Anonymisation des données à Statistique Canada cadre et aperçu

Michelle Simard
Directrice adjointe,
Division des méthodes d'enquêtes auprès des ménages

11 juillet 2016
CEFIL Libourne, France

Canada

Greenland
Iceland
Suomi
Finland
Sverige
Sweden
Norge
Norway
Россия
Russia
United
Kingdom
Polska
Poland
Deutschland
Germany
Україна
Ukraine
France
Kazakhstan
Mongolia
Canada
United
States
Italia
Italy
España
Spain
Türkiye
Turkey
中国
China
대한민국
S Korea
日本
Japan
Iraq
Iran
Afghanistan
Pakistan
India
Algeria
Libya
Egypt
Saudi
Arabia
Mauritania
Mali
Niger
Chad
Sudan
Nigeria
Ethiopia
ประเทศไทย
Thailand
México
Venezuela
Colombia
Brasil
Brazil
Perú
Bolivia
Chile
DR
Congo
Kenya
Tanzania
Angola
Namibia
Botswana
South
Africa
Madagascar
Indonesia
Papua New
Guinea
Australia
Indian
Ocean
North
Atlantic
Ocean
South
Atlantic
Ocean
South
Pacific
Ocean
rth Pacific
Ocean

Canada

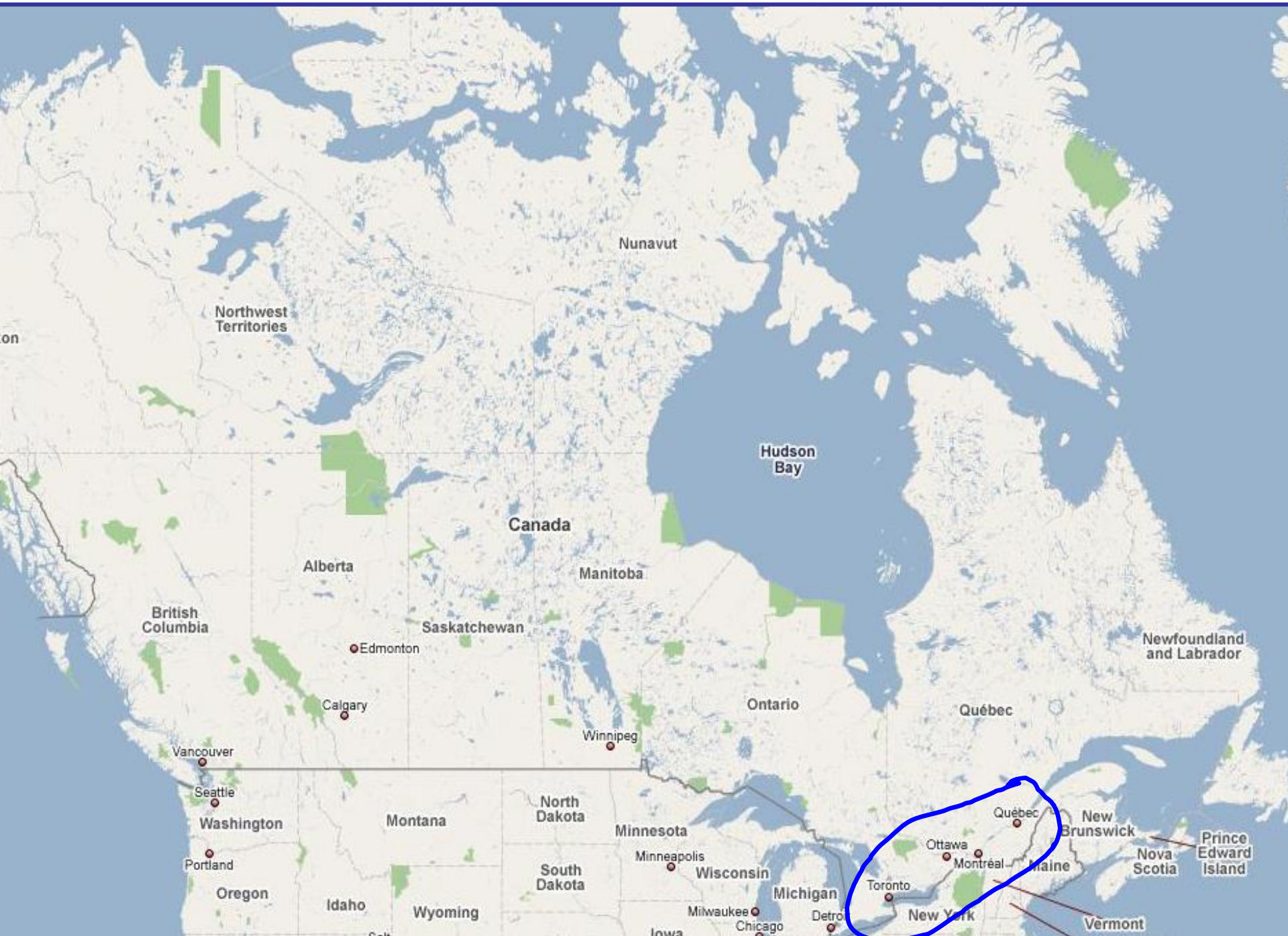
North Pole
Norwegian Sea
Norway
Russia
Greenland Sea
UK
Arctic Ocean
Ellesmere Island
Alert
Kalaallit Nunaat
(Greenland)
(Denmark)
Iceland
Bering Strait
Arctic Circle
Beaufort Sea
Banks Island
Melville Island
Prince of Wales Island
Victoria Island
Resolute (Qausuittuq)
Baffin Bay
Davis Strait
Alaska
(USA)
Inuvik
Dawson
Great Bear Lake
Yukon
Whitehorse
Northwest Territories
Yellowknife
Great Slave Lake
Nunavut
Baker Lake (Qamanittuaq)
Foxe Basin
Iqaluit
Southampton Island
Hudson Strait
Labrador Sea
Atlantic Ocean
Pacific
Fort Nelson
British Columbia
Prince George
Vancouver Island
Vancouver
Victoria
Kamloops
Fort McMurray
Lake Athabasca
Alberta
Edmonton
Calgary
Saskatchewan
Saskatoon
Regina
Churchill
Reindeer Lake
Manitoba
Lake Winnipeg
Winnipeg
Hudson Bay
Sandy Lake
Ontario
Thunder Bay
James Bay
Ungava Bay
Kuujjuaq
Ivain
Newfoundland and Labrador
Churchill Falls
Quebec
Sept-Îles
St. John's
Newfoundland
Saint-Pierre and Miquelon (France)
Gulf of St. Lawrence
Charlottetown
Prince Edward Island
Fredericton
Québec
Halifax
Nova Scotia
Montréal
New
United States of America
Canada – Political

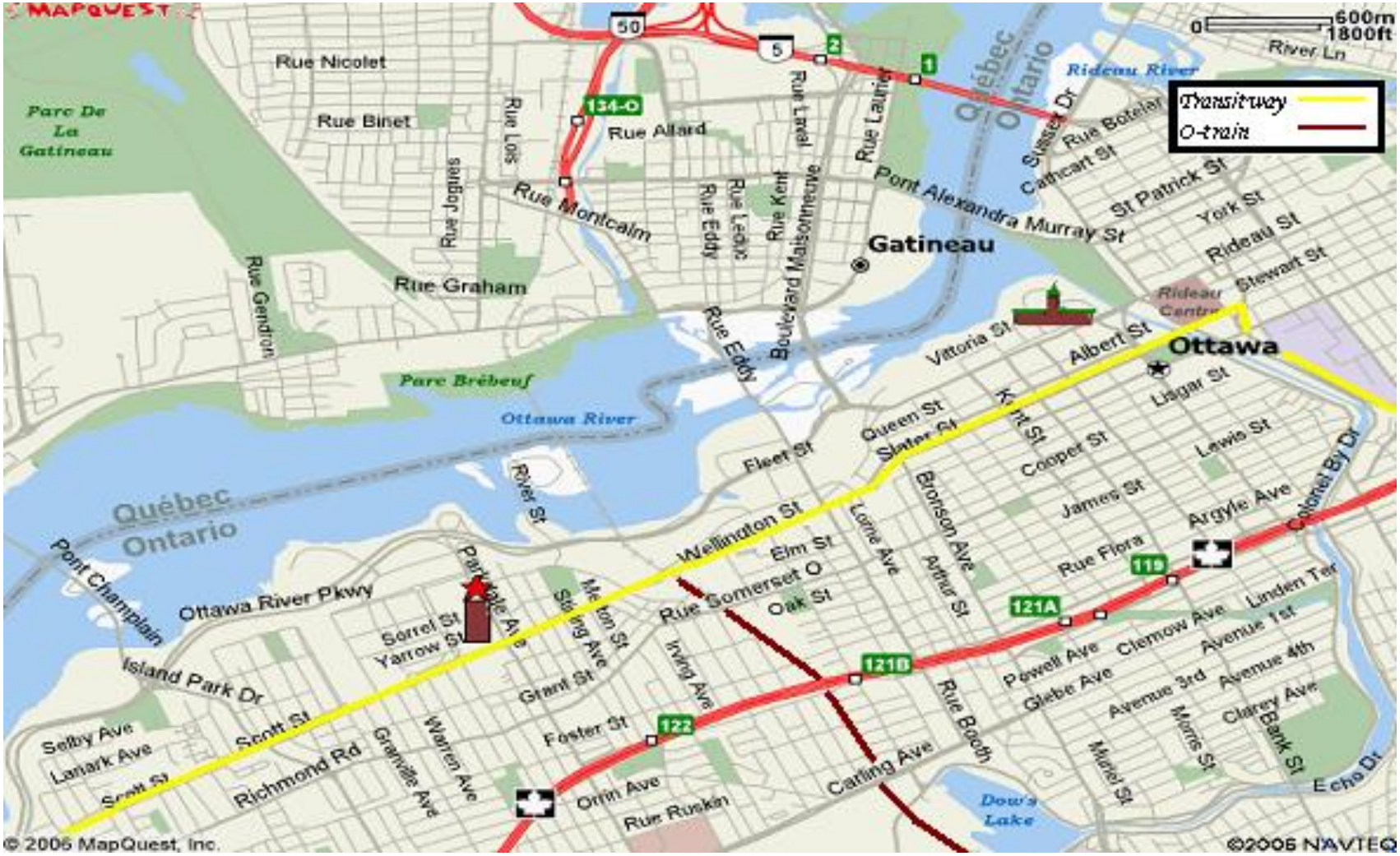
Canada

- Population : 35,2 Millions
- 10 provinces, 3 territoires
- Capitale : Ottawa
- Plus grande ville : Toronto (5.5 M)
- Langues officielles : Anglais et Français

Canada

- La partie la plus densément peuplée est le corridor Villes de Québec – Windsor, le long des Grands Lacs et du fleuve St-Laurent.


