Nº de catalogue 94-14

CONFIDENTIALITÉ DESMICRODONNÉES DE L'EDTR : APPROCHE GÉNÉRALE POUR LAMISE EN OEUVRE

Juillet 1994

Pierre Lavallée, Division des méthodes d'enquê tes sociales

Chantal Grondin, Division des méthodes d'enquê tes sociales

La série de documents de recherche de l'EDTR est conçue en vue de communiquer les résultats des études ainsi que les décisions importantes ayant trait à l'Enquê te sur la dynamique du travail et du revenu. Ils sont offerts en français et en anglais. Pour obtenir une description sommaire des documents disponibles ou un exemplaire de ces documents, communiquez avec Philip Giles, EDTR, par la poste à Édifice Jean-Talon, 11ième étage, section D8, Statistique Canada, Ottawa (Ontario), Canada, K1A 0T6; par INTERNET: GILES@STATCAN.CA; par téléphone au (613) 951-2891; ou par télécopieur au (613) 951-3253.

SOMMAIRE

Pour permettre la production d'analyses longitudinales, l'Enquête sur la dynamique du travail et du revenu (EDTR) compte diffuser des fichiers de microdonnées. Ceci entraînera inévitablement un problème relié à la confidentialité des données. Il importe donc de développer une approche générale afin que le risque de divulgation de l'identité des répondants de l'EDTR soit contrôlé.

Le document présente l'approche proposée pour le traitement de la confidentialité de l'EDTR. On donne en premier lieu un bref survol de certains principes de confidentialité où l'on explique, entre autres, la classification des variables en identifiants directs, en identifiants indirects et en variables sensibles. On passe ensuite à la description du traitement pour la confidentialité de ces identifiants et variables sensibles. Finalement, on présente la classification générale des variables que l'on propose pour l'EDTR.

TABLE DES MATIÈRES

1. INTRODUCTION

L'Enquête sur la dynamique du travail et du revenu (EDTR) porte sur les particuliers et les ménages. Elle sert à observer dans le temps leur activité sur le marché du travail ainsi que les changements touchant leur revenu et leur situation familiale. La population cible de l'EDTR touche toutes les personnes, sans distinction d'âge, vivant dans les provinces canadiennes. On exclut les territoires, les institutions, les réserves indiennes et les camps militaires pour des raisons d'ordre opérationnel. L'enquête vise d'abord à fournir des données longitudinales bien que des estimations transversales seront aussi produites.

En janvier 1993, on a sélectionné un échantillon préliminaire de 15,000 ménages tirés de deux groupes de renouvellement de l'Enquête sur la population active (EPA). Cet échantillon constitue la base de l'échantillon longitudinal de l'EDTR. Les individus de cet échantillon ont reçu en janvier 1993 un questionnaire préliminaire portant sur des variables démographiques, sur le travail et sur l'historique des répondants.

En janvier 1994 a eu lieu la première entrevue sur la mesure de l'activité sur le marché du travail. Cette entrevue a principalement porté sur l'établissement des épisodes de travail et de chômage des répondants à l'EDTR en rapport avec 1993. En mai 1994, on a procédé à la mesure du revenu et de ses composantes toujours en rapport avec l'année 1993. Ces entrevues sur le travail et le revenu sont prévues pour une période de six ans.

Pour permettre la production d'analyses longitudinales, l'EDTR compte diffuser des fichiers de microdonnées. On compte ainsi, dans un premier temps, rendre public un fichier de microdonnées avec les résultats de l'entrevue préliminaire et des entrevues sur le travail et le revenu de 1994. On diffusera par la suite des

fichiers de microdonnées cumulées pour chacune des six vagues d'enquête; soit de 1994 à 1999.

Trois types de variables se retrouveront au sein des fichiers de microdonnées: les variables fixes, les variables annuelles et les variables dynamiques. Les variables fixes comprennent les caractéristiques qui ne changent pas au cours du temps. Les variables "sexe" et "âge à l'entrevue préliminaire" peuvent être considérées comme fixes. Dans les variables annuelles, on retrouve les données provenant du questionnaire sur le revenu et toutes les variables dérivées amenant une information annuelle comme, par exemple, l'indicateur "travail ou non au cours de l'année". Finalement, parmi les variables dynamiques, on retrouve l'ensemble des vecteurs de transitions mesurés par l'EDTR.

