NATIONS UNIES **E**

Conseil économique et social

Distr.
GÉNÉRALE

ECE/CES/2007/22
2 avril 2007

FRANÇAIS
Original: ANGLAIS

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

CONFÉRENCE DES STATISTICIENS EUROPÉENS

Cinquante-cinquième réunion plénière
Genève, 11-13 juin 2007
Point … de l'ordre du jour provisoire

SÉMINAIRE SUR LE RENFORCEMENT DE L'EFFICACITÉ ET DE LA PRODUCTIVITÉ DES SERVICES DE STATISTIQUE

PREMIÈRE PARTIE

Arguments en faveur et à l'encontre de l'utilisation de fichiers statistiques par les bureaux de statistique

Communication d'Israël[1]

INTRODUCTION

1. Les technologies modernes et l'utilisation généralisée de codes d'identification dans les différentes bases de données ont créé une source de données volumineuse et relativement bon marché à la disposition des services nationaux de statistique. La liste des fichiers potentiellement utilisables comprend un registre de la population et des enregistrements du produit des activités des secteurs public et privé, par exemple le revenu et la taxe à la valeur ajoutée, les prestations sociales et la sécurité sociale, les études scolaires et universitaires, la santé, les municipalités et la police, sans compter les fichiers administratifs traditionnels dans lesquels figurent par exemple les chiffres des importations et des exportations. L'augmentation de la capacité de stockage des données et l'amélioration des méthodes de couplage des fichiers ont ouvert de nouvelles perspectives, comme en témoignent par exemple les fichiers appariés employeurs-salariés qui permettent d'analyser les résultats des entreprises concurremment avec les caractéristiques des salariés (voir Haltiwanger, Lane et Spletzer (2000) et Hamermesh (2007)). Les futures sources potentielles, qui découlent de l'utilisation des techniques modernes, comprennent les fichiers

[1] Communication établie sur l'invitation du secrétariat.

administratifs des compagnies de téléphone cellulaire qui peuvent servir à suivre l'origine, la destination et la durée des appels téléphoniques, la navigation sur le Net, la consommation d'eau et d'électricité, les déplacements de voitures au moyen d'un SIG et les droits de péage acquittés, etc.

2. L'utilisation potentielle des fichiers administratifs dépend de la législation, du progrès technique et de la tradition du pays. L'exemple exposé dans la présente communication s'appuie principalement sur l'environnement du pays que nous connaissons le mieux, Israël. Afin de l'adapter à un large éventail de pays, il se limitera à un fichier administratif bien particulier qui existe dans de nombreux pays et comprend les données sur les salaires et l'emploi communiquées par les employeurs aux autorités fiscales. Ces fichiers existent dans la quasi-totalité des pays qui prélèvent à la source un impôt sur les revenus salariaux.
Une comparaison des chiffres de la masse salariale totale calculée à partir de différentes sources montrera la confusion qui peut se produire lorsque l'on utilise simultanément des données tirées de fichiers administratifs et des données d'enquête. On constatera que, si l'on n'est pas suffisamment prudent, les différences peuvent atteindre jusqu'à 20 % selon celle des deux sources utilisées pour l'estimation de la masse salariale. De toute évidence, si on limite la comparaison à des sous-populations, la différence peut être encore plus grande. Pour autant qu'on le sache, ces conclusions valent pour d'autres pays également.

3. La présente communication a pour but de mettre en lumière les propriétés de l'utilisation potentielle des données administratives et de proposer un certain nombre de mesures destinées à remédier à leurs inconvénients.

II. AVANTAGES DE L'UTILISATION DES FICHIERS ADMINISTRATIFS

4. Les fichiers administratifs constituent une source potentielle de données nombreuses et relativement bon marché qui peut être considérée comme une mine d'or pour les bureaux de statistique ainsi que pour quiconque souhaite faire des recherches et se documenter sur le comportement social.

5. En général, les fichiers administratifs ont de nombreux points communs avec un recensement, c'est-à-dire qu'ils s'appliquent à la totalité de la population dont ils rendent compte. (Ils n'englobent pas nécessairement la totalité de la population visée.) De ce fait, les bureaux de statistique peuvent élargir leur action pour couvrir de petites populations. De surcroît, une conception minutieuse des échantillons dans le but de réduire les coûts devient moins indispensable, ce qui réduit la nécessité de disposer d'un personnel qualifié difficile à trouver pour la conception des échantillons, l'évaluation des erreurs types, etc.

6. Les fichiers administratifs peuvent remplacer les données d'enquête et, ce faisant, réduisent la charge de travail des citoyens qui doivent fournir des données et dispensent les services de statistique de la collecte directe des données. Les fichiers administratifs peuvent parfois venir s'ajouter aux données réunies par la voie traditionnelle des enquêtes. Qu'ils les remplacent ou viennent en complément est une question à déterminer pour chaque domaine, chaque fichier et chaque pays.