Nunavut
Northwest Territories
Hudson Bay
Canada
Alberta
Manitoba
British Columbia
Saskatchewan
Edmonton
Calgary
Winnipeg
Ontario
Québec
Newfoundland and Labrador
Vancouver
Seattle
Washington
Montana
North Dakota
Minnesota
Portland
Oregon
Idaho
Wyoming
South Dakota
Minneapolis
Wisconsin
Michigan
Milwaukee
Chicago
Detroit
Toronto
Ottawa
Montréal
Québec
New Brunswick
Maine
Nova Scotia
Prince Edward Island
New York
Vermont


MAPQUEST
0
600m
1800ft
Transitway
O-train
Rue Nicolet
Rue Binet
Parc De La Gatineau
Rue Lois
Rue Jogues
Rue Montcalm
Rue Allard
134-O
50
5
2
1
Rue Laurier
Rue Laval
Rue Kent
Boulevard Maisonneuve
Rue Eddy
Rue Leamy
Rue Graham
Rue Gendron
Gatineau
Québec
Ontario
Rideau River
River Ln
Sussex Dr
Rue Boteler
Cathcart St
Pont Alexandra
Murray St
St Patrick St
York St
Rideau St
Stewart St
Rideau Centre
Vittoria St
Albert St
Ottawa
Parc Brébeuf
Ottawa River
Fleet St
Queen St
Slater St
Kent St
Cooper St
Lisgar St
Lewis St
James St
Argyle Ave
Colonel By Dr
Rue Flora
119
121A
121B
122
Pont Champlain
Ottawa River Pkwy
Island Park Dr
Scott St
Wellington St
Elm St
Rue Somerset O
Oak St
Lorne Ave
Bronson Ave
Arthur St
Irving Ave
Metcalfe Ave
Sterling Ave
Grant St
Foster St
Sorrel St
Yarrow St
Parkdale Ave
River St
Selby Ave
Lanark Ave
Richmond Rd
Granville Ave
Warren Ave
Orrin Ave
Rue Ruskin
Carling Ave
Rue Booth
Powell Ave
Clemow Ave
Glebe Ave
Linden Ter
Avenue 1st
Avenue 3rd
Avenue 4th
Clarey Ave
Bank St
Echo Dr
Muriel St
Morris St
Dow's Lake
© 2006 MapQuest, Inc.
©2006 NAVTEQ

Statistique Canada

- **Au Canada, il y a :**
  - Des bureaux régionaux pour les opérations
  - Centres de données de recherché pour les analystes
  - 4,500 employés, 1,000 Économistes, 1,000 TI, 330 Mathématiciens/Méthodologistes

Introduction – Confidentialité

- La valeur d'une organisation, particulièrement celle d'une organisation publique, réside dans sa contribution à la société
- Les INS ont pour mission de recueillir de l'information à propos de la population, mais aussi de diffuser cette information
- Les utilisateurs doivent être au cœur de la stratégie de diffusion
- L'accès et la protection de la confidentialité font partie intégrante de la diffusion de l'information

Vie privée et confidentialité

- Vie privée : « le droit de vivre tranquille, d'être libre d'interférence, de surveillance et d'intrusion […] protéger la vie privée pourrait signifier ne pas recueillir de renseignements »

- Confidentialité : « relation de confiance entre la personne qui fournit des renseignements et celle qui les recueille […] assurance que ces renseignements ne sera communiqués à personne sans le consentement de l'intéressé »

- Commissaire à la protection de la vie privée du Canada

Aperçu

- Sécurité et pratiques sûres
- Confidentialité et contrôle du risque de divulgation
  - Types de divulgation
  - Caractéristiques influant sur la divulgation
- Type de données de sortie
- Méthodes de contrôle de la divulgation
- Aspects à prendre en considération pour l'établissement d'une méthode appropriée
- Protection des tableaux de données
- Protection des fichiers de microdonnées
- Protection des données de sortie analytiques
- Mot de la fin

Sécurité et pratiques sûres

Personnes sûres

Système sûr

Environnement sûr

Données sûres

Sécurité et pratiques sûres

- Personnes sûres
  - Processus d'approbation (contrat)
  - Personnes réputées être employées
  - Sensibilisation à l'importance de la confidentialité (p. ex. formation, cours, questionnaires, etc.)
  - Pratiques sûres
    - Verrouillage des postes de travail
    - Aucun accès aux données confidentielles sauf lorsque c'est nécessaire
    - Aucune transmission de données statistiques de nature délicate par courriel
    - Recours aux espaces réseaux prévus pour l'entreposage des données protégées ou confidentielles
    - Obtention d'une autorisation pour le transfert de données protégées sur des dispositifs portatifs de stockage
    - Chiffrement des données enregistrées sur des dispositifs personnels de stockage (p. ex. clés USB)

Sécurité et pratiques sûres

- Système sûr
  - Sécurité des ensembles de données entreposés (informatique)
  - Sécurité des données en transit (informatique)
  - Validation des utilisateurs enregistrés (informatique)
  - Utilisation de mots de passe
- Environnement sûr
  - Emplacement sécurisé
- Données sûres
  - Méthodes de contrôle de la divulgation

Confidentialité et contrôle du risque de divulgation

- Objectif des organisations d'information :
  - Diffuser des produits statistiques pertinents offrant des données aussi riches et complètes que possible

    -> MAIS : il faut aussi protéger le droit à la vie privée des personnes

- Une protection à 100 % est impossible – c'est le prix à payer pour diffuser des données statistiques et c'est pourquoi on parle plutôt de **contrôle du risque de divulgation**

- Il y a **divulgation** quand une personne ou une organisation reconnaît ou apprend quelque chose qu'elle ne savait pas déjà à propos d'une autre personne ou organisation à partir des données diffusées

Confidentialité et contrôle du risque de divulgation

- Statistique Canada est régi par la *Loi sur la statistique* :
  - ***Extrait : paragraphe 17(b)***

« aucune personne qui a été assermentée en vertu de l'article 6 ne peut révéler ni sciemment faire révéler, par quelque moyen que ce soit, des renseignements obtenus en vertu de la présente loi de telle manière qu'il soit possible, grâce à ces révélations, de rattacher à un particulier, à une entreprise ou à une organisation identifiables les détails obtenus dans un relevé qui les concerne exclusivement. »

Statistique Canada

- La *Loi sur la statistique* confère à StatCan les pouvoirs suivants :
  - Rendre obligatoire la participation aux enquêtes
  - Conclure des ententes sur le partage de données
  - Accéder aux données administratives

- Elle confie aussi à StatCan la responsabilité de protéger la confidentialité des données des répondants – et l'obligation de diffuser les données (à tout le monde en même temps)

Confidentialité et contrôle du risque de divulgation

- Défi

  → Tirer le maximum des renseignements tout en respectant la confidentialité des sujets (personnes)

- Les solutions représentent un compromis entre les besoins de disponibilité des données (la qualité aux analystes) et la protection de la confidentialité

Qu’est-ce que la divulgation?

- La divulgation s’entend de l’attribution inappropriée d’information à un sujet, que celui-ci soit une personne ou une organisation

- On parle de quatre types de divulgation : divulgation d’identité, divulgation d’attribut, par inférence et par recoupement

Types de divulgation

- Divulgation d'identité
  - Se produit lorsqu'une personne peut être identifiée dans des données diffusées
  - Certaines causes d'identification :
    - Personnalité ou entreprise bien connue
    - Identification par l'intéressé lui-même (p. ex. à la suite d'une plainte aux médias)
    - Tentative fructueuse de divulgation (p. ex. par couplage d'enregistrements)
    - Reconnaissance spontanée

Types de divulgation

- Divulgation d'attribut
  - Se produit lorsque des renseignements confidentiels à propos d'une personne sont divulgués
    - Divulgation d'attribut individuel
  - Ne comporte pas nécessairement l'identification d'une personne en particulier

    - « Toutes les personnes atteintes de la maladie a ont subi le traitement b. »
    - « Les personnes ayant la profession x dans l'industrie y gagnaient entre 50 000 $ et 55 000 $ l'année dernière. »
    - « 99 % des résidents de la région a bénéficient de l'aide sociale. »

Types de divulgation

- Divulgation par inférence
  - Se produit lorsque des renseignements concernant un particulier peuvent être inférés avec un haut niveau de confiance (faible niveau d'incertitude).
    - Implique une probabilité

Types de divulgation

- Divulgation par recoupement
  - Se produit lorsque des renseignements diffusés sont combinés à d'autres renseignements diffusés, ou à des renseignements provenant de sources externes, pour révéler des données confidentielles
  - **La probabilité d'une divulgation par recoupement augmente avec la quantité de données publiées par l'organisation** (p. ex. divulgation de renseignements confidentiels par recoupement)

    - Reconstitution des cellules supprimées dans les tableaux.
    - Obtention d'information concernant de petites régions en utilisant des données provenant de régions géographiques qui se chevauchent.
    - Combinaison de renseignements diffusés et de renseignements faisant partie du domaine public pour révéler des données confidentielles.

Caractéristiques influant sur la divulgation

- Types de données
  - Recensement
  - Enquête
  - Données administratives
- Types de données de sortie
  - Données tabulaires
  - Microdonnées
  - Données analytiques, graphiques, etc.
- Types de variables
  - Qualitatives ou quantitatives
  - Identificateurs directs, identificateurs indirects et variables sensibles

Couverture de la population : recensement et échantillon

- Le risque de divulgation est plus grand pour des données de type recensement
  - Aucune incertitude quand un répondant unique est identifié
  - Peu ou pas d'incertitude sous divulgation d'attribut
- L'échantillonnage réduit le risque de divulgation
  - Les répondants qui sont uniques dans un échantillon ne sont pas nécessairement uniques dans la population
  - Le plan de l'échantillon influe sur le risque de divulgation
  - L'échantillonnage peut devenir un outil en vue de réduire le risque de divulgation

Données d'enquête et données administratives

- Les données administratives présentent souvent les mêmes problèmes de divulgation que les données de recensement…
- … et les fournisseurs de données peuvent diffuser leur propre information, ce qui augmente le risque de divulgation par recoupement

- Certains organismes statistiques ont des fichiers contenant les deux types de données – que les fournisseurs de données administratives pourraient utiliser pour identifier les répondants dans les fichiers de microdonnées

Types d'unités

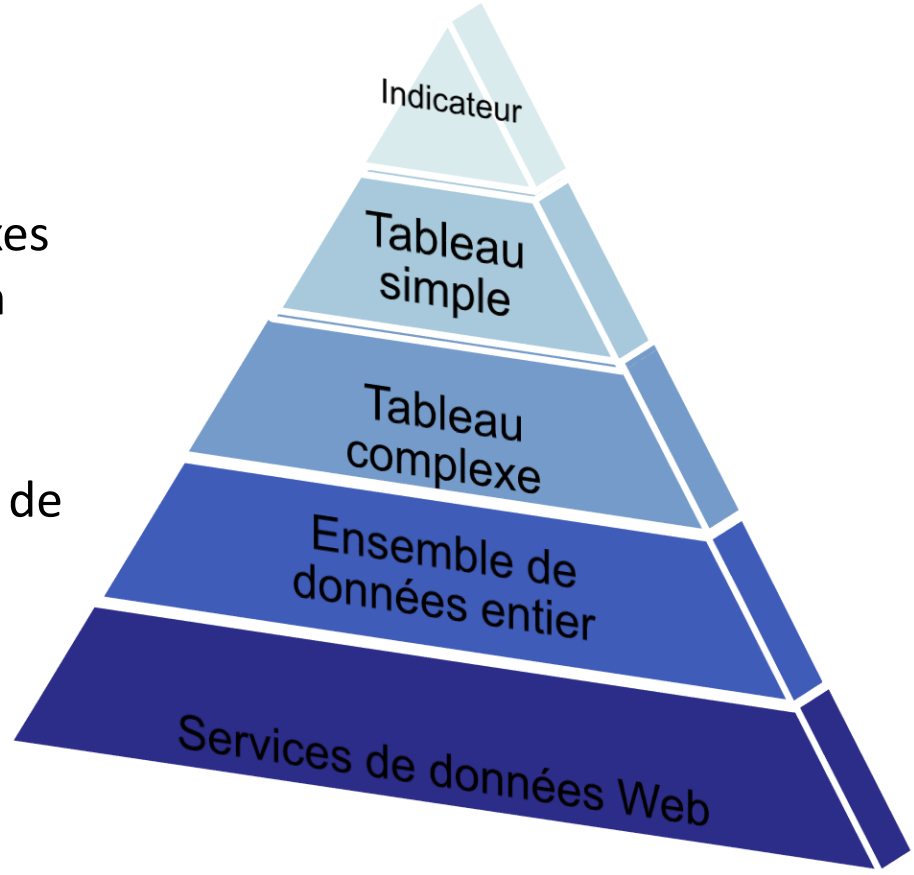
- Institutions et entités commerciales
  - Petites populations connues par domaine (industrie/région)
  - Données souvent quantitatives, asymétriques (dominance)
  - Les enquêtes comprennent souvent un recensement des unités de grande taille, importantes ou uniques
  - Incitatifs financiers ou concurrentiels à la divulgation
- Personnes et ménages
  - Nature hiérarchique des données au niveau du ménage et des particuliers
  - Les enquêtes utilisent habituellement de faibles taux d'échantillonnage mais ont un contenu de données plus « riche » (comprenant de nombreuses caractéristiques qualitatives)
  - Le concept de vie privée s'applique davantage aux données personnelles