La diffusion des microdonnées entraînera inévitablement le problème relié à la confidentialité des données. Le but de ce document est de proposer une approche générale afin de traiter le problème de la confidentialité lors de la diffusion des microdonnées de l'EDTR. On donnera en premier lieu un bref survol de certains principes de confidentialité où l'on expliquera, entre autres, la classification en identifiants directs, en identifiants indirects et en variables sensibles. On passera ensuite au traitement pour la confidentialité de ces identifiants et variables sensibles. Par la suite, on présentera la classification générale des variables que l'on propose pour l'EDTR.

2. PRINCIPES DE CONFIDENTIALITÉ

On a déjà résumé dans Lavallée (1994) des généralités concernant la confidentialité. Ces généralités étaient tirées de Schackis (1993). Dans la présente section, on rappelle certains principes principalement tirés de Lavallée (1994)

nécessaires à la mise en oeuvre des mesures de confidentialité pour les microdonnées de l'EDTR.

Il y a un problème de divulgation s'il est possible d'obtenir une estimation jugée trop précise d'une information confidentielle. On retrouve trois types de divulgation: la divulgation par identification, la divulgation d'attributs et la divulgation inférentielle. Quoique la dernière possède une certaine importance, on ne discutera ici que des deux premiers types.

Il y a divulgation par identification lorsqu'une relation bijective est établie entre les données diffusées et un répondant précis. Notons que cette bijection peut être réelle ou seulement perçue comme telle; ces deux situations étant toutes deux dommageables pour les répondants concernés. Il est primordial de noter que l'identification est impossible lorsque le répondant n'est pas unique au sein de la population.

On se retrouve devant une divulgation d'attributs lorsque l'on détermine avec une certaine précision différents attributs d'un répondant à partir des données diffusées. Par exemple, si l'on constate que tous les individus d'un certain groupe âge-sexe possèdent, par exemple, un salaire donné, le salaire d'un individu de ce groupe âge-sexe sera ainsi automatiquement divulgué. Il n'y aura pas identification mais la vie privée de cet individu sera menacée par le simple fait d'avoir diffusé certaines données. Bien que la divulgation d'attributs constitue un problème, son importance est jugée moindre par rapport à la divulgation par identification.

On distingue trois types de données confidentielles: les petits effectifs, les cas de prédominance et les cas obtenus par dérivation. Dans les cas des petits effectifs, il y a trop peu de répondants dans certains sous-ensembles de répondants pour assurer leur anonymat. Quand il s'agit d'un échantillon, le petit effectif doit aussi

correspondre à un petit effectif au sein de la population. Dans les cas de prédominance, on peut avoir un gros effectif pour un sous-ensemble de répondants donnés. Cependant l'importance d'un ou deux répondants fait en sorte que ceux-ci contribuent à un pourcentage jugé trop grand des statistiques produites. Dans les cas obtenus par dérivation, on peut avoir un problème divulgation en combinant un certain nombre de sources d'information comme plusieurs fichiers de microdonnées. On identifie ainsi des répondants en "dérivant" leurs données.

Les microdonnées correspondent aux enregistrements relatifs aux répondants de l'enquête. On peut diviser chaque enregistrement du fichier de microdonnées en trois parties:

i) Les identifiants directs:

Cette partie comprend tout ce qui peut identifier le répondant sans ambiguïté comme, par exemple, le nom, l'adresse, le numéro de téléphone, le numéro d'individu dans l'enquête.

ii) Les identifiants indirects:

Les identifiants de cette partie ne peuvent être utilisés de façon directe dans le sens où l'on ne peut identifier un répondant sans une certaine connaissance de ses caractéristiques. Leur valeur est généralement facile à obtenir auprès des répondants. Des exemples d'identifiants indirects sont l'âge, le sexe, l'état matrimonial, la région de domicile, l'occupation.

iii) Les variables sensibles:

Cette partie comprend en fait le reste des variables de l'enregistrement. Elles sont dites sensibles parce qu'elles représentent de l'information appartenant à la vie privée du répondant.

