Étendue et conséquences des erreurs de mesure dans les données individuelles d'enquête : une évaluation à partir des données appariées des enquêtes emploi et revenus fiscaux

Cyrille Hagneré, Arnaud Lefranc

► To cite this version:

Cyrille Hagneré, Arnaud Lefranc. Étendue et conséquences des erreurs de mesure dans les données individuelles d'enquête : une évaluation à partir des données appariées des enquêtes emploi et revenus fiscaux. *Economie et prévision*, 2006, 174. hal-01651144

HAL Id: hal-01651144

<https://hal.science/hal-01651144>

Submitted on 1 Apr 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Étendue et conséquences des erreurs de mesure dans les données individuelles d'enquête : une évaluation à partir des données appariées des enquêtes emploi et revenus fiscaux

Cyrille Hagneré^(*)

Arnaud Lefranc^(**)

Différents facteurs sont susceptibles d'introduire, un écart dans les données microéconomiques entre la vraie valeur des variables d'intérêt et les valeurs enregistrées dans les enquêtes : erreurs de déclaration (intentionnelles ou non), erreurs de saisie, erreurs de mémoire dans les données rétrospectives, ... Beaucoup d'études économétriques tendent encore à traiter ces erreurs de mesure comme un bruit négligeable ou sans conséquences pratiques. Pourtant, certains travaux récents ont révélé que la qualité des données utilisées et l'existence d'erreurs de mesure substantielles pouvaient avoir des conséquences critiques pour l'analyse économétrique. Les principaux enseignements théoriques en la matière sont, d'une part, que la présence d'erreurs de mesure conduit en général à biaiser les résultats d'estimations économétriques et, d'autre part, que le biais est d'autant important que le « bruit d'erreur de mesure » est grand, c'est-à-dire, en termes techniques, que la variance de l'erreur de mesure représente une part importante de la variance « vraie » de la variable considérée. La possibilité de tels biais plaide alors pour un examen empirique approfondi de l'étendue et des conséquences des erreurs de mesure dans les données recueillies dans les enquêtes individuelles. Plusieurs articles récents ont alors entrepris d'évaluer empiriquement l'ampleur des erreurs de mesure pour certaines enquêtes microéconomiques fréquemment utilisées, principalement des enquêtes nord-américaines. Les résultats obtenus sont évidemment difficilement généralisables : des différences, par exemple, dans la formulation du questionnaire ou dans les attitudes individuelles vis-à-vis des enquêtes statistiques sont en effet susceptibles d'engendrer des variations, d'une enquête ou d'un pays à l'autre, dans la qualité de l'information recueillie et dans la précision des déclarations. Toutefois, ces travaux mettent en évidence des effets substantiels des erreurs de mesure : par exemple, pour les déclarations de salaire, les biais possibles dans les estimations économétriques peuvent atteindre, voire dépasser 50%. En France, il n'existe pas d'études comparables. L'objet de cet article est de procéder à un tel examen à partir de l'enquête Emploi de l'INSEE, qui constitue l'une des principales sources de données individuelles pour l'étude du marché du travail français.

(*) OFCE, THEMA et IDEP

E-mail: cyrille.hagnere@sciences-po.org

(**) THEMA, Université de Cergy-Pontoise, IDEP et Robert Schuman Center for Advance Studies, European University Institute

E-mail: arnaud.lefranc@u-cergy.fr

Une première version de cet article a été présentée aux Journées de méthodologie statistique de l'Insee en décembre 2002. Nous remercions les participants à ces Journées ainsi qu'un rapporteur anonyme de la revue pour leurs commentaires et suggestions.

Pour évaluer l'étendue des erreurs de mesure dans les données individuelles, il convient de comparer les réponses individuelles à l'enquête considérée à des données fournissant la vraie valeur des variables enquêtées. La démarche que nous retenons consiste à confronter les réponses individuelles à l'enquête Emploi de l'Insee, à des données auxiliaires, d'origine administrative, fournissant une information supposée exempte d'erreur de déclaration. L'intérêt de cette démarche est de fournir une évaluation de la qualité des réponses individuelles, sur un grand échantillon. Les données administratives que nous retenons sont les données des enquêtes Revenus Fiscaux, issues des registres de la Direction Générale des Impôts, qui fournissent la valeur des revenus d'activité individuels déclarés à l'administration fiscale. Ces données peuvent être appariées à certaines vagues de l'enquête Emploi. C'est sur ces fichiers appariés que nous procédons à l'évaluation de l'étendue et des conséquences des erreurs de mesure.

Les données exploitées imposent plusieurs contraintes et limites à notre analyse. Tout d'abord, compte tenu de l'information disponible dans ces deux enquêtes, notre analyse se limite à l'évaluation des erreurs de mesure dans les déclarations salariales individuelles. On examine les erreurs de mesure dans les niveaux de salaire déclarés, ainsi que dans les taux de croissance individuels, cette dernière variable étant fréquemment exploitée dans les travaux économétriques sur données de panel. Ensuite, des différences dans les concepts de revenus mesurés dans les deux sources (enquête Emploi et déclaration fiscale) sont susceptibles d'engendrer un écart dans les valeurs mesurées dans les deux fichiers, même en l'absence d'erreurs de mesure. Face à ce problème, deux voies sont empruntées dans cet article. La première consiste à évaluer les erreurs de mesure à partir de l'ensemble des données disponibles : on risque alors de surestimer l'importance des erreurs de mesure dans les niveaux de salaire déclarés. La seconde consiste à évaluer l'étendue des erreurs de mesure sur un échantillon restreint pour lequel les concepts de revenu des deux sources coïncident très fortement. On obtient alors une évaluation des erreurs de mesure sensiblement moins biaisée, mais sur un échantillon qui n'est pas représentatif de l'ensemble du marché du travail.

L'évaluation empirique de la qualité des déclarations de salaire à l'enquête Emploi et des conséquences possibles pour l'estimation de relations économétriques fait apparaître des résultats contrastés, selon qu'on considère les données de salaire en niveau ou en différence première. La qualité des déclarations de salaire en niveau se révèle particulièrement bonne, au regard notamment des résultats obtenus pour des enquêtes comparables. Les valeurs déclarées sont en général assez proches de la vraie valeur du salaire perçu et les erreurs de mesure ne sont susceptibles d'engendrer que des biais assez faibles dans les modèles de régressions. En revanche, les données en différence apparaissent nettement plus bruitées, ce qui semblerait conduire à des biais économétriques très importants. Les estimations menées dans cette étude méritent toutefois, dans le cas des données en différence, d'être interprétées avec prudence : en raison de l'imparfaite adéquation des deux sources de données utilisées et des contraintes qu'elle impose sur la définition de l'échantillon sur lequel sont menées nos estimations, il est possible que notre évaluation surestime l'étendue des erreurs de mesure dans les variations inter-annuelles de salaire.

Pour finir, l'article cherche à préciser l'origine des erreurs de mesure observées. Les comportements d'arrondi sont souvent évoqués comme une source possible d'écart entre la vraie valeur du salaire perçu et la valeur déclarée dans les enquêtes. Nous proposons une méthode originale permettant de mesurer l'incidence et les conséquences des comportements d'arrondis et montrons que si les comportements d'arrondi sont vraisemblablement fréquents, leur contribution aux erreurs de mesure et aux biais d'estimation est relativement limitée.

Différents facteurs sont susceptibles d'introduire un écart entre les valeurs enregistrées dans les données individuelles d'enquête et la vraie valeur des variables enquêtées : erreurs de déclaration (intentionnelles ou non), erreurs de saisie, erreurs de mémoire dans les données rétrospectives, ... Beaucoup d'études économétriques tendent encore à traiter ces erreurs de mesure comme un bruit négligeable ou sans conséquences pratiques. Pourtant, certains travaux récents ont révélé que la qualité des données utilisées et l'existence d'erreurs de mesure substantielles pouvaient avoir des conséquences importantes pour l'analyse économétrique et dans certains cas biaiser les résultats d'estimation⁽¹⁾. La possibilité de tels biais plaide alors pour un examen empirique approfondi de l'étendue et des conséquences des erreurs de mesure dans les données recueillies dans les enquêtes individuelles. L'objet de cet article est de procéder à un tel examen à partir de l'enquête Emploi de l'Insee, qui constitue une des principales sources de données individuelles pour l'étude du marché du travail français.

L'évaluation des erreurs de mesure dans les données individuelles n'est cependant pas chose aisée. Elle nécessite en effet de confronter les réponses individuelles à l'enquête considérée à des données fournissant la vraie valeur des variables enquêtées. Pour procéder à un tel examen, plusieurs voies de recherche peuvent être empruntées.

Certaines études ont recours à des enquêtes spécifiques dédiées à l'évaluation de la qualité de l'information statistique recueillie. Il s'agit alors, le plus souvent sur un échantillon de taille limitée, de collecter simultanément les réponses individuelles au questionnaire d'enquête et la vraie valeur des variables enquêtées. Telle est par exemple la démarche suivie par l'étude de validation du *Panel Study of Income Dynamics (PSID-Validation Study)* dans laquelle le questionnaire de l'enquête est administré à un échantillon d'individus dont les réponses sont ensuite rapprochées des registres de l'employeur⁽²⁾. Ceci permet alors d'évaluer la qualité des réponses à certaines questions fondamentales pour les études empiriques en économie du travail : revenus annuels, heures travaillées, salaires horaires, ancienneté dans la firme, etc.