III. INCONVÉNIENTS DE L'UTILISATION DES FICHIERS ADMINISTRATIFS

7. L'utilisation généralisée des fichiers administratifs peut représenter un danger pour la démocratie car elle soulève de graves questions en rapport avec la protection de la vie privée et des droits de l'homme. Les données sont produites à l'occasion d'autres activités, dont certaines sont imposées au citoyen par les pouvoirs publics, sans qu'il lui soit demandé s'il consent à l'utilisation ou au stockage des données à des fins statistiques. En résumé, la source de données bon marché engendre deux types de risques. D'une part, il y a danger de créer les conditions voulues pour l'instauration d'un monde du type «big brother» d'Orwell, dominé par les bureaux de statistique et dans lequel il sera possible de suivre à la trace l'activité de chaque individu faisant partie de la société. D'autre part, la simple existence d'une source de données bon marché risque d'inciter les bureaux de statistique à réunir des données qui ne sont pas réellement nécessaires parce que la valeur ajoutée de ces sources est parfois modeste. Par exemple, de nombreux organismes corrigent ou établissent les adresses des citoyens à partir du registre de la population. Réunir des adresses en s'adressant à ces organismes n'apporte pas de «nouvelles» informations. Il s'ensuit un gaspillage des ressources publiques et il devient nécessaire de procéder à une analyse coûts-avantages de l'utilisation de chaque source de données.

8. Les fichiers administratifs présentent parfois peu d'intérêt pour un bureau de statistique car ils sont généralement orientés en fonction des priorités et des politiques de l'organisme responsable. De plus, les définitions et le contenu peuvent être parfois modifiés sans avis préalable et sans délai de grâce pendant lequel les données sont communiquées simultanément selon les nouvelles et les anciennes définitions. De ce fait, il n'y a pas de lien entre les données recueillies avant et après les modifications et il est impossible d'évaluer l'effet de la modification qui est véritablement intervenue, ce qui est particulièrement problématique en cas d'utilisation d'un fichier constamment mis à jour, par exemple un registre. Il convient de souligner que les différences entre les objectifs poursuivis par l'organisme qui produit des données et par les bureaux de statistique occasionnent ces types de problèmes. L'organisme s'intéresse à l'utilisation du fichier dans le présent, mais le bureau de statistique s'intéresse également aux données antérieures contenues dans le fichier. Ce conflit d'intérêts est l'une des causes de ces problèmes.

9. En un sens, chacun reçoit ce pour quoi il a payé. Nous partons du principe que les fichiers administratifs ne comprennent que des variables dignes d'intérêt dont l'organisme de collecte a besoin pour exercer son activité. Le plus souvent, les bureaux de statistique sont considérés comme des «utilisateurs résiduels» des fichiers administratifs et leurs besoins ne sont pas pris en compte au moment de la phase conceptuelle, d'où le manque de souplesse nécessaire pour adapter une série de données en fonction de ce qui les intéresse. En un sens, l'utilisation de données administratives relève d'une décision du type «à prendre ou à laisser».

10. Comme les données administratives consistent en éléments d'information sur des transactions effectives, on ne peut s'attendre à ce qu'elles expriment des réponses à des questions suggestives et hypothétiques qui peuvent présenter un intérêt pour le bureau de statistique. Or, il est important d'obtenir des réponses à des questions commençant par «et si» chaque fois que l'on veut modéliser la réaction de la population à des changements proposés.

IV. AUTRES PROPRIÉTÉS

11. Certaines propriétés peuvent être considérées positivement ou négativement selon les avantages et les coûts relatifs qui leur sont associés. L'avantage des fichiers administratifs en matière de couverture tient au fait qu'ils incorporent l'ensemble de la population ou tout au moins l'ensemble de la population correspondant à une classification donnée, alors que la plupart des échantillons utilisés par les bureaux de statistique portent tout au plus sur 1 % de la population. En revanche, cette même propriété accroît le risque qu'un individu souhaitant obtenir des données sur une personne ou une entité précise tentera de pénétrer illégalement dans le système d'un bureau de statistique pour avoir accès à ces données. C'est pourquoi des mesures de sécurité spéciales doivent être prises pour protéger le système contre les pirates et contre les membres du personnel qui ne sont pas autorisés à y pénétrer.

12. Jusqu'à présent, le service national de statistique doit généralement attendre, pour disposer des fichiers administratifs, que l'organisme de collecte ait terminé sa tâche. En d'autres termes, il doit attendre un ou deux ans. Cette question de temps peut donc dissuader le remplacement des données d'enquête par des fichiers administratifs. Cette question pourrait cependant perdre de son importance en raison du développement de la collecte en ligne par l'Internet des données administratives.