Les identifiants directs et indirects sont aussi appelés variables clés. On note que la frontière entre les variables clés et sensibles est relativement arbitraire. Celle-ci dépend en fait du niveau de confidentialité que l'on désire associer à chaque variable. Par exemple, on peut décider de classifier le revenu comme variable sensible si on est prêt à considérer que le revenu ne peut être utilisé de façon pratique pour identifier un individu. De façon générale, les variables clés, en particulier les identifiants indirects, correspondent à de l'information facilement disponible auprès du public. Par contre, les variables sensibles sont plus difficilement disponibles parce qu'elles appartiennent à la vie privée des répondants.

On dit qu'un répondant est unique au sein de la population si ses identifiants indirects sont uniques. En tirant un échantillon, si un répondant unique dans la population est sélectionné, il sera aussi unique au sein de l'échantillon. Le contraire n'est pas nécessairement vrai de sorte qu'après avoir noté l'unicité d'un répondant dans l'échantillon, on doit aussi faire une vérification au sein de la population. Une telle vérification peut se faire, par exemple, à l'aide de données administratives. Notons qu'un répondant qui n'est pas unique au sein de l'échantillon n'est évidemment pas unique au sein de la population.

Le risque de divulgation par identification est défini comme étant la probabilité qu'un fureteur identifie au moins une personne contenue dans le fichier diffusé de microdonnées. Ce risque est fonction de l'information véhiculée par les identifiants

indirects et par les connaissances *a priori* du fureteur. Le risque de divulgation d'attributs est la probabilité que des attributs d'un individu soient dévoilés suite à la diffusion d'identifiants indirects.

3. TRAITEMENT DES IDENTIFIANTS DIRECTS, DES IDENTIFIANTS INDIRECTS ET DES VARIABLES SENSIBLES

La confidentialité de l'EDTR sera traitée avec des méthodes de contrôle de la divulgation. Dans le cadre de la diffusion de microdonnées, on retrouve deux différentes approches. La première est de contrôler la divulgation en réduisant l'information apportée par les données. On parle ici de méthodes de contrôle de la divulgation par réduction des données. La deuxième approche consiste à modifier les données de sorte que l'identification d'un répondant devienne impossible. Cette approche dite de modification des données est généralement employée après la réduction des données.

Comme l'EDTR est une enquête longitudinale visant principalement à mesurer des changements, on privilégie les méthodes basées sur la réduction des données plutôt que celles basées sur la modification des données. En effet, en plus de perturber la relation entre les variables d'une même vague, les méthodes de modification des données peuvent dans certains cas modifier artificiellement les transitions au sein des données et ainsi fausser les analyses de changements. Pour les enquêtes longitudinales, ces méthodes doivent en fait tenir compte, en plus de la relation entre les variables d'une même vague, des valeurs des vagues précédentes et subséquentes. Par exemple, une des méthodes les plus employées pour la modification des données est la perturbation aléatoire où l'on introduit un bruit (généralement de moyenne zéro) au sein des valeurs mesurées. Généré sans tenir compte des autres vagues, le bruit introduit peut alors influencer les changements d'une vague à l'autre. Il peut de plus détruire la cohérence entre les changements

pour différentes variables. Il est malheureusement généralement difficile de modifier les données tout en respectant la contrainte de ne pas fausser les analyses de changements; c'est pourquoi l'effort déployé pour la confidentialité de l'EDTR est au départ axé sur les la réduction des données.

Le traitement pour la confidentialité des identifiants directs, indirects et des variables sensibles devra se faire selon une optique de six ans. Il faudra ainsi envisager la confidentialité non pas comme une tâche progressive année après année mais plutôt comme un tout. Le problème ici est que la quantité d'information sur un individu augmente avec le temps. Si on applique des méthodes de confidentialité sur le fichier de la première année, celles-ci risquent de ne plus être suffisantes pour le fichier contenant les six années. On fait alors face à un gros problème parce que le fichier de la première année a déjà été diffusé; c'est pourquoi l'approche de six ans est importante. Notons que cette approche pose un problème opérationnel puisque les méthodes de contrôle de la divulgation doivent être déterminées pour des données qui n'ont pas encore été recueillies.

