Techniques d'enquête, décembre 1980
Vol. 6, nº 2
Statistique Canada, nº 12-001-X au catalogue

Autres méthodes de collecte de données socio-économiques dans les années 1980 : le concept de base de données synthétique

Mukund Nargundkar et Anis Ashraf[1]

Résumé

Le présent document propose la création d'une banque de renseignements d'un nouveau genre, la « banque de données synthétiques ». Il s'agirait de coupler les renseignements de deux banques distinctes pour en créer une troisième. Il en résulterait une utilisation beaucoup plus grande des banques de données existantes dans le cadre des activités de collecte de données nouvelles. On pourrait ainsi réduire considérablement la quantité de données recueillies et, par conséquent, les coûts de collecte et le fardeau de réponse. Ce document recommande diverses considérations pour l'élaboration de techniques statistiques susceptibles de faciliter la création d'un tel concept de couplage de l'information. Certaines techniques pourraient se retrouver dans les ouvrages modernes, alors que d'autres devraient être élaborées.

Mots-clés : Banque de données synthétique; couplage de l'information; collecte des données.

1 Introduction

Depuis la nuit des temps, la collecte de données ou de renseignements joue un rôle important dans les activités humaines et les processus décisionnels. Autrefois, les seigneurs féodaux utilisaient ces renseignements pour établir les besoins en main-d'œuvre pour la protection et la préservation de leurs domaines et de leurs sources de revenus, de même que pour desservir leur système de distribution en période d'abondance ou de famine.

En ces temps modernes, le principe fondamental de la collecte de données demeure inchangé. Aujourd'hui, les données sont recueillies par tous les êtres humains en leur qualité de fonctionnaires, d'entrepreneurs privés et d'universitaires, et à l'échelle internationale par les Nations Unies. Les activités modernes de collecte de données, cependant, pourraient être classées comme suit :

a) Données administratives (recouvrement des recettes, application de la loi, etc.)

1. Mukund Nargundkar et Anis Ashraf, Statistique Canada. Les points de vue exprimés dans ce document sont uniquement ceux de l'auteur et ne représentent pas la politique organisationnelle.

b) Données socioéconomiques (recensement de la population, logement, agriculture, industrie manufacturière, balance des paiements, etc.)
c) Données sur les sciences physiques (sols, minéraux, ressources halieutiques, etc.)

Bien qu'il soit reconnu que toutes les composantes des données susmentionnées sont essentielles à l'amélioration de l'humanité d'une façon ou d'une autre, aux fins du présent document, nous souhaitons nous concentrer principalement sur les données administratives et socioéconomiques.

Tout d'abord, une question se pose : pourquoi est-il impératif de recueillir des données socioéconomiques? La réponse réside dans les besoins des différents secteurs de notre économie. Dans le secteur public, ce type de données sert à la formulation de politiques sur les plans exécutif, judiciaire et législatif. En l'absence de données fiables sur des sujets pertinents, il est impossible de légiférer ou de formuler des politiques valables pour le bien de la société dans son ensemble. Les stratégies de production, d'allocation des ressources, de marketing et de publicité dans le secteur privé seraient fortement limitées si des renseignements fiables sur les variables socioéconomiques n'étaient pas aisément accessibles. La recherche fondamentale menée par le milieu universitaire repose également sur des données similaires.

On peut se demander quelles sont les sources de renseignements socioéconomiques. Les sources pourraient être innombrables, mais en fin de compte, la seule source fondamentale reste l'humain, à l'instar de ses ancêtres dans les temps anciens. C'est lui qui, ultimement, soutient le poids de tous les organismes de collecte de données, en tant que chef de ménage, entrepreneur, consommateur et même en tant que mari payant une pension alimentaire. Cependant, il est primordial de recueillir ces renseignements auprès des personnes, car aucune société moderne ne peut continuer à progresser dans les domaines socioéconomique et administratif sans information fiable. En revanche, parallèlement, la société s'engage pleinement à ne pas empiéter indûment sur les questions perçues comme privées ou sensibles par une personne. Tel est donc le dilemme auquel sont confrontés les organismes de collecte de données et qui touche la liberté des droits et obligations individuels.