Une démarche alternative consiste à confronter les réponses individuelles à une enquête existante à des données auxiliaires, le plus souvent d'origine administrative, fournissant une information exempte d'erreur de déclaration. Tel est par exemple la démarche mise en œuvre par Bound et Krueger (1991) qui procèdent à un appariement des registres de sécurité sociale et des données des *Current Population Surveys*. L'intérêt est alors de fournir une évaluation de la qualité des données sur un échantillon plus vaste et plus représentatif. Les principales limites de cette démarche sont, d'une part, le nombre plus restreint de variables pour lesquelles on peut mesurer la qualité des réponses individuelles et, d'autre part, le fait que les données administratives disponibles ne fournissent pas toujours une information strictement comparable à celle obtenue dans les enquêtes.

Enfin, une troisième voie consiste à comparer les distributions des revenus obtenues à partir de différentes sources de données, par exemple, des données d'enquête et des données administratives. Dans ce cas, on dispose alors d'une évaluation globale de la proximité entre les réponses à une enquête et les enregistrements administratifs mais, faute d'appariement, il n'est pas possible de mener une analyse de l'étendue des erreurs de mesure au niveau individuel et de leurs conséquences.

Dans cet article, nous empruntons la deuxième voie d'analyse. L'évaluation de la qualité des réponses individuelles aux enquêtes Emploi de l'Insee entreprise dans cet article exploite les données appariées de l'enquête Emploi et de l'enquête Revenus Fiscaux⁽³⁾. Compte tenu de l'information disponible dans ces deux enquêtes, notre analyse se limite aux déclarations salariales individuelles. L'article est organisé de la façon suivante. La première partie procède à un rappel des conséquences possibles de l'existence d'erreurs de mesure pour l'analyse économétrique et présente les principales statistiques permettant d'évaluer la qualité des données déclarées. La deuxième partie présente les données utilisées. Les troisième et quatrième parties examinent la qualité des niveaux de salaires et des taux de croissance des salaires déclarés. Enfin, la dernière partie propose deux compléments : elle analyse la contribution des comportements d'arrondi aux erreurs de mesure et évalue l'incidence des erreurs de mesure pour l'analyse de la mobilité salariale.

Deux résultats principaux émergent de notre analyse. D'une part, la qualité des déclarations de salaire en niveau dans l'enquête Emploi apparaît particulièrement bonne, au regard notamment des résultats d'études similaires portant sur d'autres enquêtes de même nature. D'autre part, les variations de salaire au cours du temps, calculées à partir de l'enquête Emploi, semblent très peu corrélées aux véritables variations sous-jacentes des rémunérations individuelles. Il en découle que l'utilisation des salaires de l'enquête Emploi en différence première, dans les travaux économétriques, est susceptible de conduire à des résultats substantiellement biaisés.

Cadre d'analyse

Pour examiner les conséquences de la présence d'erreurs de mesure dans les résultats d'estimations économétriques, on suppose l'existence d'une relation linéaire entre une variable Y^* et une matrice de variables explicatives X^* :

$$Y^* = X^* \beta + \varepsilon$$

On suppose par ailleurs que les variables Y^* et X^* ne sont pas directement observées dans l'enquête mais sont mesurées avec erreurs. On observe dans l'enquête les variables X et Y , avec :

$$Y = Y^* + v$$

$$X = X^* + u$$

où v et u représentent les erreurs de mesure dans la variable dépendante et dans les variables explicatives.

L'estimateur de β par les moindres carrés ordinaires à partir des variables observées est donné par :

$$b = (X'X)^{-1} X'Y$$

Dans ce contexte, on peut montrer que les propriétés de l'estimateur des moindres carrés dépendent des propriétés des erreurs de mesure. Plusieurs cas doivent alors être distingués selon que l'erreur de mesure porte sur la variable dépendante ou les variables explicatives et que l'erreur est ou non corrélée aux variables sous-jacentes.

Erreurs de mesure classiques

Le cas d'erreurs de mesure classiques correspond à l'absence de corrélation entre l'erreur de mesure et la variable qu'on cherche à mesurer dans l'enquête. En présence d'erreurs de mesure classiques dans la variable dépendante Y , l'estimateur des moindres carrés reste sans biais mais les erreurs de mesure diminuent la précision de l'estimateur des MCO.

Par contre, la présence d'erreurs de mesure classiques dans la ou les variables indépendantes entraîne un biais dans l'estimation de β . On peut d'abord considérer le cas simple où il existe une seule variable explicative. Dans ce cas, l'estimateur des MCO est donné par :

$$\begin{aligned} b &= \frac{\text{cov}(X, Y)}{V(X)} \\ &= \beta \frac{V(X^*)}{V(X)} \\ &= \frac{\beta}{1 + \frac{V(u)}{V(X^*)}} = \beta \lambda \end{aligned}$$

Le terme $1/(1 + V(u)/V(X^*))$ est en général désigné sous le nom de ratio de fiabilité (*reliability ratio*). Ce ratio prend ses valeurs entre 0 et 1. Il est d'autant plus faible que la variance des erreurs de mesure est importante comparativement à la variance de la variable sous-jacente X^* . On voit alors qu'en présence d'erreurs de mesure classiques sur une variable explicative, l'estimateur des moindres carrés est biaisé vers 0.

Dans le cas multivarié, on peut réécrire l'estimateur des MCO sous la forme :

$$b = (X'X)^{-1} X'(X\beta + u\beta)$$

$$b = (I + (X'X)^{-1} X'u)\beta$$

On peut alors montrer que même si une seule des composantes X_i de la matrice X est mesurée avec erreurs de mesure classiques, le coefficient estimé de toutes les composantes de X corrélées à X_i sera biaisé et le sens du biais dépendra du signe de la corrélation entre X_i et la composante considérée.

Erreurs de mesure non classiques

Dans le cas général, il n'y a pas de raison de supposer que le terme d'erreur est indépendant de la variable d'intérêt. Dans ce cas, les coefficients estimés par les MCO seront toujours biaisés, que l'erreur de mesure porte sur une variable explicative ou sur la variable dépendante. Par ailleurs, la nature des biais diffère du cas d'erreurs de mesure classiques.

La présence d'erreurs de mesure corrélées à la variable d'intérêt conduit à une estimation biaisée du paramètre β dans le cas où l'erreur porte sur la variable dépendante. Ecrivons le terme d'erreur v sous la forme :

$$v = \rho Y^* + v'$$

où v' est par construction orthogonal à Y^* . L'estimateur des MCO se réécrit alors :

$$b = \beta(1 + \rho)$$

On a donc un biais proportionnel qui dépend de la valeur de ρ .

Dans le cas univarié, lorsque l'erreur porte sur la variable indépendante, l'estimateur des MCO est aussi biaisé. Cependant, le biais d'atténuation, présent dans le cas d'erreurs de mesure classiques, sera renforcé ou amoindri selon le sens de la corrélation entre l'erreur de mesure et la variable sous-jacente. En présence d'erreurs de mesure corrélées à la variable sous-jacente, on peut en effet décomposer l'erreur u de la façon suivante :

$$u = \rho X^* + u'$$

L'estimateur des moindres carrés vaut alors :

$$b = \frac{\beta}{1 + \rho + \frac{V(u')}{V(X^*)}}$$

Le biais est donc plus faible si l'erreur de mesure est négativement corrélée à X^* c'est-à-dire si $\rho < 0$ ⁽⁴⁾.

On peut aussi réécrire l'estimateur des MCO sous la forme :

$$b = \beta \frac{\text{cov}(X, X^*)}{V(X)}$$

$$\text{soit } b = \beta\gamma$$

$$\text{avec } \gamma = \frac{\text{cov}(X, X^*)}{V(X)}$$

Il importe de noter que l'évaluation de γ n'exige pas nécessairement que la vraie valeur X^* soit observée. Supposons qu'on dispose d'une deuxième variable X' telle que $X' = X^* + u'$. Alors, si l'erreur de mesure dans la deuxième variable (u') n'est pas corrélée avec la première variable ($\text{cov}(X, u') = 0$), on a $\text{cov}(X, X^*) = \text{cov}(X, X')$.

Dans le cas multivarié, on constate encore que même si une seule variable est mesurée avec erreur, le biais se reporte au coefficient estimé des autres variables, du fait de la possible corrélation entre l'erreur de mesure et les autres variables de la régression.

Indicateurs statistiques de qualité des déclarations

L'analyse ci-dessus suggère donc que les conséquences des erreurs de mesure dans les données d'enquête peuvent être examinées au travers de trois statistiques importantes :

- le ratio de fiabilité $\lambda = 1/(1+V(u)/V(X^*))$ (uniquement dans le cadre d'erreurs de mesure classiques) ;
- le coefficient de régression ρ des erreurs de mesure sur la vraie valeur de la variable enquêtée ;
- le coefficient de régression de la valeur déclarée à l'enquête sur la vraie valeur de la variable enquêtée $\gamma = \text{cov}(X, X^*)/V(X)$.

Dans la suite de l'article, ces différents indicateurs sont estimés sur les données de l'échantillon apparié enquête Emploi – Revenus Fiscaux.