L'étude de la confidentialité sur six ans n'est pas requise pour l'ensemble des variables de l'EDTR. Pour les variables fixes, le traitement pour six ans sera en fait le même que pour une seule année puisque, par définition, ces variables ne changent pas au cours du temps. Pour les variables annuelles et dynamiques, la difficulté dépendra de la classification de chaque variable en identifiant indirect ou en variable sensible. Les variables annuelles ou dynamiques classifiées comme variables sensibles ne seront soumises à des méthodes de contrôle de la divulgation que de façon annuelle; ceci permettra d'effectuer le traitement pour la confidentialité pour une année donnée et non pas pour six ans. On verra plus loin que l'on peut mettre beaucoup moins d'efforts sur la confidentialité des variables sensibles par rapport aux identifiants indirects. Par contre, les variables annuelles ou dynamiques classifiées comme identifiants indirects devront faire l'objet d'études

plus poussées basées sur des données existantes provenant de l'Enquête sur l'activité ou des données fiscales. Par ces études, on tentera de déterminer à quel niveau de changements on peut s'attendre pour une période de six ans afin de déterminer un contrôle acceptable de la divulgation. Les variables fixes, annuelles ou dynamiques classifiées comme identifiants directs sont discutées dans la section suivante.

3.1 LES IDENTIFIANTS DIRECTS

Les identifiants directs des fichiers de microdonnées seront anonymisés ou sinon simplement enlevés. Les identifiants directs qui seront anonymisés correspondent à ceux qui apportent une certaine information relative au plan de sondage. On retrouve, par exemple, les numéros de strate et les numéros de composante (ou réplique) qui sont nécessaires au calcul des variances. Pour anonymiser ces identifiants, on remplacera simplement leur valeur par un nombre aléatoire.

Les autres identifiants directs (c'est-à-dire ceux qui n'apportent rien au niveau du plan de sondage) seront simplement enlevés. Ceux-ci comprennent, par exemple, le nom, le numéro de téléphone, l'adresse et les noms des employeurs.

3.2 LES IDENTIFIANTS INDIRECTS

Les identifiants indirects sont ceux qui posent le plus de problèmes à la question de la confidentialité. De par leur nature, ces identifiants apportent une certaine information qui justifie leur présence au sein du fichier de microdonnées. On ne peut donc pas, comme les identifiants directs, les anonymiser ou les enlever sans une perte d'information. Cependant, en les conservant dans les fichiers de microdonnées, on doit s'assurer que le risque de divulgation soit minime.

De l'approche par la réduction des données, on retient les méthodes de contrôle de la divulgation suivantes pour une application éventuelle aux identifiants indirects: l'établissement de restrictions sur les effectifs de la population, la réduction du détail, le codage des extrêmes et la suppression. Parmi les méthodes de contrôle de la divulgation par modification des données, on ne retient que la méthode dite de perturbation par compensation.

L'établissement de restrictions sur les effectifs de la population consiste à appliquer un seuil limite sur la taille de population minimale que peuvent représenter les microdonnées. On se base ici sur le fait que plus une population est grande, plus la chance de trouver deux enregistrements avec les mêmes identifiants indirects augmente. On cherche donc à définir un ou plusieurs seuils tels que si une population (ou sous-population) contient un effectif inférieur à ce seuil, on cherche alors à augmenter le contrôle sur la divulgation. Pour l'EDTR, on peut penser à appliquer un seuil minimum aux poids servant à l'estimation.

L'idée de la réduction du détail est de diminuer la quantité d'information contenue dans les microdonnées diffusées jusqu'à ce que l'on obtienne des jumeaux ou doubles (c'est-à-dire des individus avec la même combinaison d'identifiants indirects) au sein de la population. En pratique, il n'est pas essentiel que les jumeaux soient caractérisés par la totalité des identifiants indirects. On peut en fait concevoir l'utilisation d'un sous-ensemble seulement d'identifiants indirects pour caractériser des jumeaux. Par exemple, deux individus peuvent être jumeaux si trois de leurs identifiants indirects sont les mêmes. Un individu sera ainsi unique si l'ensemble de tous les trios d'identifiants indirects sont uniques. Cette approche permet de réduire le nombre d'identifiants indirects traités à la fois lors du traitement des données pour la confidentialité (voir Lebrasseur (1994)). Comme il a déjà été mentionné, des jumeaux au sein de l'échantillon garantissent nécessairement des jumeaux au sein de la population; ce qui élimine le risque de

divulgation par identification. On retrouve deux différentes versions de la méthode de la réduction des données: le changement d'échelle et la réduction du nombre de catégories. Avec le changement d'échelle, on passe généralement d'une valeur numérique à une valeur catégorielle (ou qualitative) ordonnée ou non.