13. **Conclusion**: Si l'on compare l'ensemble des coûts et celui des avantages, les progrès en matière de collecte et de transmission des données amènent inévitablement à conclure qu'il faut utiliser les données administratives. Cela dit, il est indéniable également qu'elles ne remplacent pas toutes les données dont une société moderne a besoin, en particulier parce qu'on ne peut en attendre qu'elles apportent des réponses aux questions subjectives et hypothétiques indispensables pour évaluer la réaction d'une population face à diverses mesures des pouvoirs publics. Elles doivent donc coexister avec les données d'enquête. L'existence de deux sources de données pose un problème de comparabilité, dont il sera question dans le reste de la présente communication.

V. INTÉGRATION DES DONNÉES ADMINISTRATIVES ET DES DONNÉES D'ENQUÊTE

14. L'opinion la plus courante s'agissant de la comparaison entre les données administratives et les données d'enquête est exposée dans un excellent document de travail de Kapteyn et YPMA (2005). Ils font valoir avec énergie que l'on a généralement tendance, chaque fois que l'on compare les données d'enquête et les fichiers administratifs, à considérer ces derniers comme exprimant la «vérité» (Kapteyn et YPMA, 2005; Huynh *et al.*, 2002; Hotz et Schols, 2004). Cette conviction peut être attribuée au fait que les données administratives sont utilisées aux fins de l'exécution des tâches de l'administration qui les réunit et font donc l'objet d'un examen approfondi pour en déceler les erreurs. Toutefois, Kapteyn et YPMA font également valoir que cette supposition ne doit pas être interprétée comme la vérité ultime. Les erreurs systématiques commises par l'organisme qui réunit les données peuvent influer sur la qualité de ces données. Par exemple, on peut s'attendre à ce qu'une administration fiscale prête peu d'attention aux sources de revenus non imposables, même si elles doivent être consignées.

15. L'utilisation simultanée de fichiers administratifs associés à des données d'enquête peut prêter à confusion parce qu'ils font appel à des sources différentes qui peuvent être biaisées.

Les définitions, l'horizon temporel et la période considérée sont différents, et comme la source des informations de même que la collecte et le traitement des données sont différents, les erreurs relèvent de catégories différentes.

16. Les fichiers administratifs sont établis par un organisme (public ou privé) pour son propre usage. En tant que tel, l'organisme est le seul à décider des informations à réunir, du mode de collecte, etc. De ce fait, la collecte de données peut s'arrêter brusquement, en raison d'un changement intervenu dans la législation, ou bien dans les priorités de l'organisme qui réunit les données. De plus, cet organisme n'est pas obligé d'indiquer les changements à l'avance au service national de statistique, ou bien de prévoir une période pendant laquelle les deux ensembles de données seront réunis simultanément. Il peut en découler plusieurs conséquences.

17. Faute d'un accord conclu entre l'organisme et le bureau de statistique, les fichiers administratifs peuvent présenter moins d'intérêt lorsqu'il s'agit d'établir des séries chronologiques. Le bureau de statistique doit être informé des modifications des priorités ou de la législation qui peuvent influer sur la nature des fichiers qui lui sont fournis. Un exemple en est fourni par l'Institut des assurances nationales en Israël qui soumet ses relevés concernant les personnes qui occupent un emploi dont le salaire correspondant au poste est inférieur de 50 % au salaire moyen. Au début de 2005, la loi qui autorise l'application d'un taux d'imposition réduit sur les revenus inférieurs de 50 % au revenu moyen a été modifiée afin qu'un taux réduit soit appliqué aux revenus inférieurs de 60 % au salaire médian. Le Bureau central israélien de la statistique, qui n'était pas au courant de cette modification, a continué de publier les données sans en modifier les définitions, ce qui a amené à se poser un certain nombre de questions concernant l'apparition d'une nouvelle «tendance» dans l'économie, à savoir une augmentation des postes peu rémunérés. Plusieurs mois se sont écoulés avant que l'on ne découvre la raison de ce changement. Cela dit, même lorsque l'on s'en est rendu compte, il n'y avait aucun moyen d'empêcher une «discontinuité» dans la série.

VI. UNE ÉTUDE DE CAS − LA MASSE SALARIALE TOTALE

18. Chaque fois que deux sources sont utilisées pour calculer des données similaires, la question de la comparabilité se pose. Dans le présent cas, les deux sources sont les données administratives et les données d'enquête. Nous avons l'intention de montrer comment des différences dans les définitions et les méthodes de collecte peuvent aboutir à des chiffres totalement différents de la masse salariale, différence qui peut injustement réduire la confiance dans les données d'enquête et ternir la réputation du Service national de statistique.

