

Distr.
GÉNÉRALE

CES/SEM.47/15 (Summary)
30 janvier 2002

FRANÇAIS
Original: ANGLAIS

**COMMISSION DE STATISTIQUE et
COMMISSION ÉCONOMIQUE
POUR L'EUROPE**

**COMMISSION DES COMMUNAUTÉS
EUROPÉENNES**

**CONFÉRENCE DES STATISTICIENS
EUROPÉENS**

EUROSTAT

Séminaire commun CEE-Eurostat sur
les systèmes intégrés d'information statistique
et les questions connexes (ISIS 2002)
(Genève, Suisse, 17-19 avril 2002)

Thème II: Sécurité des communications et confidentialité des données

QUESTIONS DE CONFIDENTIALITÉ RELATIVES AUX SYSTÈMES D'INTERROGATION DE BASES DE DONNÉES STATISTIQUES

Communication sollicitée

du National Center for Health Statistics, États-Unis¹

Résumé

1. Les services nationaux de statistique sont chargés de collecter, de vérifier et d'affiner les données statistiques dans le but de fournir des informations fiables aux décideurs et au public. Aux termes de la loi ou de la réglementation et selon la déontologie, un service national de statistique doit s'acquitter de cette tâche, tout en préservant la *confidentialité* des données se rapportant à des entités individuelles, par exemple des personnes, des entreprises ou des dispensateurs de soins de santé. On notera que ce souci de préserver la confidentialité des données ne s'applique pas aux statistiques relatives aux entités publiques.
2. Avant les années 60, les services nationaux de statistique publiaient les données statistiques essentiellement sous la forme de résultats de calculs ou d'estimations *mis en tableaux*, définis par la classification croisée d'une seule ou de deux variables ou encore d'un petit nombre de variables. Le service national de statistique concerné sélectionnait,

¹ Document établi par Lawrence H. Cox (lcox@cdc.gov).

parmi ces présentations tabulaires, celles qui devaient être publiées, dans un premier temps sur papier, et, ultérieurement, sur support électronique également. La protection de la confidentialité des données, aujourd'hui appelée *contrôle de la divulgation des statistiques*, était assurée en supprimant ou en combinant certaines données mises en tableaux, voire des séries complètes, ou, moins fréquemment, en modifiant légèrement les tableaux par des procédés d'arrondi ou l'ajout de bruit de manière aléatoire. En réalité, le service national de statistique déterminait d'abord quels étaient les tableaux qui méritaient d'être diffusés, puis il publiait, dans chaque cas, une quantité d'informations moindre que celle contenue dans ces tableaux, eu égard aux considérations concernant la confidentialité et la qualité des données.

3. Au cours des années 60, dans le contexte tout d'abord de l'échantillon permanent relatif aux antécédents professionnels, établi par l'Administration de la sécurité sociale des États-Unis, puis des échantillons à l'usage des chercheurs constitués à partir des résultats des recensements décennaux de 1960 aux États-Unis, les services nationaux de statistique ont commencé à publier des *fichiers de microdonnées statistiques* comprenant des informations relatives à des entités individuelles (principalement des personnes). L'utilisateur de données avait désormais la faculté d'effectuer toutes les analyses imaginables à partir des enregistrements unitaires et, ce qui était tout aussi important, d'ajuster les modèles statistiques, démographiques ou économétriques aux microdonnées. Dès lors, le contrôle de la divulgation des statistiques s'est progressivement orienté vers la modification et l'élimination de certains enregistrements de microdonnées. Les données longitudinales présentaient des problèmes de confidentialité qui, pour une large part, n'ont pas pu être résolus à ce jour. Les nouvelles recherches sont axées sur l'ajustement des données à des modèles statistiques complexes et sur le remplacement des publications antérieures par la diffusion de *microdonnées synthétiques* calculées sur la base de modèles et/ou des modèles eux-mêmes. Le contrôle de la divulgation est manifestement complexe dans le cas des présentations tabulaires et des microdonnées, aussi bien en théorie qu'au niveau des calculs à effectuer.

4. Les services nationaux de statistique envisagent désormais de permettre aux utilisateurs d'avoir directement accès aux bases de données statistiques, soit librement, soit suivant des modalités d'accès restreintes, via un *système d'interrogation de bases de données statistiques*. Ce projet renforce l'acuité des problèmes et questions de confidentialité et stimulera vraisemblablement la recherche sur le contrôle de la divulgation des statistiques au cours des prochaines décennies. Dans le présent document, nous étudions au moyen d'exemples certains des problèmes de confidentialité et d'utilité des données soulevés par l'apparition des systèmes d'interrogation des bases de données statistiques. Nous nous concentrons sur deux paradigmes d'interrogation: des informations mises en tableaux issues d'une base de données s'articulant autour d'un grand tableau de contingence à multiples entrées et des modèles statistiques simples établis à partir de la base de données, en l'occurrence des modèles de données spatiales élaborés suivant la méthode de régression par moindres carrés ordinaire et en fonction du meilleur prédicteur linéaire sans biais (*kriging*).