Avec le codage des extrêmes, on cherche à réduire l'information des microdonnées en codifiant les valeurs relatives aux "queues" de la distribution. Ces valeurs correspondent en général à de très petits effectifs au sein de la population; ce qui rend le risque d'identification élevé. En prenant la variable "âge", par exemple, on peut ainsi créer une catégorie incluant les personnes de 70 ans et plus. Le codage des extrêmes est souvent effectué suite à l'utilisation de la méthode de la réduction du détail.

La méthode de la suppression consiste simplement à retirer des microdonnées certaines valeurs où le risque d'identification est extrême. On est alors en présence d'un répondant n'ayant vraisemblablement aucun jumeau au sein de la population. Il existe deux versions à cette méthode: l'omission des valeurs extrêmes et l'élimination de l'enregistrement entier.

Contrairement aux autres méthodes décrites ci-haut qui sont des méthodes dites de réduction des données, la perturbation par compensation relève de la modification des données. L'idée de base de cette méthode est de contrôler la divulgation en modifiant les microdonnées de sorte qu'une fois les identifiants directs enlevés (avec l'anonymisation), il n'y ait plus d'enregistrements avec une combinaison unique d'identifiants indirects et que les perturbations qui ne se neutralisent pas soient compensées de manière à conserver les tableaux de fréquences. On désire donc créer artificiellement des jumeaux en modifiant les identifiants indirects. Cependant, on veut aussi faire en sorte que les effectifs des tableaux de fréquences

soient les mêmes qu'avant les modifications. Pour plus de détail, on peut consulter Schackis (1993).

Suite à l'application des méthodes de contrôle de la divulgation aux identifiants indirects, on contrôlera aussi le risque de divulgation d'attributs pour certaines variables sensibles. Ce contrôle se fait en mesurant l'entropie d'une variable sensible pour la classe d'individus appartenant à une combinaison donnée d'identifiants indirects. Si l'entropie est faible, la variable sensible possède à peu près la même valeur pour tous les individus de la classe et il y a alors un risque de divulgation d'attributs. Il se peut, en effet, que pour un groupe âge-sexe donné, par exemple, que tous les individus aient le même revenu; on connaît alors le revenu des individus de ce groupe âge-sexe sans avoir à le mesurer. On corrige en général cette situation en appliquant de nouveau la méthode de réduction du détail aux identifiants indirects de manière à augmenter l'entropie de la variable sensible.

3.3 LES VARIABLES SENSIBLES

On peut mettre beaucoup moins d'efforts sur la confidentialité des variables sensibles par rapport aux identifiants indirects. Puisque ces données sont reliées à la vie privée des individus, elles peuvent en fait difficilement servir à l'identification de ceux-ci. Ceci est dû à la définition-même des variables sensibles. En effet, si la variable peut être utilisée pour identifier un individu à l'aide ou non d'une certaine connaissance raisonnable des caractéristiques de ces individus, celle-ci devrait alors être classifiée comme identifiant direct ou indirect. On procédera toutefois à certaines méthodes de contrôle de la divulgation visant à assurer un risque minimum d'identification.

La première méthode consistera à arrondir les valeurs monétaires. On désire ainsi éviter un appariement avec des fichiers de données administratives telles que les

données fiscales. Pour les données annuelles comme le revenu total, par exemple, leur valeur sera vraisemblablement arrondie au millier près. Pour les valeurs mensuelles, on propose d'arrondir au centième près de manière à assurer le plus possible la cohérence avec les valeurs annuelles. Au niveau des taux horaires comme le salaire, par exemple, on pourra arrondir au dollar près. Il est à noter que de récentes études (voir Idehem et Genest (1994)) tendent à démontrer que la puissance des tests statistiques ne diminue pas suite à un arrondissement des valeurs.