Types de données de sortie

- Tableaux de fréquences et tableaux de données quantitatives
  - Risques réduits – surtout dans le cas des données d'enquête (échantillon)
  - Possibilité de divulgation d'identité ou d'attribut
- Fichiers de microdonnées à grande diffusion (FMGD) anonymisées
  - Les microdonnées posent le plus grand risque
  - Nécessitent un traitement important en vue d'assurer la protection contre la divulgation d'identité – répercussion sur la qualité des données
- Résultats analytiques
  - Posent habituellement le risque le plus faible
  - Les données de résultats des logiciels analytiques « standard » peuvent présenter des risques (p. ex. affichage des valeurs extrêmes)

Types de données de sortie

- Du plus simple au plus avancé
  - Indicateurs
  - Tableaux de base simples
  - Tableaux dynamiques complexes permettant la personnalisation
- Utilisateurs plus avancés
  - Téléchargement de l'ensemble de données entier
  - Services de données Web



Les risques varient selon le type de données de sortie

- Les microdonnées présentent généralement le risque le plus élevé à cause de la richesse de leur contenu

- Les données tabulaires peuvent présenter un risque relativement élevé – particulièrement les totalisations multiples

- Les données de types analytiques peuvent aussi présenter certains risques (p. ex. certains modèles saturés peuvent fournir autant d'information qu'un tableau)

Le risque

Carte risque-utilité (Duncan et coll., 2001)

- Une carte **R-U** de confidentialité est l'ensemble des valeurs appariées (R,U) du risque de divulgation et de l'utilité des données qui correspondent à diverses stratégies de diffusion des données.
- **R** est une évaluation numérique du risque de divulgation non intentionnelle après la diffusion des données.
- **U** est une évaluation numérique de l'utilité des données diffusées à des fins légitimes.
- Ces stratégies consistent habituellement à appliquer des procédures de limitation de la divulgation, p. ex., l'utilisation d'un bruit supplémentaire, où :
  - R = % prévu d'enregistrements qui peuvent être réidentifiés correctement
  - U = réciproque de l'inflation de la variance-covariance

Figure 1. R-U Confidentiality Map for Kim's Method with Housing Data

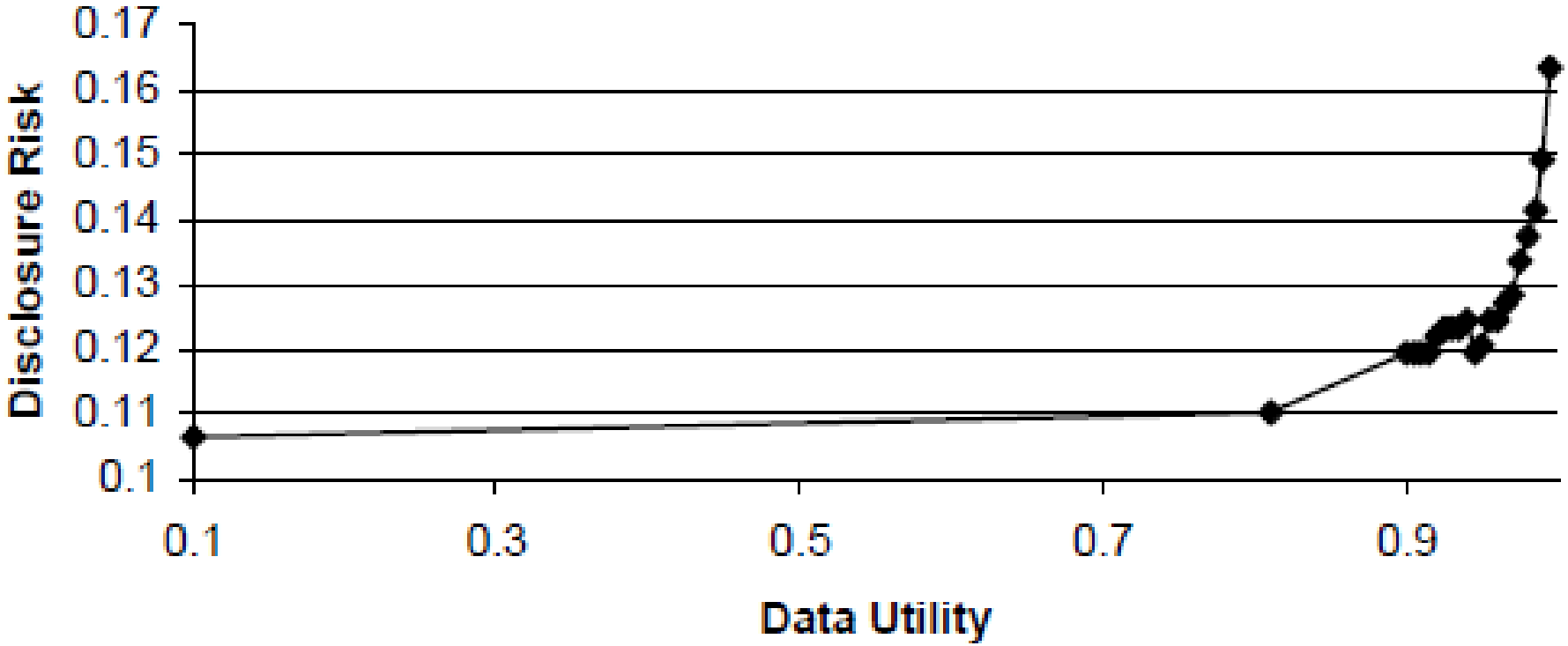


Figure 7. Empirical R-U Confidentiality Map: Comparing Risk for Two Masking Methods when Target is Known to be at 99th Percentile 1994-95 TFS, Private Schools, Part-Time Teachers

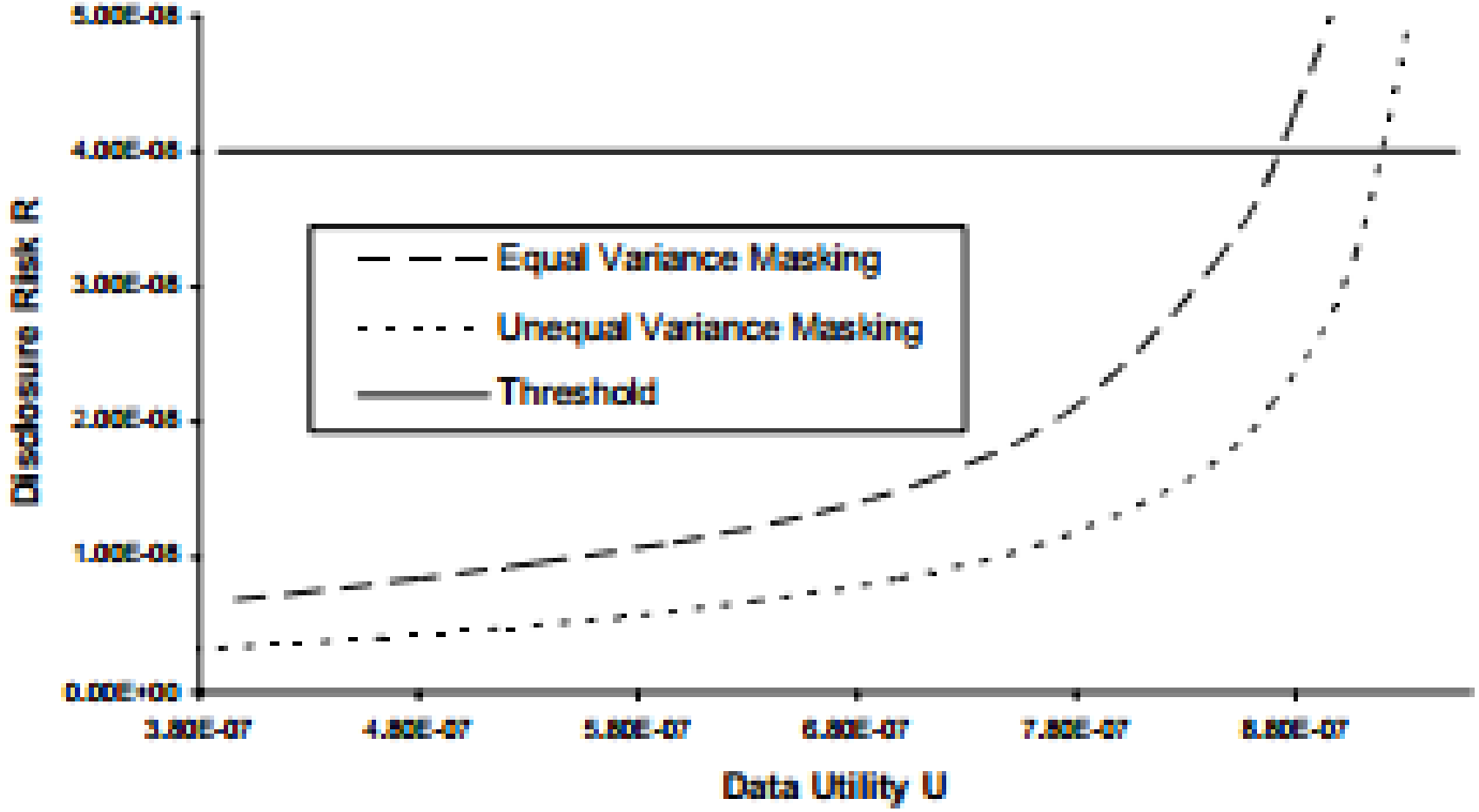


Tableau 5. Mesures de perte d'information pour des microdonnées continues

	Err. quadratique moyenne	Erreur absolue moyenne	Variation moyenne
$X - X'$	$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij}-x'_{ij})^2}{np}$	$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\lvert x_{ij}-x'_{ij}\rvert}{np}$	$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{\lvert x_{ij}-x'_{ij}\rvert}{\lvert x_{ij}\rvert}}{np}$
$V - V'$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}(v_{ij}-v'_{ij})^2}{\frac{p(p+1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\lvert v_{ij}-v'_{ij}\rvert}{\frac{p(p+1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\frac{\lvert v_{ij}-v'_{ij}\rvert}{\lvert v_{ij}\rvert}}{\frac{p(p+1)}{2}}$
$R - R'$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}(r_{ij}-r'_{ij})^2}{\frac{p(p-1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\lvert r_{ij}-r'_{ij}\rvert}{\frac{p(p-1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\frac{\lvert r_{ij}-r'_{ij}\rvert}{\lvert r_{ij}\rvert}}{\frac{p(p-1)}{2}}$
$RF - RF'$	$\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}(rf_{ij}-rf'_{ij})^2}{p^2}$	$\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}\lvert rf_{ij}-rf'_{ij}\rvert}{p^2}$	$\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}\frac{\lvert rf_{ij}-rf'_{ij}\rvert}{\lvert rf_{ij}\rvert}}{p^2}$
$C - C'$	$\frac{\sum_{i=1}^{p}(c_i-c'_i)^2}{p}$	$\frac{\sum_{i=1}^{p}\lvert c_i-c'_i\rvert}{p}$	$\frac{\sum_{i=1}^{p}\frac{\lvert c_i-c'_i\rvert}{\lvert c_i\rvert}}{p}$
$F - F'$	$\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}(f_{ij}-f'_{ij})^2}{p^2}$	$\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}\lvert f_{ij}-f'_{ij}\rvert}{p^2}$	$\frac{\sum_{j=1}^{p}w_j\sum_{i=1}^{p}\frac{\lvert f_{ij}-f'_{ij}\rvert}{\lvert f_{ij}\rvert}}{p^2}$

Mesures de la perte d'information pour les données catégoriques

Quelques mesures suggérées :

- Comparaison directe des valeurs catégoriques (somme de 0/1s, ou de la « distance » entre les catégories pour les variables ordinales)
- Nombre de cellules perturbées
- Comparaison des fréquences par cellule ($\sum c\ |Fc–Gc|$), pour F=orig, G=pert
- Distance de Hellinger : $HD(F,G) = [½\sum c\ (\sqrt{Fc}–\sqrt{Gc})^2]^{½}$
- % d'incidence sur les mesures d'association comme le V de Cramer
- % d'incidence sur la variance des fréquences par cellule
- Perte d'information basée sur l'entropie

Perte d'information basée sur l'entropie (PIBE)

- Entropie nulle, il n'y a aucune incertitude.
  - Jeu de données va toujours donner la même information
- Entropie maximale, maximum d'incertitude.
  - Jeu de données ne va jamais donner la même information.