De nos jours, les statisticiens ont mis au point des méthodes scientifiques de collecte de données qui, dans une grande mesure, répondent aux exigences des utilisateurs de données en matière : a) d'adéquation, b) d'actualité, c) de mesure de l'erreur (qualité); et d) des gains d'efficacité liés aux coûts. Jusqu'à récemment, tout cela dépendait d'un dénombrement intégral de la population étudiée. Par conséquent, la collecte, la compilation, la totalisation, l'analyse, la diffusion et le stockage des données ont mobilisé énormément de ressources. En d'autres termes, bon nombre d'employés étaient continuellement ensevelis sous une tonne de données statistiques et d'imprimés d'ordinateur. Comme vous pouvez bien vous l'imaginer, c'était un cauchemar sur le plan de la gestion que de produire des données fiables pour les gouvernements et les industries afin de les aider à formuler leurs politiques et à élaborer leurs stratégies dans un délai donné. Cela a encore une fois exercé une pression énorme sur la personne qui, aux termes de la loi, était tenue de fournir continuellement des données sur pratiquement tous les aspects de sa vie. Les législateurs, d'une part, avaient besoin de ces renseignements et, d'autre part, se sentaient contraints de veiller à l'intérêt et au bien-être de leur électorat (électeurs). La législation a obligé les collecteurs de données à mettre en œuvre les programmes de collecte de renseignements des planificateurs de politiques, qui acquiesçaient à la demande de l'électorat. Ainsi, la situation est devenue un dilemme pour tous et un cercle vicieux sans solution en vue.

Face au « harcèlement mutuel » entre les principaux segments de la société, les statisticiens de certains pays ont été contraints de trouver une solution leur permettant de réaliser leurs objectifs sans passer par l'ensemble du processus et de réduire au minimum le fardeau de réponse, tout en optimisant (ou, dans le pire des cas, en maintenant) l'efficacité, la fiabilité et l'actualité des produits statistiques pour les utilisateurs.

Les pionniers dans ce domaine, comme le professeur Mahalanolis et sir Ronald Fisher, ont avancé la théorie selon laquelle une enquête par échantillonnage probabiliste pouvait fournir des résultats satisfaisants et fiables en exploitant des ressources considérablement limitées et dans un délai plus court. C'était donc l'aube d'une nouvelle ère dans le domaine de la collecte de données fiables.

Nous devons maintenant nous pencher sur l'élaboration de la théorie de l'échantillonnage et son application dans le contexte de la société industrielle moderne, de même que sur les réponses qu'elle apporte aux questions soulevées précédemment. Les statistiques estimées à partir d'une enquête par sondage sont sujettes à deux types d'erreurs : a) l'erreur d'échantillonnage; et b) l'erreur non due à l'échantillonnage, généralement exprimée en tant qu'erreur quadratique moyenne (EQM) des statistiques estimées, plus connue comme la mesure de la qualité des données. Penchons-nous d'abord sur l'erreur d'échantillonnage. Depuis l'ébauche de la théorie de l'échantillonnage, les statisticiens d'enquêtes ont presque exclusivement consacré leur énergie à réduire au minimum les erreurs d'échantillonnage en vue d'accroître la précision, l'efficacité et la fiabilité des statistiques. De plus, on a mis au point de nombreuses méthodes de sélection d'échantillons probabilistes, de même que des « estimateurs » pour estimer les caractéristiques et leur erreur d'échantillonnage (variance). En revanche, l'erreur non due à l'échantillonnage est demeurée une entité insaisissable et inquantifiable, en ce qui concerne la mesure directe. L'élaboration théorique de l'estimation de l'erreur non due à l'échantillonnage est, à juste titre, de nature empirique et n'en est encore qu'à ses débuts. Des techniques comme le réseau d'échantillons, les plans expérimentaux ou d'autres méthodes statistiques ont connu un succès mitigé. Néanmoins, les statisticiens d'enquêtes du Canada et d'ailleurs s'emploient toujours à perfectionner ce domaine. Leurs problèmes et frustrations sont centrés sur les facteurs contribuant à l'erreur non due à l'échantillonnage, à savoir : a) l'inaccessibilité des répondants, le refus de fournir toute information ou la fourniture de renseignements incomplets; et b) les erreurs commises lors de l'enregistrement, de la transcription et du traitement des renseignements.

