garantit que toutes les utilisations sont conformes à la loi. L'option de la téléexécution représente un pas de plus. Dans ces conditions, il n'existe aucun intermédiaire entre le chercheur et l'INS. L'option de téléexécution permet au chercheur d'exécuter des formatages sans devoir télécharger les données sur son propre ordinateur. Bien que cette option soit plus efficace que le téléaccès, il reste à prouver que les systèmes de sécurité sont suffisamment puissants pour permettre d'utiliser cette technique fréquemment. À l'heure actuelle, le centre de recherche sur les politiques de Statistics Netherlands réalise, en collaboration avec le ministère des Affaires sociales et de l'Emploi, un projet pilote qui donne de bons résultats. Dans le cadre de ce projet, la téléexécution est limitée en ce sens que les employés de Statistics Netherlands continuent de vérifier manuellement les formatages qui sont transmis au centre de recherche sur les politiques.

5. DISCUSSION ET CONCLUSION

Les progiciels τ-ARGUS et μ-ARGUS sont le fruit du projet relatif au contrôle de la divulgation des statistiques (CDS) réalisé dans le cadre du quatrième Programme-cadre RTD de la Communauté européenne. Ces progiciels

semblent faciliter beaucoup l'application pratique du contrôle de la divulgation statistique. Ils permettent de résoudre nombre des problèmes de protection des données statistiques.

Les manuels (Hundepool et coll., 2002a et b) qui accompagnent les progiciels ARGUS sont fort utiles pour les utilisateurs. Cependant, ceux-ci souhaitent sans cesse obtenir des fonctions supplémentaires. Dans le cas de τ-ARGUS, il faut continuer de rechercher des moyens de traiter efficacement, de façon automatisée, les tableaux couplés et les métadonnées correspondantes. Il faut également approfondir l'étude des moyens de protéger contre la divulgation les données d'une même enquête recueillies pour des années consécutives. Dans le cas de μ-ARGUS, il est important de rendre plus claire dans le progiciel la différence entre la protection des microdonnées pour la recherche et celle des fichiers publics de microdonnées. Comme μ-ARGUS permet d'utiliser un grand nombre de critères de protection différents, il importe d'aider les utilisateurs à comprendre comment appliquer diverses stratégies en se servant du progiciel. Récemment, on s'est penché sur l'étude de méthodes de perturbation par ajout d'un bruit stochastique aux microdonnées. Il serait intéressant d'offrir dans μ-ARGUS plusieurs options de perturbation des données comme méthode de protection. Il est particulièrement intéressant pour un utilisateur de savoir dans quelle mesure les microdonnées sont protégées après le processus de perturbation.

Nous pouvons conclure que beaucoup de recherche reste à faire dans le domaine du contrôle de la divulgation statistique. Avec un peu de chance, de nouvelles versions des progiciels ARGUS (qui incluent les résultats des études en cours) seront bientôt mises à la disposition des utilisateurs des données. La production de ces nouvelles versions fait partie du projet CASC (voir Hundepool, 2001). Le projet AMRADS (mesures complémentaires en recherche et développement dans le domaine de la statistique) financé par le cinquième Programme-cadre RTD de la Communauté européenne a été entrepris en vue de promouvoir les résultats des projets statistiques réalisés aux termes du quatrième Programme-cadre RTD. Nombre de cours et de conférences portant, entre autres, sur le contrôle de la divulgation statistique sont organisés. Ces activités feront progresser la mise en application des méthodes de contrôle de la divulgation statistique dans de nombreux pays.

Quelques instituts nationaux de la statistique (INS) et bureaux d'études de marché entreprennent leurs propres projets de recherche (voir, p. ex., Bethlehem et Pannekoek, 1998). À part ces INS, Eurostat, qui est l'organisme statistique de la Communauté européenne, est devenu l'un des principaux promoteurs de la recherche en statistique. Ces dernières années, des fonds ont été réservés pour le financement de projets de recherche et de développement ciblés. Nombre de projets (p. ex. les projets CDS, CASC et AMRADS mentionnés dans le présent article) ont été subventionnés par la Communauté européenne. Avant que ne débute le cinquième Programme-cadre RTD de la Communauté européenne, le Plan européen de recherche dans le domaine des statistiques officielles (EPROS) a été lancé. Le contrôle de la divulgation statistique est l'un des sujets mentionnés dans ce plan. Eurostat a favorisé la création de groupes de chercheurs provenant d'universités, d'INS et de bureaux d'études de marché afin de permettre l'échange des idées et l'interapprentissage. Les projets subventionnés ne donnent pas tous des résultats applicables en pratique. Cependant, il est difficile de prédire lesquels seront les plus fructueux. Les facteurs critiques de réussite sont, en tout état de cause, une définition précise du but visé et une organisation efficace. Heureusement, Eurostat (et peut-être d'autres organismes internationaux) continuera de trouver des moyens de favoriser la recherche en statistique dans le futur. Bien qu'on ne sache jamais au départ quelle sera l'issue exacte des projets de recherche, il est évident que le financement des projets de recherche en statistique a donné lieu à des économies d'échelle et a accéléré le processus de production de statistiques de meilleure qualité et plus comparables.