Une deuxième méthode pour contrôler le risque de divulgation consistera à effectuer de façon systématique la détection des valeurs aberrantes. On peut voir aisément, par exemple, qu'un salaire annuel de 2 millions de dollars peut rendre le risque de divulgation d'identité assez élevé. Pour chaque valeur aberrante, on procédera au codage de cette valeur extrême ou bien à sa suppression. Par exemple, supposons qu'un salaire annuel est jugé aberrant s'il est supérieur à 200,000 $. Le codage des valeurs extrêmes consistera alors à remplacer le salaire de 2 millions par un code indiquant que le salaire est supérieur à 200,000 $. On pourra aussi remplacer la vraie valeur par une statistique telle que la moyenne des valeurs de l'échantillon supérieures à 200,000 ou même simplement remplacer la vraie valeur par 200,000. Cette dernière pratique est connue sous le nom de winsorisation.

Tout le problème de la détection des valeurs aberrantes réside dans la définition que l'on utilisera pour celles-ci. Pour certaines variables sensibles, on pourra se contenter de méthodes univariées (tableaux de fréquences, "box plot", etc.). Pour d'autres cependant, il faudra étudier chaque variable en relation avec d'autres variables. Par exemple, un revenu de 100,000 $ dans un petit village peut être aberrant si le revenu moyen du village est de 20,000 $. Dans le cas des variables de transitions, ce ne sont souvent pas les valeurs elles-mêmes qui sont aberrantes

mais plutôt le nombre de transitions. Par exemple, le vecteur de transitions matrimoniales peut être aberrant si un individu est passé de célibataire à marié plusieurs fois par année.

4. CLASSIFICATION GÉNÉRALE DES VARIABLES DE L'EDTR

4.1 RÈGLES DE BASES

4.1.1 Lien entre les individus d'un même ménage

Afin de mieux contrôler le risque de divulgation, on compte éliminer du fichier de sortie le lien entre les individus d'un même ménage. Pour ce faire, les principales caractéristiques du ménage seront ramenées au niveau des individus. Ceci entraînera deux problèmes. Premièrement, il faudra procéder à l'identification d'un sous-ensemble des caractéristiques pouvant être utiles pour des fins d'analyse. Par exemple, des caractéristiques comme la taille et le revenu du ménage sont des variables généralement considérées comme essentielles. Par contre, on peut aisément penser à d'autres caractéristiques du ménage qui peuvent n'être utiles que dans des cas rares comme, par exemple, le nombre d'enfants du ménage provenant d'un mariage antérieur. Comme le nombre de caractéristiques du ménage peut être énorme, il faudra donc effectuer un choix judicieux. Notons qu'une première ébauche se retrouve dans Butlin (1994) mais que la liste donnée dans ce document pourrait être insuffisante si l'on élimine le lien entre les individus d'un même ménage.

Le deuxième problème est qu'il faudra s'assurer que les caractéristiques du ménage fournies au sein du fichier de sortie ne permettront pas de reconstituer, et ainsi d'identifier, les ménages. Une telle chose sera possible si la caractéristique du ménage associée à chaque individu du fichier de sortie est unique pour un ménage

en particulier. Par exemple, en associant la variable "taille du ménage" à chaque individu du fichier de sortie, on pourra reconstituer le seul ménage qui aura une taille donnée d'individus. Une des caractéristiques à laquelle il faudra apporter un soin particulier est le "poids transversal" qui sera vraisemblablement unique pour chaque ménage. Ce cas particulier est discuté dans la prochaine section. On devra aussi porter une attention particulière au revenu familial qui, lorsque calculé selon la méthode traditionnelle (c'est-à-dire en fonction de la composition du ménage défini à un point précis dans le temps), sera le même pour tous les individus d'une même famille économique. Notons que dans beaucoup de cas, l'arrondissement des valeurs monétaires et le codage des valeurs extrêmes permettra de solutionner ce problème de reconstitution des ménages.