Compte tenu de ces facteurs, nous estimons que les principes de base des méthodes de collecte de données doivent être réévalués sur le plan de l'efficacité et de la fiabilité pour estimer l'erreur totale d'une statistique, c'est-à-dire les erreurs d'échantillonnage et les erreurs non dues à l'échantillonnage. Il est concevable, dans un certain nombre de situations, que le nombre d'erreurs non dues à l'échantillonnage puisse en fait dépasser le nombre d'erreurs d'échantillonnage. Dans tous ces cas, les inférences ou les décisions basées uniquement sur l'erreur

d'échantillonnage pourraient fort bien être trompeuses. Bien que ce phénomène soit à l'étude, nous ne prévoyons pas de changement radical dans l'attitude d'un répondant qui consacrerait son énergie à satisfaire les exigences insatiables des organismes de collecte de données dans un avenir prévisible, c'est-à-dire dans les années 1980. Des techniques comme l'autodénombrement, le recours à des incitatifs, les interviews téléphoniques et par la poste n'ont, jusqu'à présent, produit que des gains d'efficacité marginaux. Si les pratiques actuelles de collecte de données se poursuivent moyennant seulement des modifications disparates pour surmonter les difficultés actuelles et répondre aux changements prévus au chapitre des attitudes des répondants à l'égard de la collecte de données, non seulement le coût de la collecte de données grimpera en flèche, mais la qualité des données se détériorera à un point tel qu'elles seront peu utiles au processus décisionnel.

Pour les statisticiens des années 1980, il est donc essentiel de réduire au minimum le fardeau de réponse grâce à : a) d'autres méthodes de collecte de données; et b) l'optimisation de la valeur analytique des bases de données existantes sans compromettre les principes d'opportunité, de qualité, de pertinence et de rentabilité des statistiques nécessaires.

Les solutions de rechange proposées

À notre avis, la première option pour produire des statistiques de haute qualité sans retourner aux méthodes traditionnelles (interview sur place, entre autres) serait d'obtenir des statistiques pertinentes sur les variables socioéconomiques à partir de bases de données administratives existantes. Comme nous l'avons souligné précédemment, il existe une quantité considérable de renseignements sur une gamme de sujets, sous une forme ou une autre, dans diverses banques de données gouvernementales et industrielles centralisées. En outre, la qualité des données administratives, puisqu'elles sont fournies sur une base obligatoire plutôt que volontaire, est naturellement supérieure.

Afin de dériver des statistiques socioéconomiques à partir de banques de données administratives, il faudrait tout d'abord élaborer certaines procédures. On peut généralement les classer dans quatre catégories :

- Élaboration de méthodes afin de veiller à ce que les statistiques dérivées soient appropriées (pertinentes), à jour (période de référence appropriée) et complètes du point de vue de la couverture et du contenu.
- Élaboration d'un grand éventail de techniques de coordination afin de relier deux bases de données indépendantes.
- Synthèse de deux ou de plusieurs bases de données, administratives ou socioéconomiques, en vue de créer une nouvelle base de données pour un ensemble donné de variables socioéconomiques.
- Élaboration de banques de données statistiques « à charte ».

D'une manière générale, toutes les enquêtes socioéconomiques nécessitent, sous une forme ou une autre, des variables comme l'âge, l'état matrimonial, le revenu et des caractéristiques temporelles similaires. Ces faits peuvent sans doute être obtenus à partir de banques de données administratives sans nécessairement les intégrer dans un questionnaire d'enquête. En outre, les caractéristiques fixes dans le temps, comme la langue maternelle, le lieu de naissance, le groupe sanguin, peuvent également être obtenues à partir de sources similaires sans effort supplémentaire.