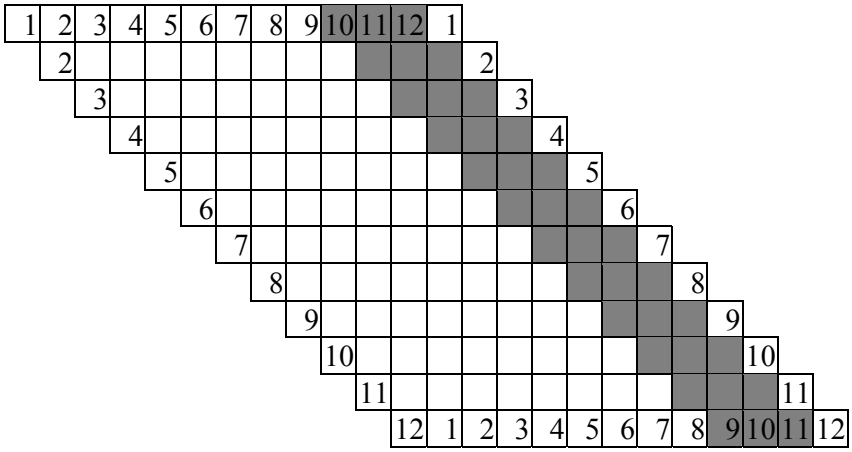
19. Pendant plus de trente ans, le Bureau central israélien de la statistique a utilisé deux sources de données pour établir des estimations de la masse salariale: une enquête basée sur les relevés communiqués par les employeurs à l'Office national de l'assurance (**A** − qui désigne ci-après la source administrative) concernant les postes occupés par des membres de leur personnel, et une enquête sur les revenus réalisée à partir d'une série d'enquêtes sur les forces de travail (ci-après LFS ou **S** qui désigne les données d'enquête).

20. Il est difficile de comparer les deux sources de données, pour les raisons suivantes:

a) **Population et couverture**: Les relevés des employeurs concernent les postes, et la LFS, les personnes. Les relations entre les personnes occupées et les postes sont parfois de

une-à-plusieurs: quelques personnes peuvent occuper plusieurs postes. De plus, le relevé concernant le poste d'une personne occupée est établi au moment d'un paiement. Cela ne signifie pas nécessairement que le travail a été effectivement réalisé au cours du mois, voire de l'exercice budgétaire, indiqué. Chaque fois qu'une personne est rémunérée pour un travail réalisé dans le passé, un versement additionnel peut être effectué six ou sept mois après la cessation de la relation de travail. Enfin, la source S ne rend pas compte de toute la population du pays. Les petits villages sont laissés de côté. Le groupe de population qui n'est pas pris en compte représente environ 7 % de la population totale;

b) **Choix du moment et période de référence**: Les employeurs établissent des relevés mensuels tout au long de l'année civile en se référant au mois précédent, alors que les personnes interrogées dans le cadre de la LFS se réfèrent à la période précédant le passage de l'enquêteur. Ces passages sont répartis uniformément sur l'année. C'est pourquoi la période de référence pour les participants à la LFS correspond non pas à un exercice budgétaire mais à une année mobile − les questions posées se réfèrent à un exercice comptable qui s'achève un mois avant le passage de l'enquêteur. En d'autres termes, la période couverte par l'enquête sur les revenus, avec une question concernant les recettes de l'année précédente, comprend vingt-trois mois avec des coefficients de pondération différents pour chaque mois. En 1986, l'exercice comptable qui était d'un an a été ramené à trois mois, et de ce fait l'enquête a porté sur quinze mois, comme indiqué ci-après:



c) La question qui est posée dans la LFS consiste à demander si la personne a travaillé le mois précédent, puis si elle a travaillé au cours de chacun des trois mois précédents. Or, dans le fichier mis à la disposition du public, les renseignements indiqués concernent le travail de l'enquêté au cours du mois précédent et son revenu au cours des trois mois précédents;

d) Couplage des relevés: Il n'est pas possible de comparer les relevés individuels car il n'est pas indiqué de numéros d'identification personnels (PIN) ni dans la LFS ni dans la source administrative. C'est pourquoi la comparaison doit être établie sur la base d'agrégats.

21. Entre 1970 et 1980, les masses salariales calculées d'après les deux sources ont été comparées à plusieurs reprises[2]. Il en est ressorti que la masse salariale totale d'après A est d'environ 15 % plus élevée que celle calculée d'après S. Depuis lors, la conclusion habituelle est que les données provenant de S comportent un biais par défaut. De ce fait, chaque fois que les chiffres de la pauvreté ou d'autres chiffres publiés ne sont pas du goût des commentateurs ou des fonctionnaires, ils font état du biais comme explication.