4.1.2 Variables au niveau du ménage

a. Caractéristiques dérivées au niveau des ménages

A cause de l'élimination du lien entre les individus et les ménages, certaines caractéristiques au niveau du ménage ajoutées au fichier risquent de devenir des identifiants indirects pour les individus de ce ménage. Ainsi, toutes les variables au niveau du ménage dérivées à partir d'identifiants indirects seront elles aussi classifiées identifiants indirects. Par exemple, une variable définissant le type de famille et faisant allusion à l'état matrimonial et à l'âge des individus qui en font partie sera classifiée comme identifiant indirect.

b. Pondération transversale

La première étape de la pondération transversale est le calcul du poids de base. Ce poids représente en quelque sorte le poids de sondage ou encore le poids avant les ajustements tels que la post-stratification. Le poids de base est calculé en faisant la

moyenne de poids initiaux au niveau du ménage. Ces derniers correspondent de façon générale à l'inverse de la probabilité d'inclusion des individus. Cette méthode est appelée le "partage des poids" et, par définition, produit des poids identiques pour tous les membres d'un même ménage. La deuxième étape du processus de pondération transversale est la post-stratification. On utilise pour celle-ci l'approche intégrée qui garantie l'obtention d'un poids final unique pour tous les membres du même ménage.

Parce que les poids initiaux ont tendance à être différents d'un individu à un autre, la pondération de base et la post-stratification intégrée feront en sorte qu'il sera hautement probable de retrouver des poids finaux différents d'un ménage à l'autre, mais identiques pour les individus d'un même ménage. Il sera donc possible de reconstituer un ménage simplement en regroupant les poids identiques. Pour éviter cette situation, on propose d'élargir le calcul du partage des poids à plus d'un ménage. On peut démontrer que cette méthode resterait sans biais à condition que le regroupement des ménages soit indépendant du processus de sélection. On peut proposer, par exemple, de regrouper les ménages à l'intérieur des strates-composantes en se basant sur le dernier chiffre du numéro de ménage; un groupe serait ainsi formé par les ménages d'une strate-composante dont le numéro se termine par 0, un autre par les ménages dont le numéro se termine par 1, etc. Rappelons que les strates-composantes forment une division de la population en rapport avec l'Enquête sur la population active.

4.1.3 Variables géographiques

Afin de mieux contrôler la divulgation, on étudie la possibilité d'éliminer du fichier de sortie les variables géographiques. Cette mesure pourrait aller jusqu'à enlever la province de résidence. L'idée de base ici est que l'EDTR est une enquête pan-canadienne ne visant pas au départ la production de statistiques régionales ou

provinciales. De manière à tenir compte des disparités géographiques, on envisage ainsi d'utiliser des variables qualitatives dérivées de l'aspect géographique. On pense, entre autres, à l'utilisation d'une variable "milieu rural/urbain", d'une indication de la pauvreté relative de la région, de la taille de la ville de résidence, d'une variable "famille sous le seuil de la pauvreté", etc. Notons que les variables linguistiques peuvent servir à cerner de façon approximative les régions à haute densité de francophones.

Une alternative envisagée pour conserver une variable géographique au niveau provincial est l'utilisation de la commutation des données ("data swapping"). Avec cette méthode, on commute à l'intérieur de chaque province les données des individus de manière à ce que les totaux régionaux soient respectés au niveau, par exemple, des effectifs démographiques. On note cependant deux problèmes relatifs à cette méthode: Premièrement, l'utilisateur ne pourrait pas se servir de la variable "région", par exemple, pour ses études longitudinales puisque les répondants n'auraient plus nécessairement leur vraie région. La commutation des données n'est en fait intéressante que du point de vue transversal qui est, rappelons-le, de seconde priorité pour l'EDTR. Il serait de plus facile pour l'utilisateur de se servir négligemment de la variable "région" et ainsi tirer des conclusions erronées. Le deuxième problème est que la commutation des données est difficile à réaliser en pratique. Le problème n'est pas de commuter les régions elles-mêmes mais plutôt de s'assurer que les effectifs démographiques soient respectés. Ces deux problèmes font en sorte que la commutation de la variable "région" n'est pas très attrayante à nos yeux.

4.1.4 Vecteurs de transitions

L'EDTR mesure un certain nombre de variables dynamiques, entre autres, sur le travail, l'activité scolaire, le chômage et les épisodes de bien-être social. Ces variables apparaissent sous forme de vecteurs de transitions au sein du fichier de sortie. Les vecteurs de transitions ne seront pas considérés comme des données confidentielles. On base cette décision sur le fait qu'il peut être très difficile, voir même impossible, de relier en pratique ces vecteurs à un individu en particuliers. Donc à moins que des vecteurs de transitions ne soient tout à fait aberrants ou particuliers, ils ne feront pas l'objet d'un contrôle de la divulgation. Les vecteurs de transitions seront en fait classés comme variables sensibles et non pas comme identifiants indirects. Notons que cette décision réduit considérablement la difficulté du contrôle de la divulgation en diminuant de façon substantielle le nombre d'identifiants indirects.

