DGINS 2005/91/2.1/FR

91e CONFÉRENCE DGINS

COPENHAGUE, 26 et 27 MAI 2005

Diffusion de statistiques au bénéfice de la communauté scientifique

Pedro DIAZ MUÑOZ

Directeur «Méthodologies et outils statistiques»
Eurostat, Commission européenne

Diffusion de statistiques au bénéfice de la communauté scientifique

Pedro DÍAZ MUÑOZ

Directeur «Méthodologies et outils statistiques», Eurostat, Commission européenne

La diffusion de données statistiques au bénéfice des chercheurs est soumise à de multiples tensions. Le danger d'identification de certaines personnes, l'énorme quantité d'informations non exploitées totalement et l'image qu'en ont les fournisseurs et les utilisateurs de données sont les facteurs les plus pertinents qui, étant en contradiction, sont porteurs de tensions. Après une analyse de ces facteurs, le présent article examine les mesures prises au niveau de l'UE en vue de faciliter l'accès des chercheurs aux données et suggère une série de mesures pour améliorer la situation.

1. Le contenu informatif des données statistiques

1.1 L'écart entre ce qui est publié et ce qui pourrait l'être

Les données statistiques contiennent bien plus d'informations que les publications ordinaires des instituts de statistique. Envisageons, pour commencer, le cas d'une enquête par sondage. En principe, il est possible de procéder à une tabulation croisée des variables classées en catégories. En d'autres termes, le nombre de cellules possibles contenant des données agrégées est égal au produit de toutes les occurrences possibles des dimensions réparties en catégories. Toutefois, toute publication aussi détaillée doit éviter deux sortes d'écueils. D'une part, l'importance des erreurs d'échantillonnage ne permet pas d'évaluer les indicateurs lorsque les domaines sont trop petits, parce que leur degré d'exactitude sera faible et que les valeurs seront trompeuses. D'autre part, le danger de pouvoir reconnaître certains individus augmente à mesure que le nombre d'unités statistiques dans l'échantillon diminue. Par conséquent, permettre l'utilisation secondaire des microdonnées expose à deux risques différents: 1) à l'identification des personnes et 2) au danger d'inférences erronées, en raison du manque de fiabilité des informations produites.

Le second cas envisagé est celui de données fixées d'une étude exhaustive, comme par exemple un recensement ou un registre. Ici encore, la possibilité de produire des données agrégées est énorme et égale au nombre de cellules de tous les tableaux à double entrée de variables collectées. En théorie, il ne devrait plus y avoir de problème d'exactitude même si d'éventuelles erreurs dans les réponses peuvent avoir un impact, surtout dans les cellules ne comptant que très peu d'individus. En revanche, les dangers d'identification de personnes sont même plus grands que dans l'exemple

précédent, parce que toute combinaison exacte trouvée renvoie avec certitude à un des individus particuliers.

Une autre caractéristique importante des données statistiques vient de leur potentiel de croiser différentes sources d'information, au niveau des microdonnées, par le biais de variables d'identification communes. Ce risque s'ajoute en l'espèce à ceux déjà évoqués. Il dérive de l'hypothèse sous-jacente à l'appariement des données, mais aussi de l'éventuel manque de cohérence entre les sources. À cet égard, il est intéressant de noter que certaines sources de données utilisées pour cet appariement ne sont parfois nullement crées à des fins statistiques. Par conséquent, leur qualité peut ne pas répondre aux exigences d'une étude scientifique.

Les exemples ci-devant montrent que la quantité d'informations susceptibles d'être extraites des données statistiques existantes est bien plus vaste que ce qu'on trouve dans les publications ordinaires des instituts de statistiques, quel que soit le canal de diffusion retenu. Toutefois, et pour autant que toutes les précautions soient prises pour prévenir les risques déjà mentionnés, il relève de la responsabilité des INS de faciliter l'exploitation secondaire des données disponibles pour en optimiser l'utilité.

1.2 Exemples d'utilisations secondaires d'informations statistiques

Heureusement, les données statistiques sont largement exploitées dans de nombreux domaines. Un relevé qui se voudrait exhaustif relève donc de l'utopie. Je me limiterai par conséquent à quelques exemples.