Pour approfondir les points fondamentaux soulevés précédemment, il semble logique de mentionner ici des exemples scientifiques pour chacune des catégories.

Voyons donc maintenant la façon dont nous pourrions adapter une base de données administratives à une étude socioéconomique donnée.

Par exemple, supposons que nous souhaitons convertir une base de données administrative générée à partir de déclarations de revenus pour créer une répartition des revenus des familles de recensement. Dans ce cas, il y aurait possibilité de réorganiser l'ensemble de données pour créer des « familles de recensement » à partir des fichiers de données en regroupant les personnes (y compris les personnes à charge) ayant les mêmes noms de famille et les mêmes adresses, pour ainsi créer un ensemble de données approprié permettant de générer la répartition des revenus familiaux. Dans une situation où l'ensemble des données ne peut pas être réorganisé de façon à créer des « familles », un modèle statistique dans lequel des renseignements auxiliaires ou externes sont accessibles

pourrait faire correspondre synthétiquement les personnes de façon à créer des « familles ». Le modèle utilisé pour créer la base de données appropriée à des fins statistiques dépend du contenu de la base de données et des exigences statistiques pour lesquelles elle doit être utilisée. Il s'agirait donc de convertir la base de données afin qu'elle soit adaptée aux statistiques sur le revenu des familles. La base de données ainsi créée est une base de données synthétique, dans le sens où elle est générée à partir d'une base de données administratives de base selon une procédure de simulation statistique et non à partir d'une procédure de collecte de données « réelle ». Il convient également de souligner que la base de données synthétique représente des renseignements sur la population à un certain niveau d'agrégation et non à un niveau micro ou individuel.

Une autre possibilité pour adapter les données administratives à un objectif statistique consiste à modifier les objectifs de l'analyse statistique de manière à ce que l'ensemble de données disponibles puisse devenir approprié. Par exemple, si les objectifs sont modifiés de façon à permettre le recours à une répartition des revenus des particuliers au lieu des revenus familiaux, l'ensemble de données générées par les déclarations de revenus deviendrait approprié.

Le recours aux données administratives dépend également de la compatibilité de la période à laquelle les données se réfèrent et de la période de référence pour laquelle les statistiques doivent être dérivées de l'ensemble de données. Par exemple, l'ensemble de données généré à partir des déclarations de revenus concerne les revenus de l'année civile précédente, tandis que les statistiques à dériver de l'ensemble de données se rapportent à l'exercice en cours. Cependant, dans la collecte de données « réelle », les renseignements recherchés sur le revenu se rapportent souvent à la dernière année d'imposition complète ou sont présentés par rapport à celle-ci.

Dans de telles situations, l'ensemble de données peut être mis à jour en utilisant des techniques de projection ou de prévision en vue de créer une base de données synthétique ayant une période de référence compatible avec les objectifs. Là encore, il existe de nombreuses procédures de projection et de prévision parmi lesquelles il est possible de choisir celle qui convient. L'adéquation d'une procédure particulière serait

déterminée en fonction de la facilité de calcul et de contrôle de la qualité des statistiques et de la disponibilité des renseignements « externes » nécessaires. Il convient de mentionner qu'une projection progressive ou régressive dans le temps peut être réalisée selon les besoins.

Le recours à des données administratives dépend également de la compatibilité de la population cible d'une étude statistique avec la population visée par un ensemble de données administratives. Par exemple, un ensemble de données administratives généré à partir des renseignements sur les allocations familiales ne porterait pas sur les familles au Canada qui n'ont pas de personnes à charge admissibles aux allocations familiales. Par conséquent, cet ensemble de données administratives peut ne pas être utile à des fins statistiques lorsque la population cible est constituée de toutes les familles de recensement du Canada. Dans de telles situations, la série de données administratives doit être complétée ou les unités manquantes doivent être générées au moyen de renseignements auxiliaires.