- Perte d'information égale un ajout d'incertitude, entropie augmente.
- Quand l'entropie augmente, on ajoute du bruit

Perte d'information basée sur l'entropie (PIBE)

- Pour la variable aléatoire X, l'entropie de l'information est la quantité prévue d'information donnée par la connaissance du résultat de X,

$$H(X) = E(-\log(p(X)) = -\sum_x p(x)\log(p(x))$$

- La perte d'information de la modification de X par Y (p. ex., après le recodage ou la perturbation) est l'entropie conditionnelle

$$H(X|Y) = -\sum_y p(y)\sum_x p(x|y)\log(p(x|y))$$

- Dans une base de données masquée, on obtient la PIBE par la somme de H(X|Y) pour toutes les unités individuelles r, c.-à-d. $\sum_r H(X|Y = jr)$

Concepts d'utilité analytique de Winkler

Selon Winkler (1998),

- un jeu de microdonnées est analytiquement valide si les conditions qui suivent sont approximativement satisfaites (certaines conditions ne s'appliquent qu'aux variables continues) :

1. moyennes et covariances sur un petit ensemble de sous-domaines;
2. valeurs marginales pour quelques totalisations des données;
3. au moins une caractéristique distributionnelle;

- un fichier de microdonnées est analytiquement intéressant si six variables sur des sous-domaines importants sont fournies et peuvent être analysées de manière valide.

Caractéristiques influant sur la divulgation

- Types de données
  - Recensement
  - Enquête
  - Autres
- Types de données de sortie
  - Données tabulaires
  - Microdonnées
  - Données analytiques, graphiques, etc.
- Types de variables
  - Qualitatives ou quantitatives
  - Identificateurs directs, identificateurs indirects et variables sensibles

Caractéristiques influant sur la divulgation

- Types de variables
  - **Identificateurs directs**
    - Variables comme le nom, l'adresse, le numéro de téléphone, le numéro d'assurance sociale, etc.

      -> liés à une seule personne
    - Ne sont jamais diffusés
  - **Identificateurs indirects**
    - Variables comme l'âge, le sexe, l'état matrimonial, la profession, le niveau de scolarité, la région géographique, etc.

      -> non liés à une seule personne, mais pourraient l'être si elles sont utilisées ensemble

      (permettent d'identifier les personnes ayant des caractéristiques rares)
    - La diffusion de données relatives à des petites régions géographiques accroît particulièrement le risque de divulgation
  - **Variables sensibles**
    - Variables se rapportant à des personnes et dont la valeur n'est pas largement connue (p. ex. le revenu et les détails de l'état de santé)

Méthodes de contrôle de la divulgation

- Méthodes
  - Accès restreint
  - Diffusion restreinte des données
  - Autres
- Dans les faits, on utilise une combinaison de méthodes

Méthodes de contrôle de la divulgation

- Méthodes d'accès restreint
  - Quelques exemples :
    - Centres de données de recherche
    - Totalisations personnalisées
    - Fichiers de microdonnées à grande diffusion
    - Accès à distance
    - Accès à distance en temps réel

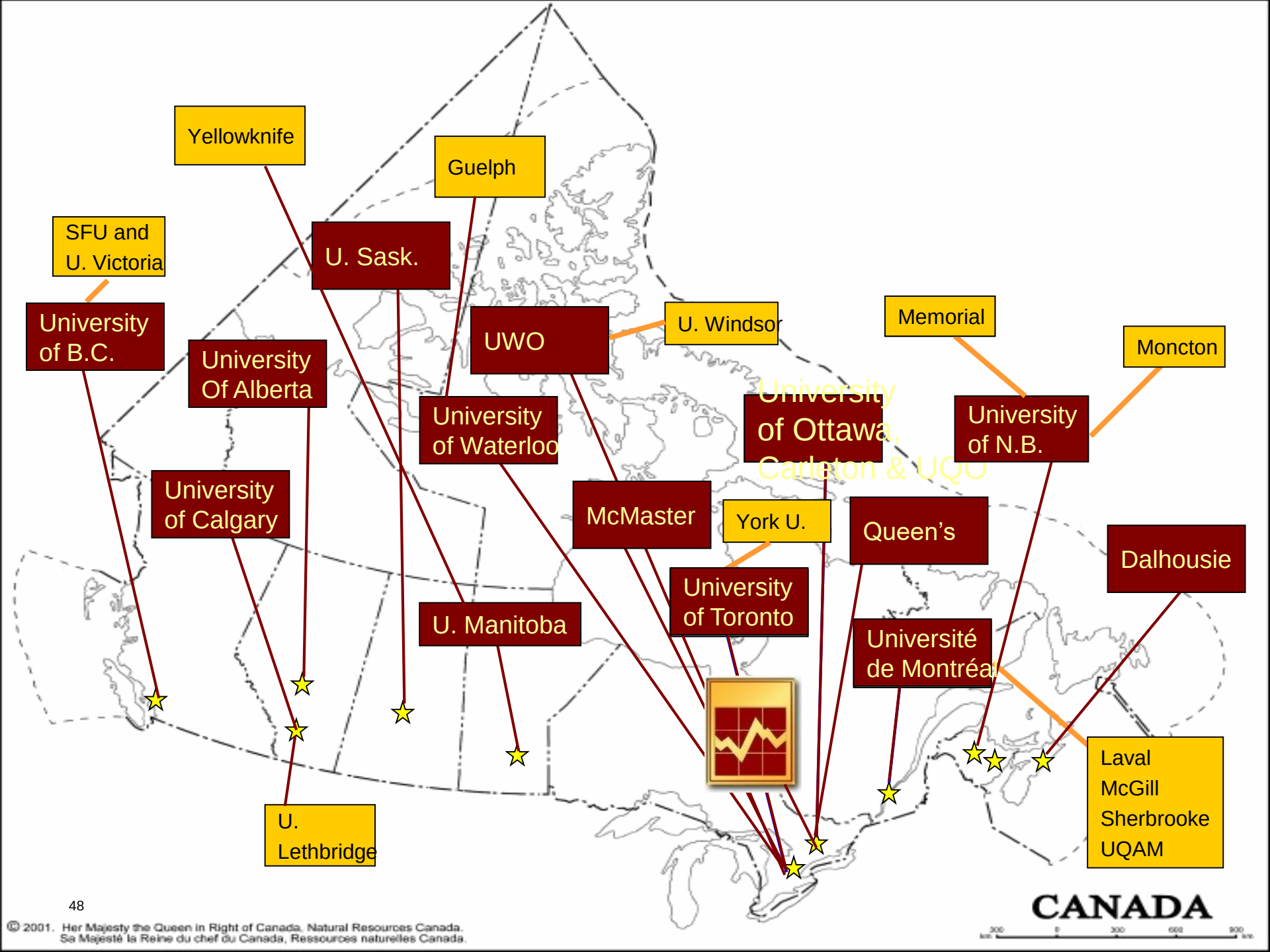
Méthodes de contrôle de la divulgation

- Méthodes d'accès restreint
  - **Centres de données de recherche (CDR)**
    - Accès aux données confidentielles dans un environnement sécurisé
    - Pour les chercheurs dont les projets ont été approuvés et qui ont été assermentés en vertu de la *Loi sur la statistique*
    - Les CDR sont toujours dotés d'un analyste de StatCan qui scrute les sorties de programmes à retirer pour préserver la confidentialité

Les centres de données de recherche

Que sont les Centres de données de recherche ?

- Environnement sécurisé similaire au bureau central (Ottawa)
  - Réseau interne
  - Portes sécurisées
- Fichiers de microdonnées accessibles aux chercheurs
- Analystes de Statistique Canada present en tout temps
- Les chercheurs doivent signer un contrat et être assermentés



© 2001. Her Majesty the Queen in Right of Canada, Natural Resources Canada.
Sa Majesté la Reine du chef du Canada, Ressources naturelles Canada.

Rôle de l'analyste

- Protection des données (micro et macro)
  - S'assurer que les procédures de contrôle de la divulgation sont correctement appliqués
  - Responsable de la protection des données sur place
- Gère les contrats et les comptes d'utilisateurs
- Fourni du support aux chercheurs
  - Explique les concepts des variables d'enquête
  - Formation sur les règles de divulgation à appliquer
  - Etc.

Contrôle de la confidentialité en méthodologie

Les procédures par lesquelles les publications sont approuvées pour sortir des CDR:

- La grande majorité des requêtes de publication des chercheurs est effectuée par les analystes
- Les cas les plus complexes sont transférés au Comité de contrôle de la confidentialité des CDR. Ce comité est formé d'environ 5 à 7 analystes et de 2 à 3 méthodologistes
- Les cas sont étudiés lors de réunions bi-hebdomadaire, puis des recommandations sont retournées aux analystes et chercheurs

Méthodes de contrôle de la divulgation

- Méthodes d'accès restreint
  - **Totalisations personnalisées**
    - Les Services aux clients préparent les totalisations personnalisées
    - Elles permettent aux utilisateurs d'obtenir des résultats personnalisés sans avoir à connaître un logiciel de totalisation
    - Avant d'être communiqués, les résultats sont examinés à des fins de confidentialité

Méthodes de contrôle de la divulgation

- Méthodes d'accès restreint
  - **Accès à distance**
    - Fournit l'accès indirect aux données d'enquête sans en compromettre la confidentialité
    - Les chercheurs autorisés obtiennent de la documentation et des fichiers fictifs (test)
    - Les programmes analytiques sont envoyés par courriel à StatCan
    - On scrute les sorties, pour la confidentialité, avant de les expédier par courriel aux chercheurs

Méthodes de contrôle de la divulgation

- Méthodes d'accès restreint
  - **Accès à distance en temps réel (ADTR)**
    - Accès en ligne permettant des analyses sur des microdonnées dans un endroit central sécurisé
    - Par un système sécurisé restreint à un environnement précis, l'utilisateur soumet des programmes SAS pour créer des tableaux personnalisés à partir de fichiers de microdonnées confidentielles et obtient les résultats
    - Contrôles de sécurité automatisés des programmes et des résultats
    - Fournit un accès plus rapide et flexible aux données
    - Opérationnel, mais encore en développement

Méthodes de contrôle de la divulgation

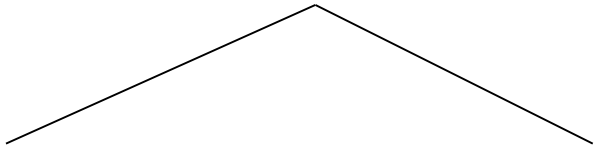
- Méthodes de diffusion restreinte des données
  - Méthodes de contrôle des entrées (intrants) : application de mesures aux microdonnées
  - Méthodes de contrôle des sorties : application de mesures aux données de sortie diffusées seulement
- Autres méthodes
  - Dérogations
  - Ententes de partage de données

Méthodes de contrôle de la divulgation

- Méthodes de diffusion restreinte des données
  - Méthodes de contrôle des entrées (intrants): application de mesures aux microdonnées
  - Méthodes de contrôle des sorties : application de mesures aux données de sortie diffusées seulement
- Autres méthodes
  - Dérogations
  - Ententes de partage de données

Méthodes de contrôle de la divulgation

- Méthodes de diffusion restreinte des données
  - Méthodes de contrôle à l'entrée : application de mesures aux microdonnées



Méthodes de réduction des données

- Recodage (regroupement des valeurs extrêmes supérieures et inférieures et recodage par intervalles)
- Suppressions locales, p. ex. valeurs extrêmes
- Sous-échantillonnage

Méthodes de perturbation des données

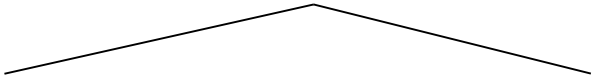
- Permutation de données entre enregistrements
- Imputation de données
- Perturbation aléatoire
- Ajout de bruit
- Création de fichiers synthétiques

Méthodes de contrôle de la divulgation

- Méthodes de diffusion restreinte des données
  - Méthodes de contrôle des entrées : application de mesures aux microdonnées
  - Méthodes de contrôle des sorties : application de mesures aux données de sortie diffusées seulement
- Autres méthodes
  - Dérogations
  - Ententes de partage de données

Méthodes de contrôle de la divulgation

- Méthodes de diffusion restreinte des données
  - Méthodes de contrôle des sorties : application de mesures aux données de sortie diffusées seulement



Méthodes de réduction des données

- Regroupement de catégories
- Suppression de cellules
- Suppression de tableaux

Méthodes de perturbation des données

- Arrondissement
- Perturbation aléatoire des cellules

  - Doit développer une méthode pour chacune des statistiques et types de sorties

REMARQUE : Il est aussi possible de jumeler des méthodes de contrôle des entrées à des méthodes de contrôle des sorties

Aspects à considérer pour l'établissement d'une méthode appropriée

- Variables diffusées :
  - Niveaux de sensibilité
  - Niveaux géographiques
- Produits et données statistiques diffusés
- Types d'utilisateurs (p. ex. étudiants, analystes des politiques, chercheurs, etc.)
- Autres produits dérivés disponibles
- Méthode de mise en œuvre des règles (automatiquement ou manuellement)
- Niveaux de priorité en termes de :
  - Qualité, Cohérence
  - Confidentialité, Reproductibilité
  - Flexibilité, Complexité

Tableaux

Protection des totaux

- Les totaux sont sensibles à cause des tendances à la dominance dans la répartition
- Un petit nombre d'unités contribuent à une grande proportion des cellules
- C'est souvent le cas des variables dans les enquêtes économiques ou des variables liées au revenu

Le répondant le deuxième plus grand en importance (vert) soustrait sa réponse du total afin de faire une estimation de la réponse du répondant plus grand.