Données utilisées

Les enquêtes Emploi et Revenus Fiscaux

Les données utilisées proviennent des vagues 1996, 1997 et 1998 de l'enquête Emploi⁽⁵⁾ (EE) et des vagues 1996 et 1997 de l'enquête Revenus Fiscaux (ERF). L'enquête Emploi représente la principale enquête microéconomique réalisée par l'Insee sur la force de travail. Dans cette enquête, les individus sont interrogés sur leur situation sur le marché du travail ainsi que sur leurs revenus salariaux. L'enquête Emploi est réalisée au mois de mars de chaque année. Par ailleurs, l'échantillon des logements enquêtés est renouvelé annuellement par tiers, ce qui permet de suivre les individus enquêtés au cours de trois années consécutives (sauf en cas de déménagement).

Les enquêtes Revenus Fiscaux de 1996 et 1997 sont, quant à elles, issues de l'appariement des fichiers de déclarations fiscales de la Direction Générale des Impôts⁽⁶⁾ et d'une partie de l'échantillon de l'enquête Emploi de l'année correspondante. Les informations recueillies se rapportent à l'année fiscale considérée. L'appariement entre les enquêtes Emploi et les fichiers fiscaux a été réalisé pour deux années fiscales et trois vagues d'enquête Emploi : l'enquête Revenus Fiscaux 1996 a ainsi été appariée avec les enquêtes Emploi de mars 1996 et mars 1997 et l'enquête Revenus Fiscaux 1997 a été appariée avec les enquêtes Emploi de mars 1997 et mars 1998⁽⁷⁾.

L'appariement des déclarations fiscales et des données de l'enquête Emploi n'est pas, faute d'identifiant individuel commun aux deux sources de données, un appariement "exact". Il s'agit d'un appariement dit statistique : les observations des deux fichiers sont appariées à partir d'informations communes aux deux sources. Ces informations sont le prénom, la date de naissance (mois et année), le sexe et l'adresse de l'individu. Pour environ 87% des ménages de l'enquête Emploi, cette procédure permet de retrouver la déclaration fiscale d'au moins un des membres du ménage. Pour 80% des ménages, on retrouve les déclarations fiscales de l'ensemble des membres du ménage⁽⁸⁾. La documentation de l'enquête Revenus Fiscaux indique toutefois que l'échec de la procédure d'appariement semble aléatoire et n'entraîne pas de biais de sélection susceptible d'affecter les résultats de notre étude⁽⁹⁾.

À partir des données de ces deux fichiers, il est possible de calculer l'étendue des erreurs de mesure dans les données de salaire déclarées à l'enquête Emploi. Sous l'hypothèse que les déclarations fiscales de salaire perçu sont exemptes d'erreur d'enregistrement ou de déclaration, l'écart entre la déclaration à l'enquête Emploi et la déclaration fiscale fournit une évaluation au niveau individuel des erreurs de mesure dans l'enquête Emploi. Il s'agit évidemment d'une hypothèse forte, dont on sera amené à discuter la pertinence, dans un certain nombre de cas particuliers⁽¹⁰⁾. On fera ici l'hypothèse d'une erreur de mesure multiplicative en niveau (et donc additive en log). L'erreur de mesure dans l'enquête Emploi sera alors définie et calculée comme :

$$(1) \ u = \ln(\text{salaire}_{EE}) - \ln(\text{salaire}_{ERF})$$

En pratique, le calcul des erreurs de mesure se heurte cependant au fait que le champ des variables de revenu diffère entre les deux sources. Ceci impose de porter une attention particulière aux variables retenues et à l'échantillon sur lequel porte la comparaison.

Variables retenues et sélection de l'échantillon

La variable de revenu retenue dans l'enquête Revenus Fiscaux correspond aux revenus d'activité perçus au cours de l'année fiscale. Il s'agit de la variable ZTSAO de l'enquête, qui reprend les cases AJ ou BJ du formulaire de déclaration de revenu. Pour l'enquête Emploi, on retient principalement le salaire mensuel perçu dans l'emploi principal au moment de l'enquête. Il s'agit de la variable SALFR de l'enquête Emploi. On utilise aussi les valeurs déclarées des primes non-mensuelles perçues par l'individu. Pour les deux enquêtes, on ne retient que les valeurs déclarées par l'individu et l'on exclut les valeurs imputées⁽¹¹⁾. Les concepts de revenu des deux sources statistiques ne sont donc pas directement comparables : l'écart entre les déclarations à l'enquête Emploi et les enregistrements fiscaux mesuré par l'équation (1) peut donc refléter à la fois les erreurs de mesure et de déclaration individuelle à l'enquête Emploi et les différences dans le champ des variables utilisées. Il est donc nécessaire de limiter l'incidence des différences de champ en restreignant l'échantillon sur lequel porte la comparaison des deux sources.

Différences dans la période de référence des deux sources

La première limite à la comparabilité des données issues des deux sources tient aux différences dans la période de référence utilisée pour les déclarations de revenu. Les déclarations à l'enquête Revenus Fiscaux se rapportent aux revenus perçus au cours de l'année fiscale considérée. *A contrario*, dans l'enquête Emploi, le salaire déclaré est le salaire mensuel courant au moment de l'enquête, c'est-à-dire, pour la quasi-totalité des personnes salariées dans l'enquête Emploi, le salaire du mois de février de l'année d'enquête. On comparera donc le revenu déclaré à l'enquête Emploi au revenu fiscal ramené à une base mensuelle⁽¹²⁾. Par ailleurs, pour chaque année fiscale considérée, il apparaît pertinent de comparer le salaire mensuel calculé à partir de Revenus Fiscaux au salaire déclaré à l'enquête Emploi en mars de l'année fiscale considérée ainsi qu'en mars de l'année fiscale suivante.

Du fait de cette différence dans la période de référence, un écart entre les deux sources de données est donc susceptible d'apparaître même en l'absence d'erreurs de déclarations dans l'enquête Emploi, dès lors que le salaire mensuel varie au cours de l'année fiscale. Plusieurs restrictions de l'échantillon utilisé dans l'évaluation permettent cependant de limiter l'incidence des différences relatives à la période de référence.

On se restreint tout d'abord aux individus ayant été employés continûment au cours de l'année fiscale considérée. Par ailleurs, afin de se restreindre à un ensemble d'individus dont le salaire mensuel en février de l'année d'enquête peut être considéré comme représentatif des salaires perçus au cours des autres mois de l'année, on exclut de notre échantillon les individus ayant changé d'établissement au cours de l'année fiscale considérée, ceux ayant changé de profession ainsi que les individus employés sous une forme contractuelle autre que le contrat à durée indéterminée. Ces restrictions permettent alors de réduire la variabilité infra-annuelle des rémunérations mensuelles et de rendre plus comparables les deux sources.

On notera que ces différents critères de sélection conduisent à exclure de notre échantillon les individus dont l'insertion au marché du travail est la plus précaire⁽¹³⁾. Notre évaluation de l'étendue des erreurs de mesure concerne donc des individus connaissant une stabilité professionnelle importante et dont les revenus sont peut-être mieux renseignés que ceux d'individus connaissant des trajectoires professionnelles instables.

Ces critères de sélection ne suffisent pas toutefois à garantir que le salaire déclaré à l'enquête Emploi est représentatif des salaires perçus au cours de l'année fiscale. Il est en particulier possible que la durée du travail ou les conditions de travail se modifient au cours de l'année, entraînant ainsi des variations de salaire. Nous examinerons l'incidence de ces facteurs dans la suite de notre analyse.

Différences dans les revenus pris en compte dans les deux sources

Le champ des variables de revenu des deux sources diffère aussi à plusieurs égards.

Premièrement, le revenu renseigné dans l'enquête Emploi est le salaire dans l'emploi principal. Les données fiscales enregistrent quant à elles l'ensemble des revenus d'activité perçus. Ceci conduit à imposer deux restrictions à notre échantillon. D'une part, on se limite aux individus salariés. D'autre part, on exclut les individus déclarant (à l'enquête Emploi) exercer des activités secondaires rémunérées.

Deuxièmement, le revenu fiscal d'activité incorpore les primes et compléments salariaux non mensuels alors que le salaire déclaré à l'enquête Emploi correspond au salaire mensuel courant hors primes. Toutefois, les individus de l'enquête Emploi sont aussi interrogés sur la perception éventuelle de primes et compléments salariaux et sur le montant de ces revenus. On calcule donc, dans l'enquête Emploi, un salaire incluant les primes et correspondant mieux au concept de revenu de l'ERF⁽¹⁴⁾. Une incertitude demeure toutefois en ce qui concerne la prise en compte des stock-options. Les revenus d'activité fiscaux incorporent une partie de la valeur des stock-options perçues par les salariés : est prise en compte, dans les revenus d'activité fiscaux, la valeur du rabais

excédentaire⁽¹⁵⁾ dont bénéficie le salarié. Celle-ci est enregistrée au moment de la levée de l'option. Le questionnaire de l'enquête Emploi stipule quant à lui que "les compléments versés sous forme d'actions ou de fonds bloqués non disponibles immédiatement" ne doivent pas être déclarés. Une incertitude demeure quant à l'éventuelle prise en compte de ces compléments, dans l'enquête Emploi, au moment de la levée de l'option.