22. D'après l'explication donnée, le biais est dû à la fraude fiscale. Cependant il vaut la peine de faire observer que cette explication ne doit pas être prise avec sérieux. Il s'agit de savoir si l'argument s'applique aux salaires et non si les revenus du travail donnent lieu à une fraude fiscale. Par définition, les salaires sont déclarés à l'administration fiscale et l'impôt est déduit à la source. C'est pourquoi ce n'est pas la fraude fiscale qui incite à ne pas déclarer les salaires à l'enquêteur.

VII. UNE COMPARAISON THÉORIQUE

23. Comme il n'est pas possible dans la pratique de comparer les salaires déclarés au niveau des relevés individuels, il faut s'en remettre à des comparaisons d'agrégats. Il est possible cependant de faire appel à une source supplémentaire de données administratives. Il s'agit des déclarations des employeurs en fin d'année concernant les impôts retenus à la source. Ces déclarations portent sur l'ensemble de l'année et sont ventilées par poste occupé chaque mois, et elles indiquent les PIN, de sorte que l'on peut les utiliser pour reproduire à la fois la source A des données déclarées par l'employeur et la source B des données provenant de l'enquête. Comme nous nous appuyons sur les mêmes données, les différences n'influent pas sur nos conclusions, qu'il s'agisse des périodes de référence utilisées, de la variabilité d'échantillonnage et des différentes définitions de l'emploi. Nous pouvons ainsi avoir une idée des différences présumées entre les deux sources, qui sont dues à la méthode de collecte des données.

24. Ainsi qu'il est mentionné plus haut, les statistiques mensuelles publiées à l'aide de la source A reprennent les données correspondant à l'année tout entière tandis que la source S fournit des données d'enquête sur les ménages de manière uniforme relative à l'année à l'aide d'une question relative à la situation au regard de l'emploi pendant les mois précédant le passage de l'enquêteur. Afin d'observer la différence entre les deux méthodes de collecte, on peut prendre l'exemple d'une personne qui ne travaille qu'un mois sur l'année. Si l'on utilise la source A, la probabilité de constater que cette personne occupe un emploi un mois par an pendant l'année est de 1. Cette probabilité, si l'on utilise les données de la source S, elle est de $1/12^{e}$ parce que la personne en question ne sera considérée comme occupant un emploi que si le passage de l'enquêteur intervient au cours du mois où elle occupe un emploi[3]. Il en ressort que l'enquête S peut être considérée comme un échantillon **stratifié** des personnes occupant un emploi au cours de l'année, la probabilité de constater qu'elles occupent un emploi étant égale au

---

[2] Voir Special publication Series n° 593, 1973, p. 20 concerning Employees Income Survey 1971, or Monthly Statistical Report, Appendix, 29 juin 1978, p. 35.

[3] Afin de simplifier l'analyse, il n'est pas posé de questions concernant plusieurs mois. L'idée générale demeure la même bien que le nombre puisse changer.

nombre de mois ouvrés au cours de l'année, divisé par 12. Par exemple, la probabilité de constater que ceux qui travaillent toute l'année ont un emploi est de 1, alors qu'elle est de 1/2 pour ceux qui travaillent six mois par an.

25. Pour étudier ce qu'implique une telle différence en ce qui concerne la population active déclarée, le tableau 1 présente les calculs effectués à partir du fichier de prélèvement de l'impôt à la source des personnes ayant un emploi selon le nombre de mois d'emploi. Ces calculs permettent de reproduire les données administratives et les données d'enquête et d'étudier l'écart dû, dans les résultats, à la différence apparemment innocente constatée dans les méthodes de collecte des données.

26. Le contenu des colonnes du tableau qui comprennent les données de base nécessaires pour nos estimations est exposé ci-après.

La **colonne [1]** présente le nombre de mois ouvrés pendant l'année[4]. Chaque rangée représente le nombre de personnes et leur salaire, en fonction du nombre de mois pendant lesquels les personnes ont travaillé au cours de l'année.

La **colonne [2]** présente le nombre de personnes qui ont travaillé pendant le nombre correspondant de mois. Si une personne a travaillé pendant toute l'année, elle est comptée dans la dernière rangée. Environ 126 000 personnes n'ont travaillé qu'un seul mois. Le nombre total figurant dans cette colonne est le nombre de personnes distinctes qui ont occupé un emploi pendant au moins un mois au cours de l'année. Le chiffre de 2 598 000 désigne le **nombre de personnes ayant eu un emploi pendant l'année**.

La **colonne [3]** présente le nombre de postes occupés chaque mois par les personnes ayant eu un emplois au cours de l'année [2]. Le nombre de postes est toujours égal ou supérieur au nombre de personnes multiplié par celui des mois ouvrés.

La **colonne [4]** présente le nombre de postes occupés par une personne ayant un emploi. Comme on peut le constater, environ 9 % des postes correspondent à des postes supplémentaires. Ce chiffre peut être rapproché des résultats de la LFS, selon laquelle 9 % des enquêtés déclarent avoir un emploi supplémentaire[5].