Confidentialité non assurée

Confidentialité assurée

Protection des totaux

- Détermination des cellules sensibles (cas de prédominance et petites cellules)
- Détermination de la méthode de protection de ces cellules, par exemple :
  - Suppression de cellules
  - Agrégation de cellules
  - Ajout de bruit (microdonnées ou agrégats)

Protection des totaux

- Il existe plusieurs types de règles pour déterminer s'il y a dominance :
  - Règles (n,k)
  - Règles de Duffett, règle C2 (règles de StatCan)
  - Règles du pourcentage p
  - Règles pq

Protection des totaux

- Règle (n,k) : la cellule est sensible si les n valeurs parmi les plus importantes représentent plus de k % de la valeur globale de la cellule
  - P. ex. (3,75) : si la valeur des trois plus importants répondants est plus grande que 75 % de la valeur globale de la cellule, alors la cellule est sensible
- Règle p % – la cellule est sensible si :

  $x[1] * (p/100) > x[3] + x[4] + x[5] + ...$
- Combinaison des règles (n,k) et p %
- Les paramètres sont confidentiels!

Suppression de cellules complémentaires

- Après la suppression des cellules sensibles, il faut supprimer des cellules complémentaires

- On choisit des compléments qui réduisent :
  - la suppression de cellules d'intérêt principal pour les utilisateurs
  - le nombre total de cellules supprimées
  - la valeur totale des cellules supprimées

- La suppression n'empêche pas les utilisateurs d'obtenir une fourchette pour la valeur de chaque cellule supprimée – pour vérifier ceci on a besoin de programmes spécialisés comme G-Confid ou Tau-argus

Protection de tableaux

- Chiffres
  - Les cellules contiennent des nombres d'unités
  - Problèmes de divulgation liés aux tableaux de chiffres :
    - Petites cellules
    - Cellules à fréquence nulle et cellules pleines
    - Divulgation par recoupement
  - Les méthodes couramment utilisées sont appliquées après la totalisation (sur les données de sortie) :
    - Suppression de cellules
    - Regroupement de catégories
    - Arrondissement

Protection de tableaux

- Chiffres
  - Les cellules contiennent des nombres d'unités
  - Problèmes de divulgation liés aux tableaux de chiffres :
    - Petites cellules
    - Cellules à fréquence nulle et cellules pleines
    - Divulgation par recoupement
  - Les méthodes couramment utilisées sont appliquées après la totalisation (sur les données de sortie) :
    - Suppression de cellules
    - Regroupement de catégories
    - Arrondissement

Protection de tableaux

- Problèmes de divulgation liés aux tableaux de chiffres
  - **Petites cellules**
    - Cellules ayant peu d'observations (p. ex. < 3)
    - Peuvent présenter un risque de divulgation pour les données de recensement
      - p. ex. une personne d'une région est un joueur de hockey et est atteint du SIDA
    - Donnent souvent l'impression qu'une information confidentielle a été divulguée – même si cela n'est pas le cas

Protection de tableaux

- Problèmes de divulgation liés aux tableaux de chiffres
  - **Cellules à fréquence nulle et cellules pleines**
    - Peuvent entraîner la divulgation d'attribut si des informations sensibles sont révélées
    - Les cellules à fréquence nulle ne renferment aucun répondant
      - Une cellule à fréquence nulle structurelle représente une combinaison non valide (p. ex. marié et moins de 15 ans)
    - Les cellules pleines renferment tous les répondants d'un total marginal
      - Par exemple, la seule cellule non nulle dans une rangée ou une colonne

Municipalité XYZ

		Revenu < 25 000 $	Revenu ≥ 25 000 $	Total
Hommes	immigrants	26	30	56
Homes	non immigrants	200	78	278
Femmes	immigrantes	39	0	39
Femmes	non immigrantes	250	61	311
Total		515	169	684

Cellules pleines

Protection de tableaux

- Problèmes de divulgation liés aux tableaux de chiffres
  - **Divulgation par recoupement**
    - Se produit lorsque des renseignements diffusés sont combinés à d'autres renseignements diffusés, ou à des renseignements provenant de sources externes, pour révéler des données confidentielles
    - Exemple :

	Immigrants	Non Immigrants	Total
Hommes	56	278	334
Femmes	**X**	311	312
Total	57	589	646

Protection de tableaux

- Chiffres
  - Les cellules contiennent des nombres d'unités
  - Problèmes de divulgation liés aux tableaux de chiffres :
    - Petites cellules
    - Cellules à fréquence nulle et cellules pleines
    - Divulgation par recoupement
  - Les méthodes couramment utilisées sont appliquées après la totalisation (sur les données de sortie) :
    - Suppression de cellules
    - Regroupement de catégories
    - Arrondissement

Protection de tableaux

- **Suppressions**
  - Exemple 1

Revenu / Âge	< 250 000 $	≥ 250 000 $	Total
15-64	311	36	347
65+	561	1	562
Total	872	37	909

Protection de tableaux

- **Suppressions**
  - Exemple 1

Revenu / Âge	< 250 000 $	≥ 250 000 $	Total
15-64	311	36	347
65+	561	X	562
Total	872	37	909

Protection de tableaux

- **Suppressions**
  - Exemple 1

Revenu / Âge	< 250 000 $	≥ 250 000 $	Total
15-64	X	X	347
65+	X	X	562
Total	872	37	909

-> Des suppressions secondaires sont nécessaires

Protection de tableaux

- **Suppressions**
  - Exemple 2

Est-il possible de déterminer Z?

	Région A	Région B	Région C	Région D	Total
Condition 1	X	X	62	34	213
Condition 2	X	X	15	4	182
Condition 3	61	**Z**	X	X	253
Condition 4	29	64	X	X	200
Total	264	215	204	165	848

Protection de tableaux

- **Suppressions**
  - Exemple 2

Est-il possible de déterminer Z?

	Région A	Région B	Région C	Région D	Total
Condition 1	X	X	62	34	213
Condition 2	X	X	15	4	182
Condition 3	61	**Z**	X	X	253
Condition 4	29	64	X	X	200
Total	264	215	204	165	848

115

395

Protection de tableaux

- **Suppressions**
  - Exemple 2

Est-il possible de déterminer Z?

	Région A	Région B	Région C	Région D	Total
Condition 1			62	34	213
Condition 2			15	4	182
Condition 3	61	**Z**	X	X	253
Condition 4	29	64	X	X	200
Total	264	215	204	165	848

280

115

395

479

Protection de tableaux

- **Suppressions**
  - Exemple 2

Est-il possible de déterminer Z?

115

	Région A	Région B	Région C	Région D	Total
Condition 1			62	34	213
Condition 2			15	4	182
Condition 3	61	**45**	X	X	253
Condition 4	29	64	X	X	200
Total	264	215	204	165	848

280

395

479

Protection de tableaux

- **Suppressions**
  - Exemple 3 (réaction en chaîne)

Région géographique n° 1

34	46	80
41	32	73
75	78	153

Région géographique n° 2

27	31	58
3	19	22
30	50	80

Total des régions géographiques

61	77	138
44	51	95
105	128	233

Protection de tableaux

- **Suppressions**
  - Exemple 3 (réaction en chaîne)

Région géographique n° 1

34	46	80
41	32	73
75	78	153

Région géographique n° 2

27	31	58
X	19	22
30	50	80

Total des régions géographiques

61	77	138
44	51	95
105	128	233

Protection de tableaux

- **Suppressions**
  - Exemple 3 (réaction en chaîne)

Région géographique n° 1

34	46	80
41	32	73
75	78	153

Région géographique n° 2

X	X	58
X	X	22
30	50	80

Total des régions géographiques

61	77	138
44	51	95
105	128	233

Protection de tableaux

- **Suppressions**
  - Exemple 3 (réaction en chaîne)

Région géographique n° 1

X	X	80
X	X	73
75	78	153

Région géographique n° 2

X	X	58
X	X	22
30	50	80

Total des régions géographiques

61	77	138
44	51	95
105	128	233

-> Sursuppression : pour protéger une cellule, il faut en supprimer huit!

Protection de tableaux

- **Suppressions**
  - **Aucune protection contre la divulgation par recoupement**
  - Exemple 4

TABLEAU A : Cancer, tous types

	Région		
ÂGE	A	B	Total
15-66	10 554	31 985	42 539
65+	2 756	4 426	7 182
Total	13 310	36 411	49 721

TABLEAU B : Cancer, types courants (excluant les types très rares)

	Région		
ÂGE	A	B	Total
15-66	10 553	31 984	42 537
65+	2 752	4 409	7 161
Total	13 305	36 393	49 698

-> L'un des tableaux (A ou B) peut être diffusé mais pas les deux, sans quoi :

TABLEAU ÉTABLI PAR RECOUPEMENT : Cancer, types très rares

	Région		
ÂGE	A	B	Total
15-66	1	1	2
65+	4	17	21
Total	5	18	23

Protection de tableaux

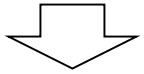
- **Suppressions**
  - **Avantages**
    - Facile à effectuer (suppressions primaires)
    - Méthode intuitive
    - Pas de perturbation (l'intégrité des données est préservée)
  - **Inconvénients**
    - Les suppressions secondaires exigent le recours à un programme hautement spécialisé
    - Méthode pratiquement impossible à mettre en œuvre dans un environnement où les demandes sont nombreuses et où un nombre illimité de tableaux peuvent être diffusés à partir d'un même fichier de base (ou d'autres fichiers)

      -> Les modèles de suppression et tous les tableaux diffusés doivent être conservés
    - Les suppressions secondaires masquent les données se trouvant dans des cellules sûres
    - Possibilité de sur-suppression, donc d'une grande perte d'information
    - Aucune protection contre la divulgation par recoupement

Protection de tableaux

- **Regroupement de catégories**
  - Exemple :

	Région A	Région B	Région C	Région D
Condition 1	**2**	72	62	34
Condition 2	6	54	15	54
Condition 3	61	89	143	97
Condition 4	29	64	45	163



	Régions A et B	Région C	Région D
Condition 1	74	62	34
Condition 2	60	15	54
Condition 3	150	143	97
Condition 4	93	45	163

Protection de tableaux

- **Regroupement de catégories**
  - **Avantages**
    - Aucune cellule n'est supprimée
    - Aucune perturbation (l'intégrité des données est préservée)
  - **Inconvénients**
    - Les tableaux doivent être redessinés; cette opération est difficile à programmer et doit être faite manuellement
    - Aucune protection contre la divulgation par recoupement

Protection de tableaux

- **Arrondissement**
  - La base d'arrondissement protège les chiffres
  - Types d'arrondissement :
    - Arrondissement classique
    - Arrondissement aléatoire
    - Arrondissement contrôlé
    - Autres

Protection de tableaux

- **Arrondissement classique**
  - Les cellules sont arrondies au multiple le plus près de la base
  - Exemple :
    - Agence du revenu du Canada – élimination graduelle de la pièce d'un cent et arrondissement des cents au multiple de 5 le plus près :