Troisièmement, les déclarations fiscales incluent la valeur de la CRDS et de la CSG non déductible alors que cette dernière n'est pas incluse dans les déclarations à l'enquête Emploi. On a donc soustrait la CRDS et la CSG non déductible de la déclaration fiscale afin de rendre comparables les données des deux sources.

Des écarts dans les concepts de revenu des deux sources peuvent toutefois subsister, qu'il est difficile de corriger. Tout d'abord, les revenus d'activité déclarés au fisc incluent aussi les indemnités journalières de maladie, de pré-retraite ou de chômage. Le fait de restreindre notre échantillon aux individus continuellement employés au cours de l'année fiscale considérée et exerçant une activité salariée permet toutefois de limiter en partie l'incidence de ces écarts⁽¹⁶⁾. Par ailleurs, la déclaration fiscale de revenus d'activité inclut aussi la valeur des avantages en nature perçus, ce qui n'est pas le cas dans l'enquête Emploi⁽¹⁷⁾.

Cas particuliers

La comparabilité des déclarations fiscales et des déclarations à l'enquête Emploi n'est pour finir pas assurée dans un certains nombre de cas particuliers.

Tout d'abord, la documentation de l'enquête Revenus Fiscaux indique que les individus travaillant à l'étranger ne déclarent souvent aucun revenu à l'administration fiscale. Nous excluons donc ces individus de notre comparaison. Par ailleurs, il peut être difficile de retrouver l'ensemble du revenu fiscal des individus ayant connu un changement d'état matrimonial (mariage, divorce,...) en cours d'année fiscale puisqu'ils remplissent alors plusieurs déclarations. Nous excluons aussi ces individus de notre échantillon.

On peut ensuite vouloir exclure de notre analyse les individus dont les salaires dans l'enquête Emploi sont "à l'évidence" erronés. À cette fin, on a comparé les déclarations individuelles sur deux années consécutives⁽¹⁸⁾. On a éliminé les individus dont le revenu à l'enquête Emploi est divisé ou multiplié, d'une année sur l'autre, par un facteur d'environ 10 ou d'environ 100⁽¹⁹⁾. On peut en effet penser que ces enregistrements présentent des problèmes de saisie.

On peut aussi penser que la comparaison est délicate dans le cas de certaines professions. Il peut s'agir de professions bénéficiant de dispositions fiscales spécifiques, d'avantages en nature importants ou dont les revenus peuvent comporter une part variable importante difficile à mesurer dans l'enquête Emploi. Afin que notre analyse ne soit pas faussée par quelques cas particuliers, nous examinons l'influence de chaque profession⁽²⁰⁾ sur nos résultats. Cette analyse de sensibilité est présentée en annexe. Elle nous conduit à exclure une seule profession : les assistantes maternelles dont les revenus d'activité font l'objet d'un traitement fiscal particulier⁽²¹⁾.

Effet des critères de sélection de l'échantillon

Les différentes restrictions imposées aux échantillons des enquêtes Revenu Fiscaux et Emploi ont pour but de rendre comparables les mesures du salaire disponibles dans les deux sources de données. Il conviendra cependant de garder à l'esprit, dans l'analyse de nos résultats, l'absence de stricte coïncidence des concepts de revenu issus des deux sources et les restrictions imposées à l'échantillon sur lequel nous évaluons les erreurs de mesure. De ce fait, nos résultats devront être interprétés, pour l'échantillon étudié, comme une borne inférieure de la qualité des déclarations de salaire individuel : l'absence de stricte coïncidence des concepts de revenu implique en effet que même si les déclarations à l'enquête Emploi étaient exactes, des différences entre les deux sources pourraient subsister. En outre, notre procédure constitue une évaluation jointe de la qualité de l'information salariale déclarée et des déclarations annexes servant à restreindre notre échantillon : ici encore, même en l'absence d'erreurs de déclaration dans les montants salariaux, des erreurs de déclaration dans les variables de sélection de l'échantillon pourraient conduire à un écart entre les deux enregistrements⁽²²⁾.

Par ailleurs, ces critères de sélection de l'échantillon nous limitent à l'étude d'une sous-population qui, du point de vue de la qualité des déclarations de salaire, n'est pas nécessairement représentative de l'ensemble des salariés. Pour les deux vagues d'enquête Revenu Fiscaux, la comparaison aux données de l'enquête Emploi ne concerne que la moitié des individus présents aux deux enquêtes et déclarant un revenu d'activité dans ces deux enquêtes. Les résultats présentés dans la suite de cet article s'appliquent uniquement aux individus connaissant une stabilité professionnelle forte. Pour le reste de l'échantillon, il n'est pas possible d'examiner de façon rigoureuse la qualité des déclarations à l'enquête Emploi, fort des différences de concept de revenu entre les deux sources. En comparant les deux sources pour l'ensemble de l'échantillon, on peut toutefois donner une bonne

inférieure à la qualité globale des déclarations à l'enquête Emploi. Les résultats de cette analyse sont présentés en annexe. Ils montrent que la qualité des déclarations pour les individus ne connaissant pas une stabilité professionnelle forte est vraisemblablement plus faible que pour l'échantillon étudié dans le reste de l'article. Toutefois, les résultats d'ensemble ne seraient pas fortement modifiés si l'on retenait un échantillon plus large⁽²³⁾.

Erreurs de mesure dans les niveaux de salaires déclarés

Les principaux résultats concernant les erreurs de mesure dans les niveaux de salaire sont présentés dans les graphiques 1 et 2 et les tableaux 1 et 2.

Les graphiques présentent les densités de salaire dans les deux sources de données, obtenues par estimation non-paramétrique⁽²⁴⁾. La comparaison des distributions indique que les valeurs faibles du salaire mensuel (entre 5 000 et 10 000 francs) sont plus représentées dans l'enquête Emploi que dans celle des Revenus Fiscaux. Inversement, les valeurs intermédiaires de la distribution (entre 10 000 et 20 000 francs) apparaissent sous-représentées dans les enquêtes Emploi, et ce, quelle que soit la paire EE-ERF retenue.

La distribution des erreurs de mesure est unimodale et symétrique. Il convient aussi de noter que les queues de distribution sont plus épaisses que dans le cas d'une distribution normale : pour l'ensemble des distributions des erreurs de mesure, la statistique de Kurtosis est de l'ordre de 100.

En moyenne, les déclarations à l'enquête Emploi tendent à sous-estimer la valeur du salaire mensualisé. Lorsqu'on compare l'enquête Revenus Fiscaux à l'enquête Emploi de la même année, la valeur moyenne de l'erreur est de l'ordre de -0,05 point de logarithme. Sur ce point, nos résultats contrastent avec ceux obtenus dans

Tableau 1 : erreur de mesure dans les niveaux de salaire - ERF96-EE96 et ERF96-EE97

Critère	Valeur	obs.	μ	σ	ρ	λ	γ	obs.	μ	σ	ρ	λ	γ
				ERF96-EE96						ERF96-EE97			
Tous		5761	-0,04519	0,18281	-0,09903	0,87314	0,95116	5796	-0,02645	0,18249	-0,10534	0,8758	0,96083
Sexe	F	2580	-0,05337	0,17655	-0,08168	0,88763	0,95337	2596	-0,0285	0,18213	-0,09295	0,88498	0,96078
	H	3181	-0,03855	0,1875	-0,14155	0,84257	0,94987	3200	-0,02478	0,1828	-0,13845	0,84897	0,95621
Répondant tiers	0	4135	-0,04447	0,18418	-0,10203	0,86917	0,94877	3818	-0,0213	0,1827	-0,10558	0,87554	0,96072
	1	1626	-0,04701	0,17932	-0,09236	0,88172	0,95598	1978	-0,03638	0,18174	-0,10351	0,87383	0,95639
Compléments salariaux	0	1800	-0,0454	0,19779	-0,09514	0,88848	0,96753	1831	-0,02522	0,20895	-0,11149	0,88102	0,97416
	1	3961	-0,04509	0,17561	-0,10938	0,85368	0,9349	3965	-0,02701	0,16891	-0,10765	0,86336	0,94633
Horaires irréguliers	0	3000	-0,05033	0,16325	-0,07013	0,88197	0,93589	3021	-0,02623	0,1618	-0,07401	0,88488	0,94289
	1	2761	-0,0396	0,20181	-0,13012	0,86202	0,96672	2775	-0,02669	0,20266	-0,13544	0,8643	0,97566
Temps de travail instable	?	663						674					
	0	3355	-0,04438	0,16023	-0,11195	0,87581	0,96748	3359	-0,02147	0,15294	-0,10869	0,88683	0,9792
	1	1743	-0,05262	0,19973	-0,07093	0,86916	0,92107	1763	-0,03525	0,20454	-0,09184	0,86382	0,93244
Temps plein	0	593	-0,0614	0,23551	-0,05889	0,87055	0,91288	593	-0,04046	0,24056	-0,08828	0,86924	0,93619
	1	5168	-0,04333	0,17569	-0,13944	0,84822	0,95612	5203	-0,02485	0,17461	-0,14203	0,85102	0,96292
Conditions de travail instables	?	34						31					
	0	5389	-0,04676	0,17428	-0,08437	0,88137	0,948	5424	-0,0266	0,1723	-0,09161	0,88609	0,96092
	1	338	-0,05068	0,22094	-0,15093	0,82859	0,93819	341	-0,03809	0,25698	-0,19578	0,78493	0,91136
Secteur public	0	4089	-0,03175	0,19166	-0,08604	0,87355	0,93963	4124	-0,01475	0,18795	-0,09254	0,87987	0,95377
	1	1672	-0,07805	0,15431	-0,13146	0,84559	0,94439	1672	-0,05528	0,16482	-0,1472	0,831	0,9382
Imposable	0	1582	-0,01404	0,24141	-0,17729	0,78179	0,88986	1582	0,00828	0,23074	-0,1748	0,80268	0,92076
	1	4179	-0,05698	0,15335	-0,06495	0,89567	0,94778	4214	-0,03948	0,15877	-0,07332	0,89074	0,94945