La **colonne [5]** présente le nombre de personnes qui auraient travaillé selon la source S, c'est-à-dire le nombre escompté d'actifs occupés lorsqu'il leur a été demandé s'ils avaient un emploi le mois précédent. La probabilité de recevoir une réponse positive est de 1/12^{e} lorsqu'une

[4] Pour simplifier l'analyse, on suppose que la période d'emploi est continue, par exemple si une personne a travaillé pendant trois mois au cours de l'année, les mois d'emploi se suivent.

[5] La définition d'un poste supplémentaire est différente dans les deux sources. Dans le fichier A, un emploi correspond à un poste supplémentaire si deux employeurs déclarent verser un salaire à la même personne. Dans l'enquête S, c'est l'enquêté qui en donne la définition. Un actif occupé qui a changé d'emploi au milieu du mois est considéré comme ayant un emploi supplémentaire selon la source A, mais pas selon la LFS. Par ailleurs, un actif occupé peut déclarer avoir plusieurs emplois, mais si son salaire est versé par la même unité, il sera considéré comme occupant un seul poste dans la source A.

personne n'a eu un emploi que pendant un mois, de 2/12^{e} pendant deux mois, etc. Si l'on définit la colonne [2] comme correspondant à la population totale de chaque groupe, on obtient une estimation de ce chiffre en multipliant la population de chaque groupe par la probabilité d'une réponse positive.

La **colonne [6]** présente la somme des salaires mensuels exprimée en NIS (nouveau shekel israélien) pour toute l'année et pour chaque groupe.
Les six colonnes présentent toutes les données nécessaires pour effectuer la comparaison théorique entre les deux sources.

VIII. RÉSULTATS CONCERNANT LE NOMBRE DES PERSONNES AYANT UN EMPLOI

27. Si l'on compare la colonne [2] et la colonne [5], on constate qu'il existe une différence entre les personnes qui ont eu un emploi pendant l'année (c'est-à-dire qui ont eu un emploi pendant au moins un mois au cours de l'année) et les personnes occupées pendant un mois (c'est-à-dire qui ont déclaré avoir eu un emploi pendant le mois précédant le passage de l'enquêteur). Le nombre de personnes ayant eu un emploi pendant l'année est de 2,60 millions et celui des personnes ayant un emploi calculé au mois est de 2,06 millions, soit une différence de 26 %. Si l'on divise le nombre total de postes au mois par 12, on obtient (26,829/12=) 2,24 millions de postes au mois. Le nombre de postes d'actifs occupés est supérieur d'environ 9 % à celui des actifs occupés qui auraient dû être déclarés dans la LFS.

Tableau 1: Postes au mois, actifs occupés et traitements, exercice budgétaire 2003

Nombre de mois	Personnes ayant eu un emploi pendant l'année	Postes au mois (en milliers)	Postes supplémentaires	Actifs occupés au mois	Total des salaires (en millions de NIS)
[1]	[2]	[3]	[4] = [3]/[2]	[5] = [2]*([1]/12)	[6]
Total	2 597 858	26 829,2	1,09	2 057 388	187 571
1,00	125 769	128,4	1,02	10 481	329
2,00	104 556	216,7	1,04	17 426	584
3,00	91 095	285,8	1,05	22 774	820
4,00	93 149	392,8	1,05	31 050	1 257
5,00	81 680	432,6	1,06	34 033	1 443
6,00	86 358	551,4	1,06	43 179	1 982
7,00	83 432	625,5	1,07	48 669	2 327
8,00	94 604	817,3	1,08	63 069	3 333
9,00	90 917	886,1	1,08	68 188	3 634
10,00	108 087	1 176,7	1,09	90 073	5 067
11,00	117 163	1 418,2	1,10	107 399	5 997
12,00	1 521 048	19 897,9	1,09	1 521 048	160 799

28. **Comparaisons des salaires moyens**: de toute évidence, comme nous utilisons les mêmes données pour établir la différence entre la source A et la source B, toute absence de concordance doit être attribuée à une erreur dans les calculs mathématiques ou une erreur de logique. Cependant, ces deux catégories d'erreurs se produisent ou peuvent se produire s'il n'est pas tenu compte de la distinction subtile dans la période de référence; l'enquête S est stratifiée selon la durée de l'emploi, et le salaire mensuel lui est fortement corrélé, comme le montre le tableau 2.