Lignes directrices sur l'arrondissement

Protection de tableaux

- **Arrondissement classique**
  - Les cellules sont arrondies au multiple de la base le plus près
  - Toutes les cellules sont traitées de façon indépendante
  - Exemple selon une base d'arrondissement de 5 :

Tableau d'origine

1	2	3
7	7	14
8	9	17

Tableau avec arrondissement

0	0	5
5	5	15
10	10	15

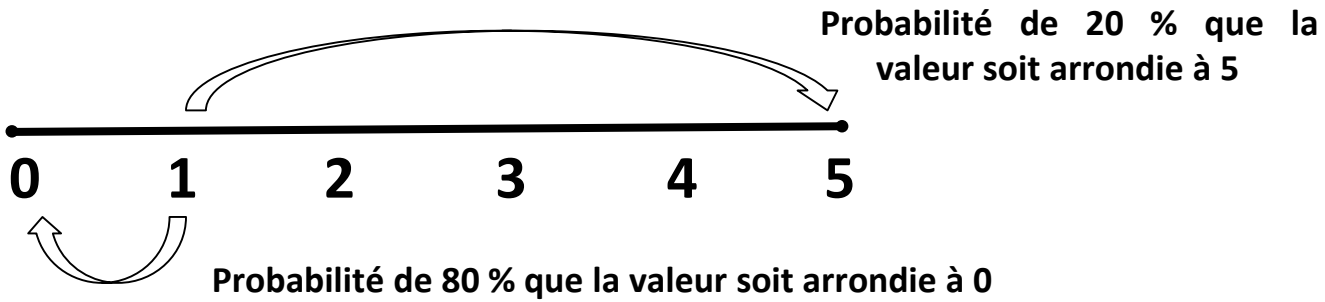
Protection de tableaux

- **Arrondissement classique**
  - Avantages :
    - Méthode bien connue et très facile à appliquer
    - Garantie d'une distorsion minimale
  - Inconvénients :
    - Les totaux ne correspondent pas à la somme des chiffres
    - Les valeurs peuvent parfois être reconstituées (par essai et erreur)

Protection de tableaux

- **Arrondissement aléatoire**
  - Les cellules sont :
    - arrondies à la hausse selon une probabilité p
    - arrondies à la baisse selon une probabilité 1 – p
  - Le choix de probabilités permet de s'assurer que l'opération n'est pas biaisée
  - Toutes les cellules sont traitées de façon indépendante

  - Exemple selon une base d'arrondissement de 5 :



Protection de tableaux

- **Arrondissement aléatoire**
  - Exemple selon une base d'arrondissement de 5 :

Tableau d'origine

1	2	3
7	7	14
8	9	17

Tableau avec arrondissement (une possibilité)

5	0	5
5	10	10
5	10	20

Protection de tableaux

- **Arrondissement aléatoire**
  - Avantages :
    - Facile à utiliser
    - Le programme d'arrondissement aléatoire est simple
    - Meilleure protection que l'arrondissement classique
  - Inconvénients :
    - Les totaux ne correspondent pas à la somme des chiffres
    - La distorsion peut être plus élevée et difficile à contrôler

Protection de tableaux

- **Arrondissement contrôlé**
  - Toutes les cellules sont arrondies de manière à préserver le caractère additif du tableau
  - Exemple selon une base d'arrondissement de 5 :

Tableau d'origine

1	2	3
7	7	14
8	9	17

Tableau avec arrondissement (une possibilité)

0	5	5
5	5	10
5	10	15

  - Problème :
    - Tous les tableaux à deux dimensions ont une solution arrondie additive
    - Certains tableaux à trois dimensions ou plus n'ont pas une telle solution (bien qu'ils puissent être additifs à ~99 %)

Protection de tableaux

- **Arrondissement contrôlé**
  - Avantages :
    - Plus complexe à programmer, mais il existe plusieurs programmes pour ce faire
    - Modérément facile à utiliser
    - Les tableaux sont additifs
    - Des tableaux additifs permettent plus difficilement aux pirates d'exploiter la structure du tableau pour obtenir les valeurs d'origine
    - Moins de distorsion qu'avec l'arrondissement aléatoire
  - Inconvénients :
    - Comme il n'existe pas toujours de solution pour certains tableaux à trois dimensions ou plus, il faut faire des compromis (mis en œuvre dans le programme)
    - Plus de distorsion qu'avec l'arrondissement classique, mais moins qu'avec l'arrondissement aléatoire

Protection de tableaux

- **Arrondissement : perte d'information**
  - Comparaison des trois méthodes :

Tableau d'origine

1	2	3
7	7	14
8	9	17

Arrondissement classique

0	0	5
5	5	15
10	10	15

1 + 2 + 2 + 2 = **7**

Arrondissement aléatoire

5	0	5
5	10	10
5	10	20

4 + 2 + 2 + 3 = **11**

Arrondissement contrôlé

0	5	5
5	5	10
5	10	15

1 + 3 + 2 + 2 = **8**

Protection de tableaux

- **Choix d'une base d'arrondissement**
  - Le choix doit être en fait en fonction du risque de divulgation que présente les données
  - Il faut trouver un compromis :
    - plus la base est élevée, plus la perte d'information attribuable à l'arrondissement est grande
    - plus la base est faible, plus il est difficile de prévenir la divulgation
  - Exemple : Statistique Canada utilise une base de 5 pour l'arrondissement aléatoire des données du recensement

Protection de tableaux

- **L'arrondissement offre plus de protection contre la divulgation par recoupement**
  - Retour à l'exemple 4

TABLEAU A : Cancer, tous types

	Région		
ÂGE	A	B	Total
15-66	10 555	31 985	42 540
65+	2 755	4 430	7 185
Total	13 310	36 415	49 725

TABLEAU B : Cancer, types courants (excluant les types très rares)

	Région		
ÂGE	A	B	Total
15-66	10 550	31 985	42 535
65+	2 755	4 410	7 165
Total	13 305	36 395	49 700

TABLEAU ÉTABLI PAR RECOUPEMENT : Cancer, types très rares

	Région		
ÂGE	A	B	Total
15-66	5	0	5
65+	0	20	20
Total	5	20	25

Protection de tableaux

Exemple :
Les valeurs nulles sont-elles vraies ou fausses?

15	15	0	30
15	0	25	40
30	15	25	70

- **Arrondissement**
  - **Avantages**
    - Des chiffres sont fournis pour toutes les cellules
    - Les valeurs nulles sont plus ambiguës
    - Facile à mettre en œuvre dans un environnement où les demandes et les fichiers sont nombreux
    - Protection contre une fausse impression de précision parfaite, particulièrement en ce qui concerne les valeurs faibles
    - Perception : des mesures sont prises pour protéger la confidentialité
    - Protection contre la divulgation par recoupement
  - **Inconvénients**
    - Données perturbées
    - Incidence plus grande sur les petits chiffres, bien que ce soit eux que l'on veuille protéger
    - À moins d'utiliser l'arrondissement contrôlé, la non-additivité peut constituer un problème

Logiciels à Statistique Canada G-CONFID, G-TAB et l'ADTR

Mise en contexte

- Statistique Canada reçoit chaque jour un nombre grandissant de requêtes ad-hoc
  - Répondre à toutes ces requêtes demande énormément de temps aux analystes
- Jusqu'à 2010, les modes d'accès aux données étaient:
  - Fichier de microdonnées à grande diffusion
  - Centre de données de recherche
- Pression pour plus d'accès aux données

  ⇒Besoin de systèmes automatisés
- Deux nouvelles initiatives (Enquêtes Sociales)
  - **Accès direct en temps réel (ADTR) – Utilisateurs externes**
  - **G-TAB - Utilisateurs de Statistique Canada**

- G-CONFID (Enquêtes Économiques)

Mise en contexte

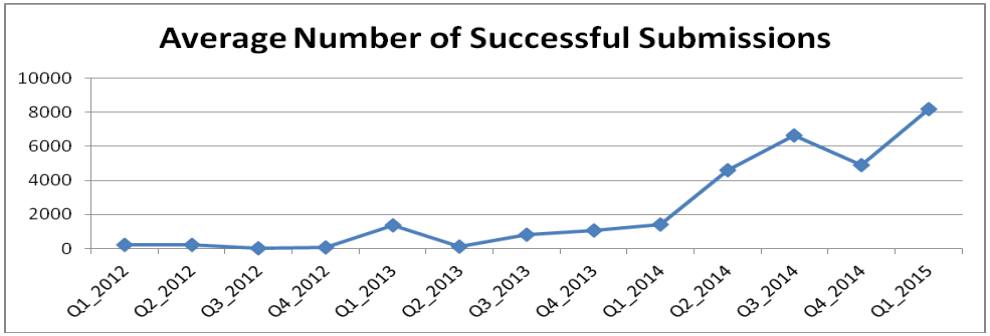
- **G- CONFID**
  - Données avec dominance – enquêtes économmiques
  - Suppressions primaires et complémentaires
  - Produit des Schémas de suppression

  - Nouveau en 2015 :
    - Arrondissement contrôlé additive (intégré avec G-TAB)

  - Automne 2016, le traitement automatisé de:
    - Données en présence de dérogation
    - Variables contenant des valeurs positives et négatives
      - Valeurs absolues
      - Combinaisons linéaires incluant des variables proxy
    - Poids de sondage

Current Status – Social Surveys

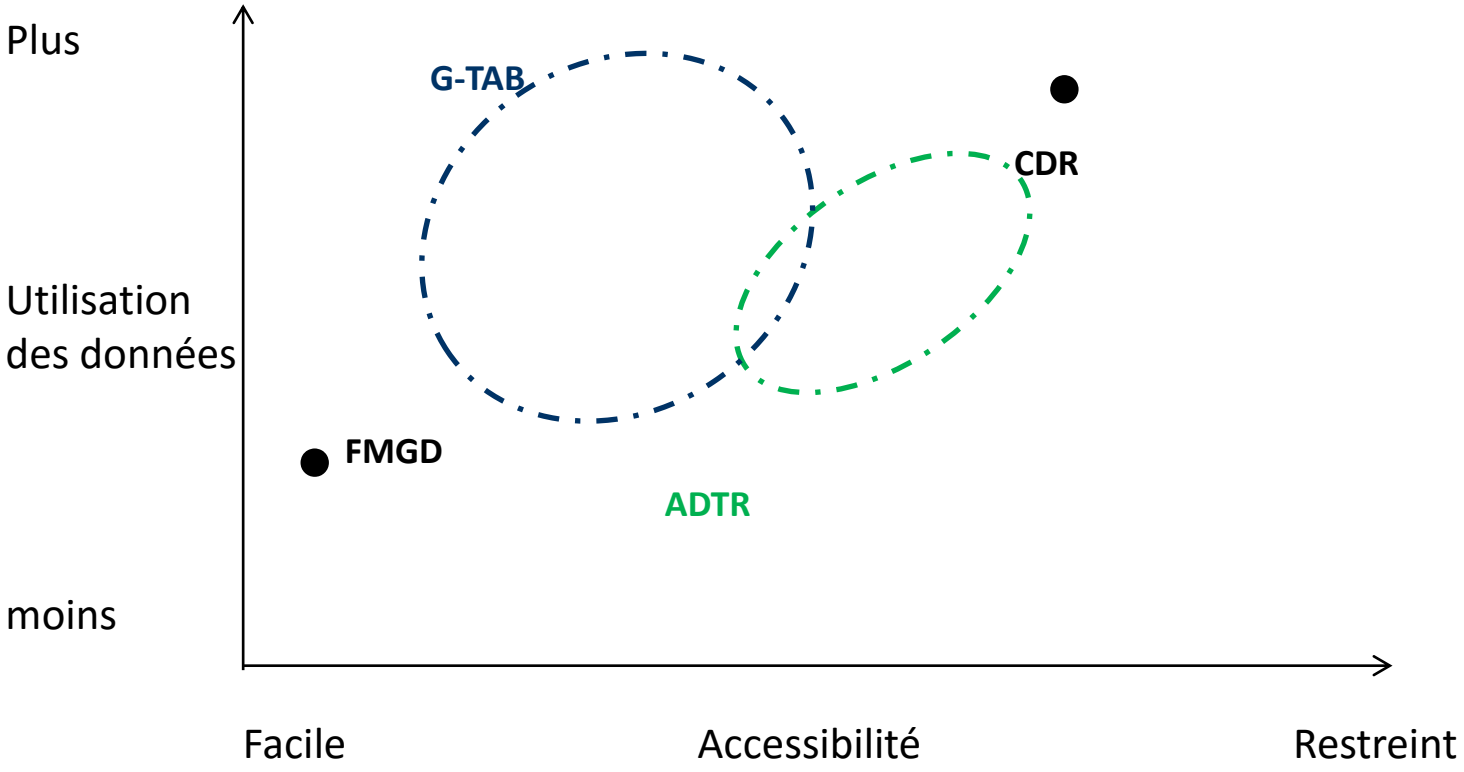
- **Accès à distance en temps réel (ADTR)**
  - Débutée en 2009
  - 29 enquêtes sociales (+ cycles), 5 fichiers administratifs
  - Utilisateurs internationaux