Dans le présent article, nous avons décrit des méthodes qui ont été mises au point pour protéger la confidentialité des données tout en permettant l'accès à ces dernières grâce à divers moyens qui soit altèrent les données, soit en limitent l'accès. Trouver le juste équilibre entre la protection de la confidentialité des données et l'accès aux données est un exercice délicat. Souhaitons que les nouvelles méthodes de recherche et les logiciels de contrôle de la divulgation statistique aident à trouver ce juste équilibre.

Dans le cadre du projet CASC, de nouvelles versions des progiciels ARGUS sont mises à la disposition des utilisateurs de données. Récemment, nous avons diffusé de nouveaux manuels concernant τ-ARGUS et μ-ARGUS (Hundepool et coll., 2002a et b) qui ont fait l'objet d'examens approfondis dans le cadre du projet CASC. Les deux manuels sont très utiles aux personnes chargées de conduire les tests. La nouvelle version de τ-ARGUS comprend la fonction HiTaS qui permet de traiter les tableaux hiérarchiques. Les options supplémentaires offertes dans la

nouvelle version de μ-ARGUS sont la randomisation a posteriori (PRAM) et les modèles individuels de risque. Les progiciels ARGUS ont évolué vers l'utilisation d'interfaces comportant plusieurs moteurs de pointe produits par des statisticiens de divers pays.

RÉFÉRENCES

Bethlehem, J. et J. Pannekoek (1998), "Statistical research activities at Statistics Netherlands", *Research in Official Statistics*, 1, pp. 131-134.

Citteur, C.A.W. et L.C.R.J. Willenborg (1993), "Public use microdata files: current practices at national statistical bureaus", *Journal of Official Statistics*, 9, pp. 783-794.

Elliot, M. (2001), "Advances in data intrusion simulation: a vision for the future of data release", *Statistical Journal of the United Nations Economic Commission for Europe*, 18, pp. 383-391.

Elliot, M. et A. Dale (1999), "Scenarios of attack: the data intruder's perspective on statistical disclosure risk", *Netherlands Official Statistics*, 14, pp. 6-10.

Fienberg, S.E. et U.E. Makov (1998), "Confidentiality, uniqueness and disclosure limitation for categorical data", *Journal of Official Statistics*, 14, pp. 385-397.

Gouweleeuw, J.M., P. Kooiman, L.C.R.J. Willenborg et P.-P. de Wolf (1998), "Post randomisation for statistical disclosure control: theory and implementation", *Journal of Official Statistics*, 14, pp. 463-478.

Groot, A. et C.A.W. Citteur (1997), "Accessibility of business microdata", *Netherlands Official Statistics*, 12, pp. 18-32.

Hout, A. van den et P.G.M. van der Heijden (2002), "Randomised response, statistical disclosure control and misclassification: a review", *International Statistical Review*, 70, pp. 269-288.

Hundepool, A.J. (2001), "Computational aspects of statistical confidentiality: the CASC-project", *Statistical Journal of the United Nations Economic Commission for Europe*, 18, pp. 315-320.

Hundepool, A., A. van de Wetering, P.P. de Wolf, S. Giessing, M. Fischetti, J.J. Salazar et A. Caprara (2002a), *τ-ARGUS, user's manual, version 2.1*, Voorburg, The Netherlands: Statistics Netherlands.

Hundepool, A., A. van de Wetering, L. Franconi, A. Capobianchi et P.P. de Wolf (2002b), *μ-ARGUS, user's manual, version 3.1*, Voorburg, The Netherlands: Statistics Netherlands.

Hurkens, C.A.J. et S.R. Tiourine (1998), "Models and methods for the microdata protection problem", *Journal of Official Statistics*, 14, pp. 437-447.

Keller, W.J. et J.A. Bethlehem (1992), "Disclosure protection of microdata: problems and solutions", *Statistica Neerlandica*, 46, pp. 5-19.

Kooiman, P., J.R. Nobel et L.C.R.J. Willenborg (1999), "Statistical data protection at Statistics Netherlands", *Netherlands Official Statistics*, 14, pp. 21-25.

Mokken, R.J., P. Kooiman, J. Pannekoek et L.C.R.J. Willenborg (1992), "Disclosure risks for microdata", *Statistica Neerlandica*, 46, pp. 49-67.

Schulte Nordholt, E. (2001), "Statistical disclosure control (SDC) in practice: some examples in official statistics of Statistics Netherlands", *Statistical Journal of the United Nations Economic Commission for Europe*, 18, pp. 321-328.

Skinner, C.J. et D.J. Holmes (1998), "Estimating the re-identification risk per record in microdata", *Journal of Official Statistics*, 14, pp. 361-372.

Waal, T. de et L.C.R.J Willenborg (1998), "Optimal local suppression in microdata", *Journal of Official Statistics*, 14, pp. 421-435.

Willenborg, L.C.R.J. (1993), "Discussion statistical disclosure limitation", *Journal of Official Statistics*, 9, pp. 469-474.

Willenborg, L.C.R.J. et T. de Waal (1996), *Statistical disclosure control in practice*, *Lecture Notes in Statistics 111*, New York: Springer-Verlag.

Willenborg, L.C.R.J. et T. de Waal (2001), *Elements of statistical disclosure control*, *Lecture Notes in Statistics 155*, New York: Springer-Verlag.