Note : μ désigne la moyenne, σ l'écart type, ρ , λ et γ sont définis dans la première partie. "répondant tiers" vaut 0 si le salaire a été déclaré à l'enquête Emploi par la personne elle-même et 1 si le salaire a été déclaré par un tiers; "compléments salariaux" vaut 0 pour les individus déclarant ne pas toucher des compléments salariaux non-mensuels et 1 sinon; "horaires irréguliers" vaut 1 pour les individus déclarant ne pas avoir d'horaire de travail régulier et 0 sinon; "temps de travail stable" vaut 1 pour les individus dont l'horaire hebdomadaire habituel ne varie pas d'une enquête à l'autre et 0 pour ceux dont l'horaire de travail varie; "temps plein" vaut 1 pour les personnes déclarant un horaire hebdomadaire habituel égal à 39 heures dans les deux enquêtes; "conditions de travail instables" vaut 1 pour les individus dont les conditions de travail (travail le dimanche, le samedi et/ou la nuit) se sont modifiées entre deux enquêtes et 0 pour ceux dont les conditions de travail sont restées stables; "secteur public" vaut 1 pour les personnes employées dans le secteur public; "imposable" vaut 1 pour les personnes appartenant à un foyer fiscal imposable.

Tableau 2 : erreur de mesure dans les niveaux de salaire - ERF97-EE97 et ERF97-EE98

Critère	Valeur	obs.	μ	σ	ρ	λ	γ	obs.	μ	σ	ρ	λ	γ
ERF97-EE97								ERF97-EE98					
Tous		11490	-0,05286	0,19065	-0,12438	0,86811	0,9665	11546	-0,02341	0,18977	-0,12022	0,86807	0,96346
Sexe	F	5195	-0,052	0,1972	-0,11855	0,87247	0,96431	5202	-0,02231	0,1919	-0,1098	0,87677	0,96369
	H	6295	-0,05356	0,18508	-0,14979	0,84563	0,9629	6344	-0,02431	0,18802	-0,15022	0,84072	0,95587
Répondant tiers	0	7911	-0,04879	0,19398	-0,12048	0,86762	0,96071	8403	-0,02146	0,18615	-0,11318	0,87534	0,96584
	1	3579	-0,06185	0,18276	-0,13384	0,86643	0,97707	3143	-0,02863	0,19906	-0,14142	0,84567	0,95434
Compléments salariaux	0	3796	-0,05234	0,20481	-0,11956	0,8866	0,98441	3801	-0,02517	0,21149	-0,12762	0,88086	0,98791
	1	7694	-0,05311	0,18327	-0,13568	0,84608	0,94921	7745	-0,02255	0,17816	-0,12156	0,85031	0,9416
Horaires irréguliers	0	6168	-0,05693	0,18177	-0,11452	0,86765	0,95314	6212	-0,02858	0,17884	-0,11491	0,87179	0,96171
	1	5322	-0,04814	0,20036	-0,14117	0,86441	0,98206	5334	-0,01739	0,20161	-0,1334	0,86024	0,96754
Temps de travail instable	?	1363						1354					
	0	6415	-0,05452	0,17717	-0,13925	0,85282	0,96272	6477	-0,02471	0,17175	-0,13113	0,86195	0,96766
	1	3712	-0,05518	0,19505	-0,11199	0,88267	0,96918	3715	-0,0266	0,20756	-0,11349	0,86701	0,95256
Temps plein	0	1251	-0,06086	0,24958	-0,12532	0,86128	0,95104	1260	-0,03485	0,23082	-0,09837	0,87695	0,95127
	1	10239	-0,05188	0,18214	-0,16474	0,83738	0,96592	10286	-0,02201	0,18408	-0,16587	0,83451	0,96256
Conditions de travail instables	?	59						56					
	0	10716	-0,05408	0,18411	-0,11475	0,87486	0,96582	10772	-0,0231	0,18785	-0,11489	0,86981	0,96036
	1	715	-0,04444	0,24078	-0,20776	0,81452	0,97544	718	-0,03433	0,1901	-0,14963	0,86767	0,99662
Secteur public	0	8199	-0,0409	0,20198	-0,12175	0,86653	0,96098	8251	-0,011	0,20059	-0,11636	0,86634	0,95685
	1	3291	-0,08265	0,15501	-0,12416	0,84156	0,93179	3295	-0,05449	0,15525	-0,12121	0,84616	0,93548
Imposable	0	2905	-0,01289	0,25184	-0,23081	0,78564	0,93719	2909	0,01658	0,2537	-0,22989	0,78629	0,94259
	1	8585	-0,06638	0,16268	-0,08063	0,8858	0,95009	8637	-0,03688	0,16047	-0,07294	0,88614	0,94347

Note : voir note du tableau 1.

Graphiques 1 : fonction de distribution des salaires et des erreurs de mesure - estimation non-paramétrique, 1996

Graphiques 2 : fonction de distribution des salaires et des erreurs de mesure - estimation non-paramétrique, 1997

des études américaines similaires. Bound et Krueger (1991) et Bound *et alii* (1994) trouvent en général une valeur moyenne de l'erreur positive et inférieure à 0,007 point de logarithme. Il est cependant possible que nos

résultats proviennent du fait que l'on compare le salaire de février au salaire moyen perçu au cours de l'année considérée et qu'on ne tient donc pas compte de la croissance infra-annuelle du salaire. Ce phénomène n'explique cependant pas l'ensemble de la sous-déclaration apparente. Lorsqu'on compare les déclarations fiscales aux salaires déclarés à l'enquête Emploi en février de l'année suivante, la valeur moyenne de l'erreur est plus faible mais reste négative, de l'ordre de -0,02 point de logarithme. En outre, si pour l'enquête Revenus Fiscaux on cherche la combinaison des salaires des enquêtes Emploi 96, 97 et 98 qui minimise l'écart (en valeur absolue) avec le montant fiscal déclaré, il apparaît que l'erreur reste en moyenne négative.

Les tableaux 1 et 2 présentent aussi la valeur moyenne de l'erreur de mesure pour différentes sous-populations. Dans la majorité des cas examinés, la valeur moyenne de l'erreur de mesure varie assez peu avec le sous-groupe considéré. On notera néanmoins des différences importantes pour les salariés du secteur public (État ou collectivités territoriales) ainsi que pour les individus non imposables. Pour les premiers, l'écart entre les salaires déclarés à l'enquête Emploi et ceux issus de l'enquête Revenus Fiscaux est en moyenne bien plus important que celui observé pour les autres salariés, alors que pour les seconds, l'écart est sensiblement plus faible, voire nul.

Le fait que les valeurs déclarées dans l'enquête Emploi soient en moyenne plus faibles que celles issues de l'enquête Revenus Fiscaux ne doit toutefois pas être sur-interprété. D'une part, il convient de remarquer que la valeur moyenne de l'erreur de mesure n'est jamais significativement différente de zéro. D'autre part, dans les applications économétriques, le fait que l'erreur soit de moyenne non-nulle n'aura d'effet que sur la valeur estimée de la constante. Pour les autres coefficients estimés, seules importent les différentes mesures de fiabilité discutées dans la première partie. Ces mesures sont présentées dans les tableaux 1 et 2.

Le ratio de fiabilité λ est pour l'ensemble des années et pour les échantillons complets de l'ordre de 0,85. Sous l'hypothèse d'erreurs de mesure classiques, le biais d'atténuation attendu dans les estimations utilisant les données de salaire de l'enquête Emploi comme variable explicative est donc de l'ordre de 15%. Toutefois, l'erreur de mesure est pour toutes les enquêtes négativement corrélée à la valeur du salaire déclarée dans l'enquête Revenus Fiscaux et il convient de remarquer que le coefficient de régression ρ (de l'ordre de -0,10) est toujours significativement différent de zéro. Les erreurs de mesure ne semblent donc pas vérifier l'hypothèse d'indépendance généralement retenue dans les travaux économétriques. En termes de biais attendus, l'existence d'une corrélation négative implique un biais vers le bas de l'ensemble des coefficients estimés lorsque le salaire est utilisé comme variable dépendante. Cette corrélation négative implique par ailleurs un biais d'atténuation plus faible que dans le cas d'erreur de mesure classique lorsque le salaire déclaré à l'enquête Emploi est utilisé comme variable explicative. Dans ce cas, le coefficient de régression γ du salaire Revenus Fiscaux sur le salaire enquête Emploi suggère un biais particulièrement faible du coefficient estimé lorsque le salaire enquête Emploi est utilisé comme variable explicative, de l'ordre de 5%⁽²⁵⁾.