Les comptes rendus du 19e séminaire du CEIES: Solutions innovantes permettant l'accès aux microdonnées [1], Lisbonne (26-27 septembre 2002) fournissent plusieurs exemples d'utilisation de données statistiques à des fins de recherche. Trois d'entre eux méritent une mention spéciale. Premièrement, Richard Blundell: il montre comment l'utilisation de données longitudinales individuelles sur la santé, le vieillissement et la retraite permet l'analyse des effets potentiels du vieillissement de la population ainsi que celle de l'impact économique de l'augmentation des besoins de soins de santé et du recours aux services sociaux, à mesure que le phénomène du vieillissement progresse. Les données utilisées dans cet exemple correspondent à celles de l'étude britannique ELSA (English Longitudinal Survey on Ageing).

Deuxième exemple: Robert Erikson. Il fournit un cas intéressant de référence croisée de trois sources de données permettant d'obtenir une répartition des notes moyennes d'élèves issus de la classe ouvrière arrivés en fin de scolarité obligatoire et la probabilité d'un passage dans l'enseignement secondaire supérieur. Les sources utilisées sont les données fournies par le

recensement de la population de 1990 et qui concernent la profession des parents, les notes moyennes provenant d'un ensemble de données d'élèves quittant l'école et des informations portant sur l'enseignement secondaire supérieur tirées d'un fichier d'admissions à ce niveau de formation.

Troisième exemple: Niels Westergaard-Nielsen. Celui-ci insiste sur l'utilisation de couples de données employeurs-salariés (EE) qui permettent d'étudier des différences entre des entreprises ainsi que les réactions face aux interventions de diverses politiques. Les données liées fournissent des réponses aux questions que se posent les responsables politiques sur les endroits où des emplois sont soit créés soit détruits, et sur la manière dont les politiques menées ont un impact sur les entreprises et la demande de main-d'œuvre. Ce document donne une description très intéressante des exigences relatives aux sources de données afin que l'appariement EE soit possible.

Le tableau ci-après présente une synthèse des projets des instituts de recherche qui ont soumis à EUROSTAT des demandes de microdonnées ECHP (panel communautaire des ménages) en 2004.

Contrats de recherche utilisant des données *ECHP* – année 2004 – principaux thèmes	
Études portant sur des sous-populations spécifiques	*Études portant sur des phénomènes spécifiques*
• Personnes âgées • Personnes pauvres • Régions • Chômeurs de longue durée • Femmes mariées • Participation des femmes au marché du travail • Divorcé(e)s • Travailleurs à durée déterminée / intérimaires • Personnes en fin de carrière professionnelle • Jeunes	• Mobilité • Inégalités de revenus • Emplois de transition <-> chômage • Taxation, subventions / aides • Transferts intrafamiliaux • Inégalités de revenus et éducation • Évolution salariale • Éducation et santé • Participation au marché du travail et fécondité • Garde des enfants • Discrimination

EUROMOD constitue un exemple très intéressant de l'utilisation de microdonnées. Il s'agit d'un modèle d'imposition/d'indemnisation fondé sur des microdonnées relatives aux ménages. Il calcule l'impact des changements intervenant dans les politiques sociales et fiscales sur la mesure des revenus personnels, et sur la prospérité et le bien-être des ménages. Le modèle a permis d'évaluer des politiques européennes et nationales.

Pour plus d'informations, cf.: http://www.econ.cam.ac.uk/dae/mu/emod.htm.

Les exemples ci-dessus ne devraient cependant pas donner l'impression que les chercheurs s'intéressent exclusivement aux microdonnées. Une catégorie importante de chercheurs est surtout intéressée par l'analyse de l'évolution économique. Ils tentent d'identifier des relations entre des macroagrégats. Leur travail requiert des quantités énormes de séries historiques sous forme brute, c.-à-d. non traitées par exemple au niveau des corrections ou des variations saisonnières ou des jours ouvrables, ce qui risquerait en effet de perturber le signal de base que l'on cherche à isoler.