Tableau 2: Salaire moyen par poste occupé et par personne occupée (NIS), 2003

Mois ouvrés [1]	Salaire moyen par poste occupé [2]	Salaire moyen par personne occupée [3]
Moyenne	6 991	7 597
1	2 565	2 619
2	2 697	2 794
3	2 869	3 000
4	3 199	3 372
5	3 336	3 533
6	3 594	3 825
7	3 720	3 984
8	4 078	4 404
9	4 101	4 441
10	4 306	4 688
11	4 228	4 653
12	8 081	8 810

29. Les chiffres indiqués dans le tableau 2 sont tirés du tableau 1. La colonne [2] du tableau 2 indique le salaire par poste occupé au mois (colonne [6]/colonne [3], du tableau 1), et la colonne [3] le salaire par personne occupée au mois {colonne [6]/{colonne [1]*colonne [5]}}. Comme on peut s'y attendre, plus la période d'emploi est longue, plus le salaire mensuel est élevé (excepté pour ceux qui ont travaillé onze mois). On constate la plus grande différence entre les personnes ayant un emploi pendant toute l'année et les autres. Comme le montre le tableau 2, le salaire moyen pour douze mois selon la source S est de 8 810, et il est, par poste occupé, selon la source A de 8 081, soit une différence de 9 %. Le salaire moyen par poste au mois est de 6 991, et il est de 7 597 par personne occupée; la différence est là encore de 9 %.

30. La manière appropriée pour établir une estimation de la masse salariale annuelle est la suivante: pour la source S, multiplier le salaire moyen par personne occupée qui a travaillé au cours du dernier mois (7 597) par le nombre moyen de personnes occupées au mois (2,06 millions)*12. Pour la source A, multiplier le nombre total de postes au mois par le salaire moyen par poste au mois. Ces deux façons différentes aboutissent à la même masse salariale totale.

31. La méthode décrite plus haut, qui consiste à utiliser uniquement le salaire payé le mois précédant le passage de l'enquêteur, en ne tenant pas compte du salaire payé pendant l'année, ne semble pas efficace et l'on est tenté d'utiliser d'autres méthodes qui pourraient améliorer l'estimation. Malheureusement, cela peut également conduire à des résultats biaisés. À cet effet, nous reproduisons les comparaisons entre les deux estimations de la masse salariale qui ont été effectuées dans les années 70 (voir la note 2 de bas de page), mais sans la retenue supplémentaire en fin d'année de l'époque, ce qui nous a permis de reproduire les deux enquêtes.
Des comparaisons ont été effectuées en utilisant les deux échantillons puisqu'il n'était pas possible de comparer les enregistrements individuels.

32. Le tableau 3 présente les postes/personnes, salaires moyens et masse salariale estimés pour différents types de méthode de calcul et les chiffres effectifs provenant des enquêtes A et S.

33. **La première ligne** du tableau 3 présente l'estimation selon la source A: afin d'obtenir le nombre moyen de postes au mois, soit 2 235, le nombre total de postes au mois indiqué pour l'année est divisé par 12. Le montant total des salaires est divisé par le nombre moyen de postes au mois et multiplié par 12. Nous qualifions les chiffres de cette ligne de «véritables résultats».

34. À **la deuxième ligne**, on calcule la masse salariale en multipliant le salaire mensuel par personne par le salaire moyen par personne occupée pendant le mois précédent (multiplié par 12), ce qui correspond à la façon correcte d'estimer la masse salariale à partir de l'enquête S.

35. **La troisième ligne** reproduit le type d'estimation réalisé dans les années 70. Le nombre de personnes occupées était établi à partir du nombre de personnes occupées au cours du mois précédent. Le revenu est le revenu perçu l'année précédente. En d'autres termes, comme dans la LFS, les chiffres publiés sont ceux soit des personnes occupées le mois précédent soit des salaires perçus l'année précédente (de 1986 − à partir des trois derniers mois), l'estimation de la masse salariale consiste logiquement à ajouter les masses salariales de toutes les personnes qui occupaient un emploi au cours du mois précédent. En divisant le salaire annuel par 12, on obtient un salaire mensuel de 7 257 nouveaux shekels. Ce salaire mensuel moyen est plus élevé de quatre points de pourcentage que le salaire mensuel par poste. En le multipliant par le nombre de personnes qui ont travaillé au cours du mois précédent, on obtient une masse salariale annuelle estimée de 179 milliards. La masse salariale totale ainsi obtenue est de 5 % inférieure à la véritable masse salariale annuelle. Cette opération reproduit la comparaison effectuée dans les années 70, qui avait amené à conclure que les données S étaient biaisées par défaut. Cela dit, cette comparaison était effectuée à partir de chiffres réels, et la différence constatée est plus grande que celle qui ressort de la comparaison théorique (14 %)[6].

[6] Par rapport aux données réelles, il faut également tenir compte de la différence de couverture entre les sources. Ce point sera abordé ultérieurement.