Les résultats obtenus pour l'ensemble de notre échantillon sont assez proches de ceux obtenus dans des évaluations similaires menées dans le cas américain. Bound et Krueger (1991) font état d'une valeur de λ de l'ordre de 0,85 à partir de données annuelles appartenant à l'enquête sur la population active et registres de Sécurité Sociale. Bound *et alii* (1994), à partir des données du *PSID-Validation Study*, aboutissent quant à eux à une valeur de λ comprise entre 0,70 et 0,85. Par ailleurs, ces deux études font aussi état d'une corrélation négative entre l'erreur et la vraie valeur du salaire. Le coefficient de régression ρ de l'erreur sur la vraie valeur est de l'ordre de -0,20 pour les premiers et de -0,10 pour les seconds. En conséquence, leurs mesures de γ sont aussi assez proches des nôtres.

L'évaluation désagrégée sur différents sous-échantillons permet de préciser les facteurs affectant la précision des déclarations de salaire.

En tenant compte du sexe de l'individu, il apparaît que les salaires des femmes sont en général mieux déclarés que ceux des hommes. Le ratio de fiabilité λ pour les premières est plus élevé d'environ 3 à 4 points de pourcentage. Par ailleurs, l'erreur de mesure est aussi moins corrélée à la vraie valeur du salaire dans le cas des femmes. Ces deux écarts se compensent alors presque entièrement dans le calcul de γ : pour cet indicateur, les différences entre hommes et femmes sont assez faibles, même si elles restent à l'avantage des femmes.

La ventilation des résultats selon l'origine des déclarations (distinguant les déclarations faites à l'enquête par l'individu percevant le salaire de celles faites par un tiers) suggère que les erreurs de déclaration sont en moyenne plus importantes lorsque les réponses sont faites par un tiers. Toutefois, l'échantillon ERF97-EE98 mis à part, les valeurs du ratio de fiabilité ne font pas apparaître de différences notables dans la qualité des deux types de déclaration. Ce résultat est comparable à ceux obtenus par Mellow et Sider (1983) et Bound et Krueger (1991)⁽²⁶⁾.

A contrario, le fait de déclarer des compléments de rémunération ne modifie pas particulièrement l'erreur moyenne mais a tendance à altérer la fiabilité du salaire. Il est cependant important de noter que l'influence des compléments n'est ici que partiellement traitée. En particulier, les résultats ne prennent pas en compte le fait que certains individus ne déclarent pas de compléments salariaux alors qu'ils en perçoivent ou en percevront dans l'année. On peut par exemple noter que parmi les 25% des professeurs agrégés ou certifiés – catégorie de salariés encline à percevoir des compléments – qui sous-déclarent le plus, seuls 20% (au maximum sur l'ensemble des échantillons) déclarent des primes. En comparaison, ce pourcentage est au moins le double parmi les 25% qui sous-déclarent le moins.

Le fait de tenir compte des heures travaillées semble avoir un effet sur la qualité des réponses, même si l'effet mesuré n'est pas toujours facilement interprétable. De façon attendue, les données indiquent que les déclarations faites par les individus dont l'horaire de travail hebdomadaire est irrégulier sont en général moins précises que les déclarations faites par les individus soumis à des horaires plus réguliers. Mais la différence entre les deux groupes, qui est tout de même limitée, ne tient peut-être pas tant à des différences intrinsèques dans la qualité des réponses fournies qu'à une plus grande imprécision de notre procédure de calcul du salaire mensuel à partir des déclarations fiscales dans le cas d'individus soumis à des horaires variables. Plus étonnamment, les résultats sur la stabilité des heures de travail n'indiquent pas une meilleure qualité de réponse des individus qui déclarent le même nombre d'heures de travail à deux enquêtes consécutives. Par ailleurs, les individus déclarant travailler à temps complet dans deux enquêtes consécutives ont un ratio de fiabilité un peu plus faible que les autres. Cependant, en raison d'une corrélation plus importante entre l'erreur et la vraie valeur, le coefficient de régression du salaire Revenus Fiscaux sur le salaire enquête Emploi est plus important pour les personnes ayant travaillé à temps complet deux années de suite.

S'agissant de l'évolution des conditions de travail (travail le dimanche, le samedi et/ou la nuit) entre deux enquêtes consécutives, il apparaît (sauf sur l'échantillon ERF97-EE98) que la qualité des déclarations des individus ayant connu un changement significatif de conditions d'une année sur l'autre est nettement plus faible que pour les autres salariés. Sur l'échantillon ERF96-EE97, l'écart est d'environ 10 points de pourcentage pour λ et de 5 points pour γ .

Outre des écarts de mesure importants, les salariés du secteur public présentent des déclarations de moins bonne qualité que les autres salariés : la valeur de λ est plus faible et celle de ρ également ; les deux écarts ne se compensant que partiellement, la valeur de γ est elle aussi plus faible pour les salariés du secteur public. L'inclusion de prestations non imposables dans le traitement des fonctionnaires peut expliquer, tout au moins en partie, ces résultats. Il s'agit notamment des allocations familiales qui, contrairement au cas du régime général, ne sont pas versées par les CAF mais directement par l'État lors de la paye⁽²⁷⁾. En distinguant les individus susceptibles de percevoir des allocations familiales (ceux qui ont au moins deux enfants) de ceux qui ne peuvent pas en bénéficier, il apparaît que l'écart de qualité des déclarations entre secteur public et secteur privé est effectivement plus important pour les individus censés percevoir des allocations familiales (cf. tableau 3).

Enfin, les tableaux 1 et 2 indiquent que les individus appartenant à un foyer fiscal non imposable présentent des erreurs de mesure plus faibles en moyenne mais des biais d'atténuation plus élevés. Ces résultats montrent les limites de l'hypothèse selon laquelle les déclarations fiscales sont exemptes d'erreurs. Pour les foyers non imposables, une erreur dans la déclaration des revenus n'a (généralement) aucune conséquence fiscale ; on peut ainsi penser que ces foyers sont moins impliqués dans la déclaration et de ce fait omettent, par exemple, de déclarer certains compléments salariaux. Il est donc probable qu'une partie de la mauvaise qualité (relative) des résultats observés sur les foyers non imposables est à imputer au revenu fiscal.

Tableau 3 : erreur de mesure dans les niveaux de salaire, selon le nombre d'enfants et l'appartenance au secteur public - ERF96-EE96 et ERF96-EE97

Nombre d'enfants	Secteur public	obs.	μ	σ	ρ	λ	γ	obs.	μ	σ	ρ	λ	γ
		ERF96-EE96						ERF96-EE97					
≤ 1	0	2260	-0,02625	0,20465	-0,09366	0,86245	0,93228	2293	-0,01088	0,19776	-0,09566	0,87147	0,9458
	1	928	-0,07871	0,13177	-0,09813	0,88048	0,95997	925	-0,05064	0,15347	-0,13535	0,84771	0,95126
≥ 2	0	1829	-0,03854	0,17408	-0,07499	0,88872	0,94851	1831	-0,0196	0,17482	-0,08794	0,89077	0,96336
	1	744	-0,07723	0,17856	-0,17125	0,80738	0,92486	747	-0,06103	0,17782	-0,16145	0,81227	0,9233
		ERF97-EE97						ERF97-EE98					
		obs.	μ	σ	ρ	λ	γ	obs.	μ	σ	ρ	λ	γ
≤ 1	0	4715	-0,03657	0,21379	-0,1261	0,85537	0,9472	4731	-0,00755	0,20979	-0,12275	0,85883	0,95135
	1	1826	-0,07909	0,15775	-0,12473	0,84211	0,9331	1821	-0,05346	0,15919	-0,10834	0,84648	0,92431
≥ 2	0	3484	-0,04677	0,18467	-0,11516	0,88292	0,98067	3520	-0,01563	0,18743	-0,10692	0,87731	0,96444
	1	1465	-0,08708	0,15145	-0,12377	0,84096	0,93059	1474	-0,05575	0,15028	-0,13919	0,84571	0,95216

Erreurs de mesure dans les variations du salaire

Les résultats concernant la qualité des niveaux de salaire déclarés peuvent être complétés par l'examen de la qualité des variations de salaire individuelles mesurées à partir de deux enquêtes consécutives. L'intérêt d'une telle évaluation est d'autant plus grand que les estimations à partir de données de panel utilisent fréquemment des données en différence première pour tenir compte de la présence d'effets fixes individuels.

L'erreur dans les déclarations de variations de salaire peut être calculée comme l'écart entre la variation du log du salaire déclaré à l'enquête Emploi et la variation du log du salaire enregistré dans Revenus Fiscaux. Afin de tenir compte de la croissance infra-annuelle des salaires, nous prenons comme référence pour le salaire enquête Emploi de l'année fiscale t la moyenne du salaire déclaré dans l'enquête Emploi en février t et du salaire déclaré en février $t+1$ ⁽²⁸⁾.

Les différentes statistiques de fiabilité des variations de salaire calculées à partir de l'enquête Emploi sont présentées dans le tableau 4. En moyenne, les variations de salaire calculées sont assez proches de celles enregistrées dans les déclarations fiscales : la valeur moyenne de l'erreur sur l'échantillon le plus complet est de -0,0027 et n'est pas significativement différente de zéro.

Par contre, les différentes mesure de fiabilité des variations obtenues à partir de l'enquête Emploi indiquent une très faible qualité des données recueillies.