1.3 Utilisation abusive de données statistiques

Heureusement, il n'existe que peu de cas d'utilisations abusives de données statistiques. En ce qui concerne les violations de confidentialité, il se pourrait que les mesures protectrices adoptées par les instituts statistiques se révèlent suffisantes. Mais c'est peut-être aussi parce que les informations statistiques ne permettent guère de procéder à des identifications à une échelle suffisante pour que l'opération devienne rentable. Il paraît toutefois utile de s'attarder sur certains abus fictifs de données statistiques. En effet, leur détournement est susceptible de porter gravement atteinte à la confiance que la société a dans les systèmes statistiques. C'est pourquoi tout doit être mis en œuvre pour prévenir les abus. C'est l'objet des paragraphes suivants. À noter que les scénarios évoqués vont au-delà de la violation de confidentialité et envisagent d'autres abus en matière d'information trompeuse et de traitement incorrect des données.

- Des listes de personnes bénéficiant de revenus élevés peuvent être extraites des registres de personnes comportant des informations économiques, par exemple des registres fiscaux. En outre, ces derniers pourraient être associés à d'autres ensembles de données comme des échantillons de recensements ou des enquêtes sur le budget (des ménages).

- Des attaques de journalistes contre des personnalités en vue (en vertu du principe du droit à l'information). Exemples: en Finlande, certains journaux prennent la peine de traiter des données relatives aux ressortissants les plus riches (et de les publier ensuite sur la toile) ; il s'agit généralement des mille personnes bénéficiant des revenus les plus élevés pour lesquelles des chiffres détaillent les revenus annuels, le revenu du capital et les avoirs patrimoniaux. D'autres journaux à diffusion locale [1] ont publié l'intégralité des données fiscales disponibles pour tous les ressortissants habitant dans leur zone de chalandise. Ces mêmes variables de revenus sont également utilisées pour les données finlandaises EU-SILC (statistiques communautaires sur le revenu et les conditions de vie).

1 Ces journaux ont été cités en justice et mis en examen pour violation de la loi sur la protection de la vie privée!

- Le scénario des «voisins fouineurs»: le but est d'identifier son voisin dans un microfichier, sur la base de quelques variables clés seulement (type d'habitation, nombre de personnes, sexe, âge, occupation, etc.).

- Le croisement de microdonnées ou d'informations économiques détaillées avec des registres de personnes, en vue de dresser des listes à des fins de marketing.

- La recherche d'informations sur des ratios économiques portant sur des entreprises en vue d'identifier certains secteurs, permettant ensuite une analyse spécifique de sociétés individuelles, à des fins administratives.

- Des interprétations erronées de résultats statistiques par des personnes ignorant les principes et règles statistiques de base (par exemple, confondre relation causale avec simple fausse corrélation).

- L'incohérence entre les statistiques dérivées de microdonnées perturbées publiées et les agrégats dans les statistiques officielles.

- Production d'agrégats de données sur la base d'observations trop peu nombreuses, ce qui leur enlève tout sens.

1.4 Protection de la confidentialité

La littérature scientifique abonde d'articles décrivant des méthodes de protection des données et analysant le risque de divulgation statistique. Je fais ici référence à la session de travail sur la confidentialité des données statistiques organisée à Luxembourg les 7-9 avril 2003 [2], lors de laquelle la plupart de ces méthodes ont été présentées. À noter également le projet CASC du cinquième programme-cadre de recherche, qui a été récemment achevé [3]. Le tableau ci-après donne un aperçu schématique des différentes méthodes utilisées à l'heure actuelle.

Méthodes de protection de la confidentialité des fichiers de microdonnées	
• Recodage global	Il consiste à fusionner plusieurs catégories d'une variable en une seule.
• Suppression locale	Une ou plusieurs valeurs d'une combinaison peu sûre sont remplacées par une valeur manquante.
• Regroupement des valeurs extrêmes supérieures et inférieures	Recodage global dans le cas de catégories ordinales.
• Randomisation a posteriori	Classifications perturbées délibérément par la modification des valeurs pour une ou plusieurs catégories de variables.
• Micro-agrégation	Remplacement des valeurs quantitatives individuelles par d'autres calculées sur de petits agrégats.
• Agrégation aléatoire avec bruit	Ajout d'un bruit aléatoire aux données quantitatives.
• Permutation des données	Permutation de certaines variables entre deux registres présentant des catégories communes pour certaines variables prédéfinies.