De même, dans les situations où un ensemble de données administratives compte la plupart des variables requises pour une étude mais où certaines variables font défaut, il faut introduire d'autres bases de données, renseignements auxiliaires ou techniques statistiques. Les renseignements sur les variables manquantes peuvent être générés à l'aide de modèles de régression ou d'autres méthodes d'information. Les données supplémentaires destinées à combler les lacunes au chapitre de la couverture ou du contenu d'une base de données particulière en générant une base de données synthétique peuvent se situer à un niveau micro ou macro. À plus grande échelle, cela nous conduit à la deuxième phase de notre exposé.

Le problème du couplage des ensembles de données était, par le passé, lié au couplage au niveau micro. C'est ce que l'on appelle plus communément le problème du deuxième couplage. Cependant, le couplage pris en compte pour générer une base de données synthétique ne doit pas se limiter au niveau micro; il peut être étendu au « couplage statistique ». L'extension du couplage statistique permet d'accroître l'utilisation des ensembles de données.

Des travaux approfondis ont déjà été réalisés dans le domaine du couplage des microenregistrements (données) et ne seront pas traités davantage dans le présent document. Certains travaux ont également été réalisés en ce qui a trait aux liens entre les données, que l'on peut appeler le « couplage statistique ». Les avancées dans les domaines de la simulation de données, de l'analyse de modèles, de la régression multidimensionnelle et d'autres techniques statistiques peuvent être mises à profit pour réaliser le couplage statistique.

Le couplage statistique est une technique statistique qui peut être utilisée efficacement pour générer un ensemble de données synthétiques en « combinant » deux ou plusieurs ensembles de données tout en fournissant des mesures de la qualité pour les statistiques dérivées de l'ensemble de données synthétique.

Compte tenu de cette définition générale d'un système de couplage statistique, on pourrait affirmer que les possibilités de couplage statistique sont illimitées. Il n'est pas possible, dans le cadre du présent document, d'examiner en détail les façons d'adopter les techniques statistiques qui peuvent servir à générer un ensemble de données synthétique. Nous en donnerons toutefois un aperçu à l'aide de quelques exemples.

Il convient de mentionner que les techniques statistiques actuellement disponibles ne sont pas toutes conçues pour le couplage statistique en soi; elles doivent être adaptées. Il faut mettre au point d'autres nouvelles techniques statistiques propres à l'utilisation du couplage de données. C'est le nouveau domaine de développement de l'information pour la prise de décisions que les statisticiens d'enquêtes doivent reconnaître et mettre au point. Sinon, les sociologues, les psychologues, les responsables d'études de marché et les autres décideurs, etc., qui ne sont pas statisticiens mais qui sont de gros utilisateurs de données, devront mettre au point leurs propres méthodes pour générer l'information, comme ils l'ont fait pour concevoir et adopter des techniques statistiques dans des domaines comme la psychométrie, l'analyse conjointe, etc.

Il nous faut bien préciser une fois de plus que la base de données synthétique générée représenterait la population cible au niveau agrégé et ne serait utile que pour la dérivation d'inférences statistiques. Nous

réitérons également à ce stade qu'une unité dans un ensemble de données synthétique n'existe pas à titre individuel, mais représente la population au niveau agrégé. De même, les liens statistiques dont nous traitons ne se limitent pas aux ensembles de données administratives, mais s'étendent aussi à des ensembles de données statistiques et à d'autres types d'ensembles de données. Par conséquent, lorsque nous faisons référence à un ensemble de données, il peut s'agir d'un ensemble de données administratives, statistiques ou autres. Les seules conditions auxquelles un ensemble de données doit satisfaire pour faire partie d'un couplage statistique sont les suivantes :

i) Chaque unité (enregistrement) est définie de manière unique par un ensemble de renseignements.

ii) La « population cible » de l'ensemble de données est clairement définie.

iii) La forme et le format dans lesquels l'ensemble de données est disponible peuvent facilement être traités par ordinateur.

La troisième phase de cet exercice consiste à synthétiser plus de deux ensembles de données.

Nous allons maintenant examiner quelques exemples afin de classer la génération d'une base de données synthétique.