Tableau 3. Estimation de la masse salariale (en NIS), 2003

	Postes ou personnes occupées au mois (en milliers)	Salaire moyen	Masse salariale (en milliards par an)
1) Source A: postes	2 236	6 991	188
2) Source S: personnes occupées au mois	2 057	7 597	188
3) Source S: revenu annuel	2 057	7 257	179
Ratio (3/1)	1,09	0,96	1,05
4) Emploi occupé (postes)	2 340	6 972	196
5) Personnes occupées: un mois en moyenne	1 946*	7 112*	166
6) Personnes occupées − ajustement en fonction de la couverture	2 082	7 112	178
7) Personnes occupées	1 946*	6 908**	161
Ratio (4/1)	1,04	0,997	1,04
Ratio (6/1)	1,16	1,01	1,18

* Ces chiffres correspondent au nombre de personnes qui ont travaillé pendant le mois précédant le passage de l'enquêteur, et non à la moyenne sur trois mois publiée par le Bureau central israélien de la statistique.

** Moyenne sur trois mois, publiée par le Bureau central israélien de la statistique.

36. Les lignes 4) et 5) indiquent les chiffres effectifs communiqués par le Bureau central israélien de la statistique à partir des échantillons A et S. Le nombre effectif de postes occupés par mois qui a été indiqué est de 2 340, et le nombre de personnes qui étaient occupées pendant le dernier mois est de 1 946. Cependant, comme l'enquête porte sur 93 % seulement de la population, une révision rudimentaire consisterait à multiplier ce chiffre par 1,07. C'est ce chiffre révisé qui figure à la ligne 6). La masse salariale globale est de 178 milliards, soit 6 % de moins que la masse salariale fictive. D'après l'échantillon A, le nombre de postes occupés (ligne 4) s'établit en moyenne à 2,34 millions par mois, soit 4,6 % de plus que notre «véritable» estimation[7], ce qui correspond à une estimation biaisée par excès de 4 % de la masse salariale.

37. La ligne 7) indique le type de biais que l'on obtient si l'on utilise l'enquête S sans procéder à un ajustement, ce qui reproduit ce qui a été fait dans les années 70. Le biais de la masse salariale serait proche de 15 % (161/188), chiffre identique au biais exprimé en pourcentage qui a été constaté dans les années 70.

[7] Même si les chiffres des lignes 1 et 4 sont établis à partir des données administratives fournies par l'administration fiscale, l'écart entre les estimations peut s'expliquer en partie par les différences de méthodes utilisées par les entreprises pour communiquer les déclarations mensuelles et annuelles. Cela vient s'ajouter aux erreurs normales d'échantillonnage.

38. On trouve des arguments supplémentaires en faveur de la conclusion concernant la qualité des données S dans Romanov et Furman (2006) qui ont comparé les données administratives et les données de recensement, qui contiennent le PIN, et n'ont pas constaté de biais sérieux par défaut.

IX. CONCLUSION

39. La présente communication a pour but d'exposer les problèmes que l'utilisation de fichiers administratifs associés à des données d'enquête peut présenter pour les services nationaux de statistique. Le fait que deux sources, comprenant des données apparemment connexes, peuvent produire des résultats totalement différents parce que des différences d'une sorte ou d'une autre n'ont pas été prises en compte peut entamer la crédibilité accordée aux données d'enquête et ternir la réputation des bureaux nationaux de statistique. Ce problème se trouve aggravé par la tendance à divulguer des microdonnées, qui peut amener des chercheurs inexpérimentés à formuler des conclusions erronées concernant la qualité des données.

BIBLIOGRAPHIE

Furman, O. (2005). Comparative analysis of wages and work indicators from the income tax records of wage earners. CBS, CBS, Technical Paper No. 14.

Haltiwanger, J. C., J. I. Lane et J. R. Spletzer (2000). Wages, Productivity, and the Dynamic Interaction of Businesses and Workers, Working Paper No. 7994, National Bureau of Economic Research, Cambridge, MA.

Hamemesh, D. S. (2007). Fun with Matched Firm-Employee data: Progress and Road Maps, IZA Discussion Paper No. 2580, (janvier).

Hotz, V. J. et J. K. Scholz (2004). Measuring Employment and Income for Low-Income Populations with Administrative and Survey Data. http://www.econ.wisc.edu/~scholz/Research/NRC_Income_Measurement_Paper_Final_Draft.pdf.

Huynh, M, K. Rupp, J. Sears (2002). The Assessment of Survey of Income and Program Participation (SIPP) Benefit Data using Longitudinal Administrative Records, Working Paper No. 238, Office of Research, Evaluation, and Statistics, Social Security Administration, Washington, DC: US Census Bureau, 2002.

Kapteyn, A. et J. Y. YPMA (2005). Measurement Error and Misclassification (A comparison of Survey and Register Data), Working Paper, Rand, WR-283, (juillet).

Romanov, D. et O. Furman (2006). Analysis of wage data from the 1995 census by using wage file of the National Insurance Institute, CBS, Working Paper No.