2. Fourniture de données aux chercheurs

2.1 Les compromis

Le présent chapitre résume ce qui a déjà été dit. Chaque fois que des données sont communiquées aux chercheurs, plusieurs compromis sont à envisager:

(1) Le risque de reconnaissance d'informations et leur sensibilité. On peut distinguer deux types de dangers:

- le scénario des «voisins fouineurs» mentionné ci-dessus: à partir de la connaissance de quelques caractéristiques d'une unité déclarante, tentative d'identification et divulgation de l'information.

- Attaque d'un registre: une banque de données commerciale est accessible sous forme d'un registre. Celui-ci permet d'identifier l'unité déclarante (nom, adresse) grâce à quelques variables clés, par exemple le sexe, l'activité, la région, etc. Le croisement des données du registre et du microfichier pourrait permettre la reconnaissance d'un nombre significatif de fournisseurs de données statistiques.

Le deuxième type de danger représente, aujourd'hui, une source d'inquiétude majeure, étant donné les développements technologiques actuels qui permettent de constituer des bases de données parallèles à des fins non statistiques.

La protection contre les violations de confidentialité est garantie par des lois qui contiennent des dispositions régissant le contrôle de la divulgation. Toutefois, ces lois n'énumèrent que des règles générales. Leur interprétation sur le terrain varie énormément de pays à pays et débouche sur de grandes différences quant aux niveaux de protection.

(2) La capacité d'analyse des informations. Comme il est dit au point 1.1 ci-dessus, les instituts statistiques sont parfaitement conscients qu'ils ne publient qu'une petite fraction de l'ensemble des informations contenues dans les données. Ils ont par conséquent la grande responsabilité de faciliter l'exploitation secondaire de ces précieux renseignements en vue de répondre à certains besoins spécifiques.

(3) La perception de la vie privée dans la société. Au-delà du danger réel de reconnaissance, il y a le risque tel que perçu par la société. Même si les dispositions législatives régissant le contrôle de la divulgation sont suffisantes, voire même excessives, la perception du risque par les unités déclarantes – face aux abus potentiels des données qu'elles rapportent – est très élevée. D'aucuns pourraient même aller jusqu'à altérer la qualité des réponses fournies pour essayer de dissimuler ou de fausser certaines caractéristiques qu'ils ne souhaitent pas voir identifiées.

(4) L'intérêt porté aux informations par la société. Dans de nombreux cas, les analyses statistiques détaillées se rapportant à de petites populations ou à des régions/domaines limités revêtent un très grand intérêt pour la prise de décisions politiques ce qui, comme c'est souvent le cas, entraîne certaines contradictions: alors que les individus sont très préoccupés par le caractère privé de leurs propres réponses, ils se plaignent par ailleurs que ces informations ne soient pas suffisamment exploitées dans le but d'identifier des besoins sociétaux, de corriger des déséquilibres dans la répartition des richesses ou d'améliorer l'organisation des services publics.

2.2 L'accès aux microdonnées à des fins scientifiques dans l'Union européenne Le règlement 831/2002

Afin de répondre aux besoins des chercheurs dans l'Union européenne (UE), deux instruments ont été mis en place dans le contexte des textes de base sur la confidentialité des données (les règlements 1588/90 et 322/97). Il s'agit, d'une part, du comité du secret statistique qui a compétence d'exécution pour tout ce qui a trait à la confidentialité et du règlement (CE) n° 831/2002 de la Commission en ce qui concerne l'accès aux données confidentielles à des fins

scientifiques. Le lecteur trouvera une description et une analyse détaillées de ce texte réglementaire dans l'article de John King et de Jean Louis Mercy, présenté lors de la session de travail conjointe CEE-NU/Eurostat sur la confidentialité des données statistiques, organisée à Luxembourg les 7-9 avril 2003. Alors que ce règlement a suscité de grands espoirs pour la mise à disposition de microdonnées au bénéfice de la communauté scientifique, sa mise en œuvre s'est heurtée à plusieurs difficultés, si bien que les progrès initiaux sont plutôt lents.