Supposons qu'il existe deux ensembles de données sur une population. Un ensemble de données, D1, présente les caractéristiques démographiques de base des unités (personnes) ainsi qu'un ensemble de caractéristiques C1, et l'autre ensemble de données D2 présente également certaines caractéristiques démographiques de base des unités (personnes) et un autre ensemble de caractéristiques C2. Supposons également que l'objectif d'une analyse statistique exige que les renseignements des ensembles de caractéristiques C1 et C2 soient tous deux disponibles sur un ensemble de données, en plus des caractéristiques géographiques de base.

Si toutes les unités de D1 recoupent D2 et qu'il existe un identifiant commun pour relier deux ensembles au niveau micro, et si les ensembles de caractéristiques C1 et C2 se rapportent à la même période, l'ensemble de données requis avec les deux ensembles de caractéristiques C1 et C2 peut être obtenu sans aucun couplage statistique. L'une des autres

possibilités peut être de combiner une ou plusieurs des situations suivantes :

i) Les périodes de référence de C1 et C2 ne sont pas les mêmes.

ii) Ce ne sont pas toutes les unités de D1 et D2 qui sont communes; donc, ni D1 ni D2 ne rendent complètement compte de la population cible, P. Toutefois, P = D1UD2.

iii) Il n'existe pas d'identificateur d'enregistrement commun permettant de relier les deux ensembles de données à l'aide d'une procédure de couplage des microdonnées.

On peut d'abord déterminer les caractéristiques des ensembles C1 et C2 qui sont fixes dans le temps (la langue maternelle, le lieu de naissance, etc.). Celles qui varient dans le temps peuvent être ajustées en fonction de la période de référence requise au moyen de diverses méthodes d'estimation ou de projection. Pour réaliser l'estimation ou la projection, il est possible d'avoir recours à des renseignements auxiliaires. Par exemple, supposons que l'un des éléments de C1 est une donnée de revenu annuel qui remonte à un an alors que le revenu actuel est requis. La procédure la plus simple (mais rudimentaire) pour actualiser le revenu serait d'ajuster l'ancien revenu au moyen d'un facteur dérivé de renseignements auxiliaires comme l'augmentation du produit intérieur brut (PIB), l'Indice des prix à la consommation (IPC), etc. On peut également avoir recours à un ensemble de courbes de régression basées sur des données antérieures pour projeter le revenu sur la période actuelle. Les renseignements auxiliaires utilisés et la procédure adoptée pour effectuer la projection dépendraient certainement des objectifs de l'étude statistique. De même, les renseignements démographiques de base peuvent être mis à jour pour les deux ensembles de données.

L'ensemble de données D1 ou D2 ne rend pas compte de la population cible P, mais comme nous supposons que P = D1UD2, cela signifie que, collectivement, ils couvrent la population cible. Par conséquent, avant de procéder à un couplage statistique des ensembles de caractéristiques C1 et C2, nous devons ajuster les ensembles de données afin que chaque ensemble couvre la population cible.

Pour ce faire, nous pouvons générer f_1 et f_2, deux ensembles d'unités fictives à ajouter à D1 et à D2 de manière à rendre compte de la population totale où :

$$f_1 = D2 - D1 \cap D2$$
$$f_2 = D1 - D1 \cap D2$$

Pour générer f_1 et f_2, il faut déterminer les unités communes entre D1 et D2. Cela peut se faire, par exemple, en comparant la répartition selon l'âge et le sexe des unités des ensembles de données D1 et D2. Il convient de souligner que la détermination des unités communes peut ne pas être basée sur la correspondance exacte ou la valeur selon l'âge et le sexe, mais sur une règle statistique comme les règles utilisées pour cerner les valeurs aberrantes ou non aberrantes dans un ensemble de données.

Après avoir ajusté les ensembles pour rendre compte de la population cible totale, une technique d'imputation par enregistrement donneur peut être employée pour imputer les caractéristiques des unités fictives dans chaque ensemble. Pour réaliser le couplage statistique des données, il est possible d'employer la technique d'imputation par enregistrement donneur (hot deck) pour imputer les caractéristiques « manquantes » d'un ensemble à l'aide de l'autre ensemble en tant qu'enregistrement donneur et en utilisant les caractéristiques démographiques de base comme variables corrélées, et en maintenant les répartitions des caractéristiques imputées identiques à celles des ensembles de données originaux. En utilisant cette procédure, nous pourrions obtenir deux ensembles de données synthétiques. Ces deux ensembles de données synthétiques représentent la même population et peuvent donc servir à évaluer la qualité des statistiques dérivées de ces ensembles.