- **Generalized Tabulation System (G-TAB)**
  - Integré avec le nouveau modèle de diffusion (Open data)
  - Interface convivial

- Pris les mêmes modules TI pour les deux outils

Mise en contexte

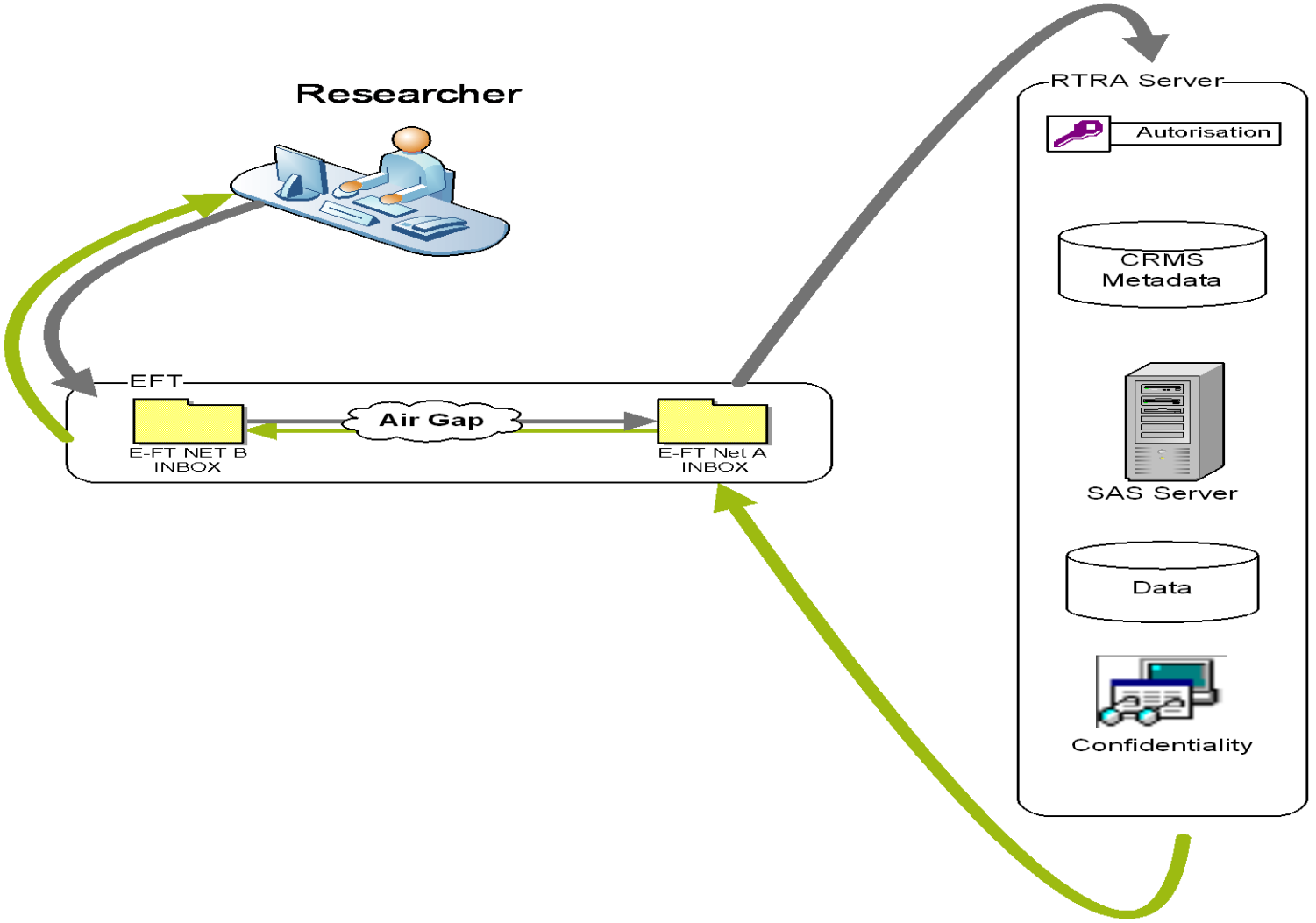


Accès direct en temps réel (ADTR)

Accès direct en temps réel (ADTR)

- Accessible de n'importe quel ordinateur ayant un accès à Internet en utilisant un nom d'utilisateur et mot de passe
- Les utilisateurs doivent faire une demande décrivant l'objet de leur travail
- L'ADTR est disponible 24h / 7 jours
- Disponible depuis le printemps 2010
- Accessible en SAS seulement
  - Ensemble de macros prédéfinies
  - data step, proc sort, modified proc freq, etc.

Accès direct en temps réel (ADTR)



Statistics Canada | Statistique Canada

Welcome to
Electronic File Transfer

Bienvenue au
Service de transfert de fichiers électroniques

English | Français

Canada

Important Notices | Avis importants

Statistics Canada
Statistique Canada
Canada
Statistics Canada
www.statcan.gc.ca
Français
Home
Contact Us
Help
Search
canada.gc.ca
Home > Electronic File Transfer >
EFT
Frequently Asked Questions (FAQ)
About Us
Logon page
To get started, enter your Username and Password. Then press Logon.
Username:
2416
Password:
Logon
Date Modified: 2009-12-30
Top of Page
Important Notices

Statistics Canada Statistique Canada

Canada

Statistics Canada

www.statcan.gc.ca

Français | Home | Contact Us | Help | Search | canada.gc.ca

Home > Electronic File Transfer >

EFT

Logout

Search

Safe List

RTRA-Testing Project

FromStatcan

ToStatcan

Safe MAD_RTRA_tst_366_2416 - ToStatcan

Select the action you want to perform on the file

Upload file

No File exists in this folder.

Date Modified: 2009-12-30

Top of Page

Important Notices

Statistics Canada
Statistique Canada
Canada
Statistics Canada
www.statcan.gc.ca
Français
Home
Contact Us
Help
Search
canada.gc.ca
Home > Electronic File Transfer >
EFT
Logout
Search
Safe List
RTRA-Testing Project
FromStatcan
ToStatcan
Safe RTRA-Testing Project - ToStatcan
File (required)
Browse...
Upload
Date Modified: 2009-12-30
Top of Page
Important Notices

gss2005_demotest.sas

Text | Properties

File Tools

Find

```
data work.Table1(keep=spa_sp91 actlimit id);
   set RTRAdata.c19analm_num;
run;

%RTRAFreq(
   InputDataset=work.Table1,
   OutputName=DemoTest,
   ClassVarList=spa_sp91 actlimit,
   UserWeight=wght_per);
```

Row: 11, Col: 1

c:\users\belbmou\desktop\gss2005_demotest.sas

Statistics Canada Statistique Canada

Canada

Statistics Canada
www.statcan.gc.ca

Français | Home | Contact Us | Help | Search | canada.gc.ca

Home > Electronic File Transfer >

EFT

Logout

Search

Safe List

RTRA-Testing Project

FromStatcan (5)

ToStatcan

Safe MAD_RTRA_tst_366_2416 - FromStatcan

Select the action you want to perform on the file

Files list; Page 1 of 1

File name	Creation date	Last used date	Size	
RTRA036597_DemoTest.htm	18/11/2011 12:02:03 PM	18/11/2011 12:02:03 PM	56KB	Download
RTRA036597_demotest.sas7bdat	18/11/2011 12:02:04 PM	18/11/2011 12:02:04 PM	5KB	Download
RTRA036597_GSS2005_DemoTest.log	18/11/2011 12:02:05 PM	18/11/2011 12:02:05 PM	1KB	Download
RTRA036597_GSS2005_DemoTest.sas	18/11/2011 12:02:05 PM	18/11/2011 12:02:05 PM	1KB	Download
RTRA036597_GSS2005_DemoTest.txt	18/11/2011 12:02:05 PM	18/11/2011 12:02:05 PM	1KB	Download

Date Modified: 2009-12-30

Top of Page

Important Notices

RTRA036597_GSS2005_DemoTest.log - Notepad

File Edit Format View Help

```
18  +data work.Table1(keep=spa_sp91 actlimit id);
19  +   set RTRAdata.c19analm_num;
20  +run;

NOTE: There were observations read from the data set RTRADATA.C19ANALM_NUM.
NOTE: The data set WORK.TABLE1 has observations and 3 variables.
NOTE: DATA statement used (Total process time):
      real time           0.03 seconds
      cpu time            0.01 seconds


21  +
22  +%RTRAFreq(
23  +   InputDataset=work.Table1,
24  +   OutputName=DemoTest,
25  +   ClassVarList=spa_sp91 actlimit,
26  +   UserWeight=wght_per);

NOTE: RTRAFreq macro called with input data set: work.Table1
NOTE: La macro RTRAFreq appelée avec l'ensemble de données d'entrées: work.Table1


NOTE: RTRAFreq macro completed successfully.
NOTE: macro RTRAFreq complétée avec succès.

27  +
```

spa_sp91 * actlimit

		actlimit				
		1	2	8	9	Total
spa_sp91						
2	N	3,000	7,000	0	0	10,000
	SE	396	885	x	x	973
	QI	d	d	x	x	d
7	N	3,137,000	22,366,000	93,000	350,000	25,946,000
	SE	13,800	14,900	2,640	4,760	3,650
	QI	a	a	a	a	a
8	N	2,000	66,000	63,000	0	131,000
	SE	270	2,370	2,260	x	3,340
	QI	d	a	a	x	a
9	N	0	15,000	0	4,000	19,000
	SE	x	1,000	x	557	1,180
	QI	x	d	x	d	a
Total	N	3,142,000	22,454,000	156,000	354,000	26,106,000
	SE	13,800	14,700	3,420	4,810	0.0000001
	QI	a	a	a	a	a

Accès direct en temps réel (ADTR)

- Présentement, les utilisateurs peuvent accéder à des fréquences et 6 statistiques descriptives (données pondérés seulement)
- Enquête et source administrative disponible:

-ACS
-APS
- MES
- LSIW
- SAEP
- ASETS
- CFMHS
- UCR

-PALS
-CCHS
-NGS
- HSAS
- PSES
- AETS
- PSIS
- SOW

- CFCS
- CUIS
-LFS
-GSS
- CCR
- RAIS
- SFS

- SLCDC
- PEPS
- NAS
-SLID
- EDS
- LASS
- SYC

Accès direct en temps réel (ADTR)

- **Mesures de protection complémentaires aux règles de contrôle de la divulgation**
  - Niveau de géographie le plus détaillé est la province
  - Pour les enquêtes, seulement des statistiques pondérées sont permises
  - Tous les “logs” sont revus et modifiés (ex: Supprimer le nombres d’observations des fichiers)
  - Nombre maximal de requêtes par jour: 10
  - Nombre maximal de tables par requête: 10
  - Nombre de variables de domaines: 5
  - Les variables de domaines doivent avoir moins de 500 valeurs distinctes (ex: afin d’éviter de faire un tableau de fréquence par revenu)

G-TAB

G-TAB

- G-TAB est un système interne qui standardise la façon dont les clients:
  - Gère et créer les spécifications pour leurs publications
    - Tableaux CANSIM
    - Requêtes ad-hoc
  - Génère des statistiques et des indicateurs de qualité (Fréquence, Moyenne, Percentile, Proportion, Variance, CV, etc.)
  - Applique des règles de confidentialité
  - Crée des sorties formatées pour les outils de publication (G-Export et CANSIM)