En décembre 2004, le comité du secret statistique a procédé à l'analyse des progrès de mise en œuvre du règlement: il a décidé la création de procédures rapides pour le traitement des demandes des chercheurs et pour l'éligibilité des institutions de recherche. Aujourd'hui, des microdonnées ne peuvent être fournies aux chercheurs que pour deux domaines statistiques. Il s'agit du panel communautaire des ménages (ECHP) et, depuis peu, des enquêtes sur les forces de travail (EFT). En outre, le groupe de travail de l'enquête communautaire sur l'innovation débat actuellement des critères de distribution des fichiers de microdonnées de cette enquête. Enfin, un groupe de travail a été mis en place pour se livrer au même exercice à propos de la prochaine enquête EU-SILC (statistiques communautaires sur le revenu et les conditions de vie). Parallèlement, des ressources ont été prévues pour réduire l'arriéré de traitement des demandes et assurer un suivi des processus dans le but d'améliorer la situation à court terme.

3. La réflexion internationale. La Task force NU/UE Confidentialité et microdonnées

Au mois de juin 2003, la Conférence des statisticiens européens crée une Task force sur la confidentialité et les microdonnées, présidée par Dennis Trewin du Bureau australien des statistiques (ABS). L'objectif de ce groupe est d'élaborer des principes communs régissant l'accès aux microdonnées, ainsi que de présenter des études de cas illustrant de bonnes pratiques conformes à ces principes.

Au mois de mai 2004, la Task force a produit un document de discussion [4] portant sur les perspectives des INS et celles des chercheurs, pour étudier comment les tensions entre ces deux approches pouvaient être aplanies. Différentes méthodes ont été débattues, la réflexion portant par exemple sur des séries de microdonnées anonymisées, des centres d'accès à distance et des laboratoires de données qui pourraient servir à cette fin. Un ensemble de principes et certains points de discussion ont été proposés.

Au mois d'octobre 2004, la Task force a rassemblé les commentaires reçus des différents pays dans un document [5] en y ajoutant un résumé des remarques transmises. Il inclut également une liste de

points faisant l'objet d'une large convergence des vues et ceux pour lesquels un grand nombre d'avis avait été exprimé.

La Task Force a l'intention d'achever son mandat en présentant au CES un document contenant des lignes directrices en matière de bonnes pratiques lors de la réunion plénière du mois de juin 2005.

4. Perspectives

Plusieurs pistes parallèles devraient permettre de rendre les données statistiques plus utiles aux chercheurs, tout en offrant des garanties suffisantes de non divulgation. En voici quelques-unes:

(1) Harmonisation des critères relatifs au danger de divulgation. De manière générale, les législations nationales et européenne sont relativement harmonisées en ce qui concerne lesdites données confidentielles. Toutefois, quand il s'agit d'appliquer ces législations sur le terrain, on observe d'importantes différences dans les pratiques nationales. Parfois certains critères bénéficient d'un lourd poids historique; d'autres fois la base scientifique ne semble guère solide; et dans la plupart des cas, on a recours à des solutions conservatrices parce qu'on ne maîtrise pas suffisamment le risque réel.

Ces différences d'interprétation résultent du fait qu'il n'existe aucune approche harmonisée du phénomène de risque de divulgation. Pour se mettre d'accord sur ce risque, il convient en effet de commencer par dégager un consensus sur la sensibilité des données (à quel point les variables du fichier sont-elles «privées»?), ainsi que sur la possibilité d'apparenter ces données à d'autres sources externes (présence ou non de variables clés et de variables d'identification). Ensuite, il faut encore trouver une méthode harmonisée pour mesurer le risque. Des travaux méthodologiques seront donc requis pour concilier les différentes approches ou pour justifier une préférence pour l'une d'entre elles.

La nécessité de dégager des critères de base communs est évidente. Tout en garantissant un niveau d'harmonisation satisfaisant, ils devront permettre un degré de souplesse suffisant pour les adapter à la perception spécifique des collectivités nationales. Cette flexibilité combinée avec une base internationale solide et convenue aurait également pour avantage de mieux justifier certains choix nationaux pour la publication de microdonnées.

(2) L'éligibilité des chercheurs et des projets de recherche. Aujourd'hui, les critères d'éligibilité, tant pour les projets de recherche qu'au niveau des chercheurs eux-mêmes, ne sont pas clairs. Il va dès lors de soi qu'ils ne sont nullement homogènes dans l'Union européenne.