Ainsi, l'examen du ratio de fiabilité λ indique que la plus grande partie de la variance interindividuelle dans les taux de croissance des salaires calculés provient d'erreurs de mesure. Sous l'hypothèse d'erreurs de mesure classiques, la part des erreurs de mesure dans la variance totale observée des variations de salaire serait supérieure à 50%. Cette moindre qualité des données d'enquête en différence première par rapport aux données en niveau se comprend aisément. Il est en effet raisonnable de penser que la vraie valeur du salaire individuel sera fortement corrélée d'une année à l'autre. De ce fait, la variance de la différence première des salaires effectivement perçus sera relativement faible. *A contrario*, il n'y a pas de raison de penser que les erreurs de déclaration soient fortement corrélées d'une année à l'autre et la variance des erreurs de mesure en différence première représentera donc une part importante de la variance observée des variations de salaire.

Tableau 4 : erreur de mesure dans les variations de salaire - ERF9697-EE9698

Critère	Valeur	obs.	μ	σ	ρ	λ	γ
Tous		3701	-0,00266	0,19344	-0,96136	0,48193	0,25372
Sexe	F	1680	-0,00015	0,20169	-0,95342	0,48428	0,29467
	H	2021	-0,00474	0,18634	-0,96923	0,47968	0,21039
Répondant tiers	0	1913	-0,00084	0,19138	-0,95873	0,481	0,25546
	1	1788	-0,0046	0,19566	-0,96403	0,48287	0,25174
Compléments salariaux	0	1047	0,0001	0,21062	-0,95577	0,49236	0,37016
	1	2654	-0,00375	0,18626	-0,96444	0,47646	0,20928
Horaires irréguliers	0	1525	-0,00746	0,12114	-0,93041	0,44018	0,16931
	1	2176	0,00071	0,23097	-0,9662	0,48928	0,30333
Temps de travail instable	?	533					
	0	1669	-0,00421	0,17463	-0,99806	0,4748	0,01765
	1	1499	-0,00157	0,19703	-0,93622	0,48253	0,31896
Temps plein	0	350	0,02111	0,2858	-0,99209	0,47717	0,07094
	1	3351	-0,00514	0,18098	-0,95312	0,48335	0,28818
Conditions de travail instables	?	24					
	0	3228	-0,00174	0,18782	-0,9636	0,48168	0,24449
	1	449	-0,00865	0,22781	-0,95974	0,48385	0,27334
Secteur public	0	2562	-0,00552	0,20486	-0,95969	0,48148	0,25587
	1	1139	0,00379	0,16481	-0,96671	0,48362	0,24788
Imposable	0	777	-0,00729	0,23944	-0,96276	0,49464	0,38729
	1	2924	-0,00143	0,17927	-0,96062	0,47573	0,21784

Note : voir note du tableau 1.

L'hypothèse d'erreurs de mesure classiques ne semble par ailleurs pas vérifiée dans nos données. Le coefficient de régression ρ de l'erreur de mesure sur la vraie valeur obtenue dans l'enquête Revenus Fiscaux est toujours significativement négatif et très proche de -1. Compte tenu du mode de calcul retenu pour l'erreur de mesure, ce résultat indique aussi que le coefficient de régression de la variation du log du salaire de l'enquête Emploi sur la variation du log du salaire dans l'enquête Revenus Fiscaux est très faible. Ceci suggère la possibilité d'un important biais vers zéro de l'ensemble des coefficients estimés dans les équations utilisant la variation de salaire dans l'enquête Emploi comme variable dépendante.

Enfin, l'examen du coefficient γ permet d'évaluer l'ampleur des biais possibles lorsque les salaires de l'enquête Emploi en différence première sont utilisés comme variable explicative. Les résultats sont un peu meilleurs que dans le cas précédent mais le biais demeure très important : dans l'ensemble, les valeurs que nous obtenons suggèrent un biais d'atténuation de l'ordre de 75 à 80 %.

Nos résultats peuvent là encore être comparés à ceux obtenus dans des études américaines similaires. À partir de déclarations annuelles de salaire dans la *PSID-Validation Study*, Bound et alii (1994) trouvent que la part des erreurs de mesure dans la variance totale des variations du log du revenu est de l'ordre de 30%. Ils obtiennent par ailleurs un coefficient de régression de l'erreur sur la vraie valeur négatif, significatif mais notablement plus faible que celui obtenu ici (-0,2 contre -0,95 dans nos données). Dans le cas des variations annuelles du salaire horaire, ces mêmes auteurs trouvent une valeur de λ de l'ordre de 20% et une valeur du biais d'atténuation, lorsque les valeurs déclarées sont utilisées comme variables dépendantes, assez proche de celle obtenue ici. Par contre, même dans le cas des variations annuelles du salaire horaire, le coefficient de régression de l'erreur reste plus faible en valeur absolue que celui reporté dans le tableau 5.

À partir de déclarations annuelles de salaire en différence première issues des *Current Population Surveys* et des registres de Sécurité Sociale, Bound et Krueger (1991) aboutissent, pour la différence première du salaire, à une valeur de ρ de l'ordre de 0,65 et à une valeur de γ de l'ordre 0,77, indiquant une assez bonne fiabilité des données en différence première.

Nos résultats indiquent donc une qualité particulièrement faible des données de salaire de l'enquête Emploi en différence première. Pour partie, cette faiblesse s'explique sûrement par les critères de sélection de l'échantillon retenus pour mener à bien notre évaluation des erreurs de mesure. Le fait de se restreindre à un échantillon d'individus n'ayant pas connu de changement d'établissement ou de changement de profession conduit vraisemblablement à ne retenir que des individus dont les vraies variations de salaire sont plus limitées que dans l'ensemble de la population. On impose donc un critère de sélection de l'échantillon qui a pour effet de réduire la variance vraie des variations de salaire mais qui n'affecte pas, *a priori*, la variance des variations d'erreurs de mesure. De ce fait, la part des erreurs de mesure dans la variance totale des variations de salaire déclarée à l'enquête Emploi est vraisemblablement surestimée dans notre échantillon, de même que l'ampleur des biais d'atténuation.

Extensions

Effet des comportements d'arrondi

Parmi les différents phénomènes qui peuvent expliquer l'existence d'erreurs de mesure dans les déclarations de salaire, il est courant d'évoquer les comportements d'arrondi. Les individus déclareraient une valeur arrondie de leur véritable salaire⁽²⁹⁾. L'examen du graphique 3 semble confirmer *a priori* l'importance de ces comportements. Le graphique présente l'estimation de la densité des salaires déclarés pour l'année 1996. On s'est restreint aux individus ne déclarant pas de compléments salariaux non mensuels. La densité est estimée par *kernel* et on a délibérément choisi une fenêtre de lissage étroite⁽³⁰⁾. On voit alors apparaître des pics importants dans la distribution pour les valeurs multiples de 1000 francs. Des pics intermédiaires apparaissent aussi, notamment pour des valeurs multiples de 500 francs.

La fréquence des déclarations de valeurs "rondes", par exemple de valeurs multiples de 100, 500 ou encore 1000 francs, ne suffit toutefois pas, en elle-même, à établir l'importance des comportements d'arrondi. On ne peut en effet considérer qu'on est en présence d'un comportement d'arrondi que si deux conditions sont réunies : d'une part, la valeur déclarée est un multiple d'une certaine valeur, qui représente le pas d'arrondi ; d'autre part, l'écart entre la vraie valeur du salaire et la valeur déclarée est inférieur à la valeur du pas d'arrondi. En appliquant cette définition, il est possible de mesurer l'incidence des comportements d'arrondi dans les données de l'enquête Emploi. On restreint le champ d'analyse aux individus ne déclarant pas de compléments salariaux non

Graphique 3 : distribution des salaires déclarés à l'enquête Emploi 1996 - hors primes

mensuels⁽³¹⁾. On considère cinq pas d'arrondi possibles. La procédure de définition du pas d'arrondi consiste à retenir pour chaque observation le plus petit pas d'arrondi possible⁽³²⁾. Le tableau 5 donne la fréquence des comportements d'arrondi. Selon la vague d'enquête considérée, la fréquence des comportements d'arrondi imputés varie entre 42 et 35%. Environ 40% des valeurs arrondies le sont pour un pas de 500 francs. La valeur moyenne du pas d'arrondi relatif (pas d'arrondi minimal divisé par la valeur du salaire) varie de 10,6 % à 13,3%. Enfin, l'examen de la distribution des comportements d'arrondi par quintile de salaire (graphique 4) suggère que la fréquence des comportements d'arrondi varie assez peu en fonction du niveau de salaire mais que le pas d'arrondi augmente avec le salaire.