G-TAB

- **Statistiques disponibles dans G-TAB**
  - Fréquence pondérée
  - Proportion
  - Moyenne
  - Percentiles (P1, P5, P10, P20, P25, P30, P40, P50, P60, P70, P75, P80, P90, P95, P99)
  - Ratios (Type 1 and 2)
  - Coefficient de Gini
  - Moyenne géométrique
  - Niveau de changement
  - Pourcentage de changement
  - Test statistique (Z-test)
  - Moyenne Mobile
  - Quantile (Quartile, Quintile et Décile)

Microdonnées

Microdonnées

- Les chercheurs ont besoin de microdonnées pour leurs analyses – les données agrégées ne sont utiles que pour des analyses grossières
- Le défi pour les organismes statistiques est de donner accès aux microdonnées sans compromettre la confidentialité de ces données
- Les organismes statistiques peuvent faire appel aux moyens suivants :
  - Fichiers de microdonnées anonymisées
  - Fichiers de microdonnées synthétiques
  - Accès à distance
  - Accès aux données confidentielles dans des laboratoires de données

Fichiers de microdonnées à grande diffusion

- Fichiers de données individuelles anonymisées provenant d'un échantillon d'unités
- Destinés au grand public – les données tabulaires tout usage ne permettent pas de réaliser des analyses approfondies

Protection des fichiers de microdonnées

- On vise à protéger l'identité des répondants
- Les risques de divulgation sont plus importants lorsqu'il s'agit de FMGD
- Ils sont « riches » en information (>100 variables)
- On s'intéresse particulièrement aux uniques dans la population qui se trouvent dans l'échantillon (on ignore de quelles unités il s'agit)

- L'objet est de réduire le nombre d'uniques dans l'échantillon et/ou d'introduire un élément d'incertitude dans les données
- Les méthodes peuvent être appliquées globalement (à l'ensemble des enregistrements) ou localement (à des enregistrements spécifiques)

Protection des fichiers de microdonnées

- Retrait de tous les identificateurs directs
- Réduction du nombre d'identificateurs indirects (géographiques)
- Regroupement des valeurs extrêmes supérieures, recodage, etc.

- Sous-échantillonnage, post-randomisation, microagrégation
- Permutation de données, permutation de rangs, recalage, données synthétiques

- Évaluation des enregistrements uniques
  - Analyses unidimensionnelles – vérification de la présence d'observations aberrantes
  - Examen des combinaisons de variables : âge et état matrimonial, profession et industrie, différence d'âge entre les conjoints...
  - Analyse des tableaux à trois dimensions

Usage de la multiplicité à StatCan

- La multiplicité est utilisée comme autre outil pour repérer les enregistrements à risque.
- Étant donné le partitionnement d'un échantillon (p. ex., par géo. et sexe), produire tous les croisements à trois dimensions des variables clés pour obtenir un score de multiplicité pour chaque unité (nombre de tableaux où elle est unique).
- Appliquer la suppression locale aux unités ayant les scores les plus élevés (p. ex., pour chaque unité, supprimer la variable contribuant le plus à la multiplicité).
- Méthode très heuristique – peu d'orientation théorique.

Protection des microdonnées

- Autres techniques disponibles :
  - Sous-échantillonnage
  - k-anonymat (examiné antérieurement)
  - Post-randomisation (PRAM)
  - Microagrégation
  - Utilisation d'un bruit additif ou multiplicatif
  - Permutation de données
  - MASSC
  - Masquage de matrice
  - Microdonnées synthétiques

Sous-échantillonnage

- Offre une protection contre les connaissances d'initié, c'est-à-dire le fait de savoir qu'une personne spécifique se trouve dans le fichier de microdonnées
- Utile quand les données proviennent d'un recensement ou d'un fichier administratif, quand une sous-population est échantillonnée à un taux élevé, ou quand un fichier de microdonnées est produit par appariement de deux fichiers de données, dont l'un est un fichier de recensement ou un fichier administratif
- Empêche aussi l'appariement de fichiers de microdonnées à des données de sortie agrégées provenant d'autres sources pour essayer d'identifier des particuliers ou recouvrer des données supprimées ou dont on a réduit le niveau de détail

Post-randomisation (PRAM)

- Perturbation de microdonnées catégoriques en utilisant un mécanisme probabiliste (Kooiman et coll., 1997)
- Le mécanisme probabiliste est connu, de sorte qu'on peut estimer les caractéristiques des données réelles en partant du fichier perturbé
- La PRAM peut être appliquée indépendamment à différentes variables ou être appliquée simultanément à plusieurs variables

Microagrégation

- Remplacement de données quantitatives pour des groupes de k enregistrements par une valeur moyenne pour le groupe (p. ex. k = 3 ou 5 ou 10)
- Protection offerte si aucune donnée faisant partie du groupe n'est dominante
- Peut être univariée (Defays et Nanopoulos, 1993) ou multivariée (Domingo-Ferrer et Mateo-Sanz, 1999)

Microagrégation

- La microagrégation univariée est facile à effectuer, mais considérée comme une approche médiocre

- On a montré que la microagrégation univariée utilisant les classements des unités individuelles est associée à un risque élevé de divulgation si chaque variable est microagrégée indépendamment – il est facile d'apparier les enregistrements de microdonnées aux enregistrements originaux

- L'utilisation de la microagrégation multivariée est préférable, p. ex., pour des groupes de variables provenant du partitionnement du fichier de données

Ajout d'un bruit à des données continues

- Ajout d'un bruit aléatoire ayant la même structure de corrélation que les données originales afin de préserver essentiellement les corrélations (Kim, 1986)
- Les problèmes comprennent la non-préservation des valeurs nulles, la création de valeurs négatives, les niveaux insuffisants de protection même pour les grandes valeurs de c, et l'effet sur les corrélations au niveau des sous-domaines
- Afin de protéger les enregistrements encore à risque (valeurs aberrantes et unités inhabituelles), les enregistrements originaux et masqués peuvent être appariés en utilisant le programme de couplage d'enregistrements – en faisant une permutation de données quantitatives basée sur le taux de réidentification (Kim et Winkler, 1995)
- Un bruit multiplicatif peut mieux protéger la confidentialité – surtout pour les grandes valeurs

Permutation de données

- Introduite à la fin des années 1970 – utilisée correctement, elle fournit une protection adéquate des fichiers de microdonnées sans altérer les fréquences marginales (Dalenius et Reiss, 1982)
- Facile à mettre en œuvre, élimine la relation entre l'enregistrement et le répondant; peut être appliquée à des variables sensibles seulement et être conçue pour fournir une protection là où elle est le plus nécessaire
- Selon les relations entre les variables (p. ex. contrôles), les calculs peuvent devenir assez difficiles
- La permutation peut aussi fausser les statistiques des sous-domaines, ce qui rend les fichiers de données inappropriés pour l'analyse – la permutation de rangs fondée sur la proximité peut être utile (p. ex. en préservant les relations)

Permutation aléatoire

1. Identifier les candidats pour la permutation – considérer des taux de permutation plus élevés dans les sous-populations à haut risque de divulgation
2. Identifier les partenaires de permutation – semblable à la sélection du donneur pour l'imputation
   - Définir des contrôles de cohérence et veiller à ce que les permutations soient exécutées logiquement
   - Former des cellules de permutation en utilisant des variables clés ou d'autres variables fortement associées aux variables permutées – l'utilisation des poids de sondage pour former les cellules de permutation préservera mieux les distributions
   - Possibilité d'utiliser une fonction de distance (p. ex. norme L2) pour trouver les partenaires – chercher des partenaires ayant des valeurs différentes pour les variables permutées
   - Étendre la recherche si le bassin de donneurs est petit (plus grandes cellules de permutation)

Permutation de rangs

- S'applique aux variables continues/ordinales (de façon indépendante)
- Consiste à classer les répondants, à ajouter un bruit aléatoire à la valeur de rang (p. ex. ±*p*%) et à attribuer les nouvelles valeurs
- Par exemple, 13 enregistrements utilisant une variable classée par ordre alphabétique (p=15)

Id	1	2	3	4	5	6	7	8	9	10	11	12	13
Valeur	Q	W	E	R	T	Y	U	I	O	P	Q	W	E
Rang	6	11	1	8	9	13	10	3	4	5	7	12	2
+bruit	5,6	12,9	0,7	6,1	7,8	14,7	8,9	1,3	2,1	6,0	6,5	12,0	0,5
Nouv. rang	5	12	2	7	9	13	10	3	4	6	8	11	1
Nouv. valeur	P	W	E	Q	T	Y	U	I	O	Q	R	W	E

- Assez bon compromis entre risque de divulgation et perte d'information
- Ne masque pas les valeurs maximales et minimales

MASSC (brevet américain US7058638B2)

- La méthode comprend quatre étapes (Singh et coll., 2003)
- **Microagglomération :** partitionner le jeu de données en strates de risque en se basant sur les variables d'identification (VI) sélectionnées (enregistrements uniques selon les VI de base, uniques selon toutes les VI, doubles, triples, autres)
- **Substitution :** remplacer les valeurs des VI pour des enregistrements sélectionnés aléatoirement par des donneurs sous des contraintes de biais prédéfinies
- **Sous-échantillonnage :** sous des contraintes de variances prédéfinies
- **Calage :** des poids pour les domaines sociaux et démographiques
- Contrôle simultané de la perte d'information et du risque de divulgation
- Peut être appliquée aux microdonnées ou aux résultats agrégés

Données synthétiques

- Jeux de données synthétiques avec imputation multiple (DSIM) créés en remplaçant les valeurs sensibles par des tirages répétés à partir d'un modèle ajusté aux données originales
- L'approche est étroitement liée à l'imputation multiple – si le modèle d'imputation est correct, des inférences valides peuvent être obtenues
- Avantages par rapport aux méthodes de CDS : préserve mieux les relations entre les données; peut être appliquée aux données catégoriques et continues; permet d'imputer les données manquantes avant de générer les DSIM; plus de souplesse pour traiter les contraintes et les enchaînements de questions; l'imputation multiple est plus transparente pour les analystes
- Voir l'ouvrage de Jörg Drechsler (2011)

Données de sortie analytiques

Risques de divulgation liés aux données de sortie analytiques

- Les données de sortie comprennent les statistiques de base, les données de sortie de modèle, les graphiques, etc.
- Certaines données de sortie sont très semblables à des données de sortie tabulaires, par exemple :
  - les histogrammes sont des représentations de tableaux de fréquences
- Certains produits de sortie fournissent des valeurs individuelles, par exemple :
  - valeurs minimales et maximales, intervalles de valeurs, certains quantiles
  - nuages de points, boîtes à moustaches, graphiques des résidus
  - mesures d'influence (p. ex. distance de Cook)

Règles proposées d'après ESSNet SDC

- Règles empiriques globales
  1. 10 unités minimum pour les données de sortie tabulaires ou similaires
  2. 10 degrés de liberté minimum pour toutes les sorties de modèle
  3. Aucune cellule ne peut contenir plus de 90 % des unités de sa ligne ou colonne
  4. La contribution de l'unité la plus grande ne peut pas dépasser 50 % du total d'une cellule
- Les ratios sont sûrs s'ils satisfont aux règles 1, 3 et 4
- Les moments d'ordre plus élevé (variance, asymétrie, aplatissement) sont sûrs si la règle 2 est satisfaite
- Les coefficients de régression sont sûrs si l'un d'entre eux n'est pas diffusé
  - Les résultats complets des régressions linéaires peuvent être diffusés si ≥10 ddl et qu'ils ne sont pas fondés seulement sur des données catégoriques et que la régression ne porte pas sur une seule unité

Mot de la fin

- Le contrôle de la divulgation est complexe et subjectif
- Il n'existe pas de solution universelle
- Il importe de connaître les paramètres spécifiques (type de données, données de sortie, variables)
- Les solutions représentent un compromis entre les besoins de disponibilité des données (la qualité aux analystes) et la protection de la confidentialité
- Il n'existe pas de risque nul, sauf si on ne diffuse aucune donnée
- L'importance de la mise en œuvre de pratiques sûres ne doit jamais être sous-estimée

Protection des chiffres

- **Perception**

  « […] Enfin, la façon dont le public perçoit la vigilance avec laquelle nous protégeons la confidentialité des statistiques est à tout le moins aussi importante que les mesures réelles que nous prenons pour empêcher la divulgation des données des répondants. »*

* ***Statistique Canada : lignes directrices concernant la qualité,*** *Cinquième édition – octobre 2009, page 85.*

Merci !!

Michelle.simard@Canada.ca