Dans notre exemple, nous supposons qu'il existe un ensemble commun de caractéristiques démographiques sur les deux ensembles, mais il n'est pas nécessaire que ce soit ainsi. Toutefois, en l'absence d'un ensemble commun de caractéristiques, il faut avoir recours à des renseignements auxiliaires externes, par exemple des paramètres de corrélation entre C1 et C2, si corrélation il y a.

De même, on a supposé que D1UD2 = P. Cependant, l'ensemble D1 ou D2 pourrait être une base de données-échantillon. Dans cette situation, l'ensemble de données synthétique dérivé pourrait également être une base de données-échantillon; il faudrait donc générer des « pondérations » aux fins d'estimation.

Un couplage statistique similaire peut s'étendre à plus de deux ensembles de données pour générer une base de données synthétique.

La création de banques de données statistiques « à charte » est le quatrième et dernier processus de l'exposé sur les autres sources d'information.

Les banques de données dont il est question ici ne sont pas seulement les banques de données telles qu'on les conçoit actuellement, mais des banques de données au sens propre, comme les banques à charte, où divers organismes de collecte de données (c'est-à-dire le gouvernement, l'industrie, le milieu universitaire) déposeraient les données. Ensuite, les banques créées monteraient les bases de données synthétiques pour les utilisateurs de données à des fins statistiques.

Il ne s'agit pas d'une proposition farfelue, puisque de telles banques de données existent actuellement au sein du gouvernement fédéral, par exemple, le Système canadien d'information socioéconomique (CANSIM) à Statistique Canada. Dans le secteur privé, le Bureau d'éthique commerciale est une banque de données de ce type. Les banques d'information de l'avenir pourraient être des banques de données coopératives dans lesquelles les membres déposeraient des données afin de générer des bases de données synthétiques plus vastes en exploitant les ensembles de données de divers membres qui pourraient être retirés par les membres aux fins d'utilisation dans leur processus décisionnel. Depuis l'avènement de technologies comme Télidon, une telle réalité est potentiellement imminente.

Conclusion

Il ressort ce qui suit des observations ci-dessus :

a) Il est possible d'obtenir des données socioéconomiques à partir de données administratives déjà accessibles auprès des organismes gouvernementaux.
b) Le couplage des bases de données peut être élaboré sans trop d'efforts, bien qu'il soit nécessaire de mettre au point des procédures hautement raffinées et sophistiquées. Par ailleurs, il faudrait donner de la formation et acquérir des compétences techniques supplémentaires dans ce nouveau domaine.
c) La création de banques de données statistiques « à charte » nécessite des recherches supplémentaires pour les procédures de mise en œuvre. Pour que celles-ci fonctionnent efficacement, il est nécessaire d'obtenir la pleine collaboration de toutes les parties concernées, qu'il s'agisse des administrateurs de données administratives ou de données socioéconomiques.

Nous estimons, sur la base de la présentation de notre analyse de ce domaine, qu'en nous assurant que ces méthodes sont adoptées et rendues fonctionnelles, nous économiserons non seulement des millions de dollars, mais aussi des centaines d'heures de travail. Qui plus est, nous protégerons la source d'information — dans ce cas, le répondant lui-même — de la pression continue exercée par les organismes de collecte de données.

C'est maintenant à nous, les statisticiens, d'apaiser la crainte de répondre des répondants et de nous éviter, à nous les organismes de collecte de données, les erreurs mathématiques. Peut-être que l'application des solutions de rechange suggérées ci-dessus pourrait contribuer à la résolution du problème.

Remerciements

Les auteurs souhaitent remercier monsieur S.M. Ashraf pour sa précieuse contribution à l'élaboration du présent document.