Lorsqu'il s'agit d'évaluer un chercheur, on a souvent tendance à évaluer d'abord l'institut de recherche auquel il appartient pour ensuite essayer d'établir un lien solide entre la personne en question et l'organe de recherche responsable d'elle. Une évaluation a priori constitue une étape qui mérite une attention minutieuse, mais une analyse ex post basée sur des procédures standardisées de suivi des instituts, ainsi que la création d'archives individualisées par institution concernée, semblent plus prometteuses du point de vue administratif et qualitatif. La possibilité de dresser des listes «noires» reprenant tous ceux qui ont mal utilisé ou abusé des données communiquées devrait maintenir une pression salutaire sur les instituts de recherche.

Pour évaluer un projet de recherche, il est essentiel d'impliquer les unités techniques qui pourraient déjà être en contact avec les chercheurs et leurs travaux qui ont déjà pu mener des études sur le sujet ou qui pourraient avoir un intérêt direct pour le sujet traité. Il est à noter que dans certains pays, par exemple au Royaume-Uni, l'intérêt propre de l'institut statistique intervient souvent. D'autres paramètres, comme le caractère original de la recherche, la nécessité réelle d'avoir accès aux données confidentielles ou l'absence de conflits d'intérêt entre les INS et le chercheur, sont également à prendre en considération.

Certains critères d'éligibilité ont été élaborés dans le cadre du règlement 831/2002 et les questionnaires d'évaluation correspondants existent. Ils pourraient constituer la base d'un ensemble transparent de critères transparents et garantir ainsi l'égalité de traitement de toutes les communautés scientifiques, partout dans l'Union.

(3) Les dispositions légales en cas d'abus. Un aspect important de la protection des données consiste à rendre l'utilisateur conscient de sa responsabilité juridique propre et des actions en justice qu'il encourt effectivement en cas d'abus. Lorsque l'exploitation des données est internationale, il convient de préciser les responsabilités de chacun et de les communiquer de manière précise.

(4) Le rôle du CENEX et le contrôle de la divulgation statistique. La Task force sur les centres d'excellence instaurés par le CPS a proposé le lancement, en 2005, d'un projet pilote sur le concept des centres d'excellence (CENEX). En résumé, le but est de mettre en place une équipe d'organisations statistiques nationales capables de fournir une expertise sur un domaine spécifique et de développer des outils ou des savoirs au service non seulement des membres de l'équipe, mais à disposition de l'ensemble des acteurs SSE. Le contrôle de la divulgation statistique est l'un des deux thèmes qui seront inclus dans la phase pilote du CENEX. Eurostat prépare actuellement la documentation nécessaire au lancement du projet. Les résultats escomptés à la fin de la phase pilote, dont la durée prévue est d'un an, devraient être un inventaire des besoins, le développement d'outils

informatisés pour y répondre et un manuel des pratiques habituelles en matière de contrôle des divulgations.

(5) Le code de bonne pratique récemment adopté par le comité du programme statistique inclut quelques dispositions sur l'utilisation de données statistiques par les chercheurs et sur la protection de la confidentialité. Ce code de bonne pratique fournira le cadre permettant d'harmoniser l'activité dans ce domaine.

(6) L'accès à distance pour les chercheurs. Il existe un large consensus parmi les pays pour considérer que cette approche est très prometteuse. Les expériences actuelles sont cependant plutôt isolées. Deux types d'accès à distance sont à envisager: premièrement, l'exécution à distance, où le chercheur soumet son programme et reçoit ensuite un courriel avec les résultats ; deuxièmement, l'accès à distance à proprement parler, où le chercher procède lui-même à l'analyse et obtient immédiatement les réponses sur son écran. Au mois d'octobre 2004, Eurostat a présenté son analyse de trois expériences d'accès à distance en cours devant le comité directeur TI. Deux d'entre elles, celle du Centre national US de recherches sur la santé (Center of Health Research) et l'Étude sur les revenus au Luxembourg concernent des exécutions à distance, tandis qu'un troisième projet, celui du Denmark Statistics constitue un exemple d'accès à distance.