4.1.5 Champs contenant du texte

La plupart des champs contenant du texte seront éliminés du fichier de sortie, surtout si l'information contenue dans le texte augmente le risque de divulgation. On retrouve parmi ces variables les notes de l'interviewer, le nom de l'employeur, le genre d'entreprise, d'industrie ou de service, la description du genre de travail effectué, les principales activités ou tâches effectuées au travail et le principal domaine d'études. Ces renseignements seront pour la plupart codifiés et seuls les valeurs codifiées seront gardées au sein du fichier de sortie. Les valeurs codifiées seront, bien entendu, sujettes au processus de confidentialité de l'EDTR.

4.2 CLASSIFICATION GÉNÉRALE DES VARIABLES

4.2.1 Identifiants directs

De façon générale, seront classés comme identifiants directs:

- les identifiants de personne et de ménage provenant de l'EPA et de l'EDTR (à éliminer).
- le nom, prénom, adresse et numéros de téléphone (au travail et à la maison) du répondant (à éliminer).
- le nom, prénom, adresse, numéro de téléphone et identifiant de la personne contact (à éliminer).
- le nom de l'employeur du répondant (à éliminer).
- information provenant de la base de sondage (unité primaire d'échantillonnage (UPE), groupe de rotation, strate, composante, numéro de tâche, numéro d'interviewer (à anonymiser).

4.2.2 Identifiants indirects

De façon générale, seront classés comme identifiants indirects:

- les variables géographiques (à éliminer, sauf peut-être la province de résidence).
- la date de naissance, l'âge, le sexe, la langue de préférence, la langue de l'entrevue et l'état matrimonial du répondant, et toutes autres variables reliées.
- la relation du répondant avec les autres membres du ménage.
- le nom de l'employeur, le type de travail et les codes d'industrie et d'occupation pour tous les emplois.

- les variables reliées au type de logement (provenant de l'EPA).
- les variables reliées à l'origine ethnique, au pays d'origine et à la langue maternelle.
- le nombre d'enfants.
- les variables reliées aux études: principal domaine d'étude, obtention d'un diplôme en médecine ou en optométrie, obtention d'un doctorat, etc...
- variables au niveau du ménage dérivées à partir d'identifiants indirects.
- items à cocher par l'interviewer reliés à des identifiants indirects comme, par exemple, cocher si le répondant est âgé de plus de 70 ans (à éliminer).

4.2.3 Variables sensibles

Seront classées comme variables sensibles toutes les autres variables, c'est-à-dire les variables ayant trait à la vie privée des gens. Ceci inclut les variables non mentionnées ci-haut et reliées au travail et au revenu, ainsi que certaines variables au niveau des ménages non reliées à des identifiants indirects (par exemple le revenu familial).

5. BIBLIOGRAPHIE

Butlin, G. (1994), *Variables de l'EDTR relatives aux ménages et aux familles*, publication n° 94-06, Série de documents de recherche de l'EDTR, 1994.

Idehem, I., Genest, C. (1994), *L'utilisation de données regroupées et son impact sur l'inférence en analyse de la variance*, Actes du colloque sur les méthodes et applications de la statistique 1994 (à paraître), Bureau de la statistique du Québec, Québec, 1994.

Lavallée, P. (1994), *Confidentialité des microdonnées de l'EDTR: Approche générale*, document interne, 11 mars 1994.

Lebrasseur, D. (1994), *Différents aspects de la méthodologie dans la production des fichiers de microdonnées à grande diffusion dans le cadre du recensement de 1991*, Actes du colloque sur les méthodes et applications de la statistique 1994 (à paraître), Bureau de la statistique du Québec, Québec, 1994.

Schackis, D. (1993), *Manual On Disclosure Control Methods*, Eurostat - D3, Luxembourg, septembre 1993.