(7) Fichiers accessibles au grand public. On pose souvent la question de savoir s'il est possible de protéger la confidentialité d'un fichier de manière à ce que le risque de violation puisse être considéré comme suffisamment faible pour permettre l'accès public aux données anonymisées. Dans l'Union européenne, on trouve des exemples de tels fichiers publics dans certains pays comme l'Autriche, la France, l'Italie et le Royaume-Uni. Certes, ces fichiers publics ne remplacent pas les fichiers confidentiels, parce que ces derniers offrent bien plus de possibilités d'analyse. En revanche, le coût de production des fichiers accessibles au grand public est minime et leur distribution aux utilisateurs peut être immédiate.

Au mois de décembre 2004, Eurostat a soumis au comité du secret statistique sa stratégie de création de fichiers accessibles au grand public ainsi qu'une proposition spécifique relative aux enquêtes sur les forces de travail (EFT). Toutefois, cette proposition n'a pas obtenu un soutien suffisant de la part du comité.

5. Conclusions

(1) L'écart entre les informations contenues dans les données statistiques et le contenu des publications statistiques officielles est très grand. Une manière de combler ce fossé est de fournir des fichiers de microdonnées aux chercheurs.

(2) De nombreux risques doivent cependant être maîtrisés. Ils concernent la protection juridique contre la reconnaissance de personnes individuelles, la lutte contre les utilisations abusives ou illicites des données et aussi la perception qu'ont les individus des manipulations abusives de leurs propres informations. Il importe de bien gérer ces risques.

(3) L'objectif vise par conséquent à combler ce fossé pour autant que les risques soient bien maîtrisés. Des mesures juridiques et techniques peuvent être explorées à cette fin. Les mesures juridiques concernent l'éligibilité des chercheurs et des projets de recherche. Les mesures techniques font référence aux différentes méthodes de protection de la confidentialité.

(4) Des réflexions internationales montrent que bien qu'il existe un large consensus en faveur de la fourniture de microdonnées aux chercheurs, les avis sont très différents dès que l'on aborde la plupart des aspects plus détaillés. Plus spécifiquement, les critères pour décider si un fichier est suffisamment sûr et peut être diffusé varient énormément.

(5) Le règlement 831/2002 de la Commission constitue un cadre juridique utile pour la fourniture de microdonnées aux chercheurs. Après des débuts de mise en œuvre difficiles, la procédure et les délais prescrits par ce règlement devraient être réexaminés de manière à dégager des suggestions d'amélioration.

(6) Plusieurs voies peuvent être explorées pour améliorer la diffusion des microdonnées à la communauté scientifique: premièrement, l'élaboration de critères harmonisés pour anonymiser et réglementer l'éligibilité des chercheurs et des projets, deuxièmement, un cadre légal contre les abus, troisièmement, la mise en œuvre d'un code de bonne pratique, quatrièmement, l'étude des possibilités d'accès à distance aux données et enfin, la création de fichiers accessibles au grand public.

Bibliographie:

[1] (2003), 19e séminaire du CEIES: Solutions innovantes permettant l'accès aux microdonnées, Lisbonne, 26 et 27 septembre 2002, Luxembourg: Office des Publications officielles des Communautés européennes.

[2] (2004), Monographies sur les statistiques officielles : session de travail conjointe CEE-NU/Eurostat sur la confidentialité des données statistiques, Luxembourg, 7-9 avril 2003, Luxembourg: Office des Publications officielles des Communautés européennes.

[3] Domingo-Ferrer, J. & Torra, V. (2004), Privacy in Statistical Databases – Projet CASC, conférence de clôture, PSD 2004, Barcelone, Catalogne, Espagne, juin 2004, Actes, Éditions Springer-Verlag.

[4] Trewin, D. (2004), groupe de travail CES Confidentialité et microdonnées – document de discussion, 52e session plénière, Paris, 8-10 juin 2004, NU, Conseil économique et social,

Commission de statistique et Commission économique pour l'Europe, Conférence des statisticiens européens.

[5] Trewin, D. (2004), groupe de travail CES Confidentialité et microdonnées – première rencontre du Bureau 2004/2005, Washington, DC, 18-19 octobre 2004, Commission de statistique et Commission économique pour l'Europe, Conférence des statisticiens européens, point 10.