Une fois établie l'incidence des comportements d'arrondi, il convient de s'interroger sur leur contribution aux erreurs de mesure observées. Pour mesurer cette contribution, nous procédons de la façon suivante : lorsque la valeur déclarée peut être considérée comme une valeur arrondie, on assigne comme valeur du salaire dans

Graphiques 4 : fréquence des arrondis par quintile de revenu

Tableau 5 : fréquence des arrondis - en %

	ERF96		ERF97		ERF97		ERF97	
	EE96		EE97		EE97		EE98	
Aucun arrondi	65,66		57,17		63,77		58,06	
Arrondi à 50 FF	4,77	(51)	6,34	(50)	4,08	(56)	5,56	(50)
Arrondi à 100 FF	4,55	(62)	5,4	(58)	4,06	(64)	5,39	(48)
Arrondi à 500 FF	13,2	(70)	17,18	(59)	14,48	(69)	17,97	(58)
Arrondi à 1000 FF	7,45	(75)	8,97	(71)	8,5	(78)	8,9	(67)
Arrondi à 5000 FF	4,37	(78)	4,93	(72)	5,11	(82)	4,12	(77)
Pas relatif moyen	0,1258		0,1154		0,1331		0,1061	
Observation	2242		2351		4778		4752	

Note : les chiffres en parenthèses et en italique donnent, pour chaque pas d'arrondi, le pourcentage de valeurs arrondies par défaut.

Tableau 6 : effet des arrondis sur les erreurs de mesure

Fichiers		obs.	μ	σ	ρ	λ	γ
ERF96EE96	sans correction	2242	-0,051498	0,19722	-0,09878	0,88684	0,96901
	avec correction	2242	-0,041888	0,18861	-0,08505	0,89549	0,96656
ERF96EE96	sans correction	2351	-0,032628	0,2049	-0,11536	0,87973	0,97643
	avec correction	2351	-0,025737	0,19523	-0,09686	0,88959	0,9707
ERF96EE96	sans correction	4777	-0,060732	0,20873	-0,12349	0,87855	0,97827
	avec correction	4777	-0,049302	0,20075	-0,1093	0,88663	0,97448
ERF96EE96	sans correction	4751	-0,029575	0,21255	-0,13527	0,87376	0,98646
	avec correction	4751	-0,022911	0,20286	-0,12013	0,8837	0,98416

Note : voir note du tableau 1.

l'enquête Emploi la valeur déclarée dans l'enquête Revenu Fiscaux. L'effet des comportements d'arrondi sur la qualité des déclarations de salaire peut alors être mesuré en comparant la valeur des indicateurs de qualité des déclarations avant et après correction des arrondis. Les résultats, pour le sous-échantillon des individus ne déclarant pas de prime, sont donnés dans le tableau 6. La qualité mesurée des déclarations n'est pas sensiblement affectée par la correction des arrondis. Par ailleurs, l'effet de la correction des arrondis est ambigu. Sous l'effet des comportements d'arrondis, le coefficient de corrélation ρ augmente, ce qui tend à augmenter les biais lorsque le salaire est la variable expliquée d'un modèle économétrique. Par contre, la valeur de l'indicateur γ augmente légèrement : les biais d'erreurs de mesure sont alors diminués lorsque le salaire est utilisé comme variable explicative. Au total, la contribution des comportements d'arrondi aux erreurs de mesure semble donc faible et a des effets ambigus sur la qualité des déclarations de salaire.

Effets des erreurs de mesure sur la mobilité salariale

En dehors des biais possibles pour l'estimation de modèles économétriques à partir de données en différences premières, les erreurs de mesure dans les déclarations de variations de salaires sont aussi susceptibles d'introduire des biais dans d'autres types de travaux statistiques. Les salaires déclarés à différentes dates sont par exemple souvent utilisés dans l'étude de la mobilité salariale. Si une part importante des variations de salaires provient d'erreurs de mesure, alors on peut penser que l'utilisation des données de l'enquête Emploi conduira à surestimer la mobilité salariale. Toutefois, l'ampleur du biais dépendra aussi de l'étendue des erreurs de mesure comparativement à la variance de la distribution de salaire sous-jacente. À titre illustratif, nous présentons dans le tableau 7 des matrices de mobilité interdéciles calculées à partir des deux sources de données de salaire. La comparaison des deux matrices suggère que l'enquête emploi surestime l'étendue de la mobilité salariale. Les écarts entre les deux matrices de transition sont surtout marqués du décile 5 au décile 9.

Tableau 7 : matrices de mobilité salariale interannuelles

A - enquêtes Emploi											
décile EE97		décile EE98									
		1 (5533)	2 (6500)	3 (7041)	4 (7800)	5 (8558)	6 (9333)	7 (10500)	8 (11917)	9 (14500)	10
1	(5416)	72,4	22,1	2,6	1	1	0,5	0,3	0	0	0
2	(6280)	7,3	52,1	32,2	6,4	2	0	0	0	0	0
3	(6975)	1,9	9,7	50,5	27,8	8,6	0,3	0,3	0	0,8	0
4	(7583)	1	1,8	8,5	50,8	33,7	1,3	1	1	0,3	0,8
5	(8316)	0,3	1	1,8	9,2	61,5	10,7	12,7	2,3	0,2	0,2
6	(9200)	0,8	0	0	1,6	18,1	22,8	48	7,9	0,8	0
7	(10075)	0,3	0,6	0,3	0,8	7,2	6,9	47	32	4,4	0,6
8	(11500)	0,3	0	0	0	2,2	1,9	11,2	52,7	30,3	1,4
9	(14000)	0,3	0	0	0,3	0,5	0	1,1	11,8	67,7	18,3
10		0	0	0	0	0,3	0,3	0	0,5	6	92,9
B - enquêtes Revenus Fiscaux											
décile RF96		décile RF97									
		1 (5602)	2 (6565)	3 (7316)	4 (8077)	5 (8841)	6 (9748)	7 (10879)	8 (12326)	9 (15110)	10
1	(5452)	74,9	19,9	1,6	1,4	1,1	0,3	0,3	0	0,3	0,3
2	(6366)	6,2	55,6	32,2	3,8	1,2	0	0,6	0,3	0	0
3	(7107)	1,7	4,9	53,9	33,3	5,5	0,3	0,3	0	0	0
4	(7873)	0,7	1,7	6,7	51	38,3	0,7	0,7	0,3	0	0
5	(8580)	0,7	1,5	0,7	3,1	71,3	14,2	5,8	2,3	0,2	0,3
6	(9512)	0	0	0	0,7	10,7	31,3	52,7	4,7	0	0
7	(10556)	0	0,3	0,3	0,3	4,4	4,7	56,2	32,1	1,8	0
8	(11942)	0,7	0	0	0	0,5	0,2	5,4	67,3	24,8	1
9	(14627)	0	0,2	0,5	0	0,5	0	0,5	4,9	79,5	13,9
10		0	0	0,2	0,2	0,2	0	0,2	0,2	2,3	96,5

Note : les chiffres en parenthèses et en italique donnent la limite supérieure des déciles.

Conclusion

L'évaluation de la qualité des déclarations de salaire à l'enquête Emploi et des conséquences possibles pour l'estimation de relations économétriques apporte donc des résultats contrastés, selon qu'on examine les données en niveau ou en différence première.

La qualité des déclarations de salaire en niveaux se révèle particulièrement bonne, et ce, d'autant plus que la nature différente des données dans les deux sources utilisées (annuelles dans l'enquête Revenus Fiscaux et mensuelles dans l'enquête Emploi) laissait présager une sous-estimation de la qualité des déclarations à l'enquête Emploi.

En revanche, les données en différence apparaissent nettement plus bruitées, ce qui semblerait induire des biais économétriques importants. Les estimations de biais menées dans cette étude méritent toutefois, dans le cas des données en différence, d'être interprétées avec prudence en raison de l'imparfaite adéquation des deux sources de données utilisées et des contraintes qu'elle impose sur la définition de l'échantillon sur lequel sont menées nos estimations.

Notes

- (1) Voir par exemple sur ce point Griliches (1986).
- (2) Voir notamment Bound *et alii* (1994) et Duncan et Hill (1985).
- (3) Nordberg, Penttilä et Sandström (2001) procèdent à une analyse similaire à partir de l'appariement de données administratives finlandaises et des données de l'*European Community Household Panel survey*.
- (4) On peut en fait montrer analytiquement que le biais d'atténuation sera plus faible si et seulement si $-1/3 < \rho < 0$.
- (5) L'enquête Emploi a connu une refonte importante depuis 2002. La collecte des données a désormais lieu "en continu" durant toutes les semaines de l'année. Concernant l'information sur les revenus d'activité, quelques changements ont été introduits dans le questionnaire. La principale variable reste le salaire mensuel perçu, mais le traitement des primes est différent et l'enquête identifie précisément la nature des montants déclarés (salaire brut ou net). Les revenus tirés d'activités secondaires font maintenant l'objet de questions. Les revenus non-salariaux sont aussi déclarés, sur une base annuelle. Enfin, l'information sur la durée de travail est affinée. Ces changements sont susceptibles de modifier la qualité de la mesure des salaires dans l'enquête Emploi mais vraisemblablement de façon marginale.
- (6) Il s'agit des fichiers correspondant à la déclaration 2042 à l'impôt sur le revenu et des fichiers relatifs à la taxe d'habitation.
- (7) L'appariement des déclarations fiscales de l'année 1996 concerne le tiers médian de l'échantillon de l'enquête Emploi 97 ; pour les déclarations de revenus de 1997, il concerne les tiers médian et sortant de l'échantillon de l'enquête Emploi 98.
- (8) Les taux de réussite de l'appariement se réfèrent au tiers médian de l'échantillon de l'enquête Emploi 97.
- (9) En particulier, la documentation de l'ERF indique que les salaires moyens sont identiques pour les ménages appariés et les ménages non-appariés. Le seul biais de sélection dans l'appariement mis en avant dans la documentation concerne les ménages ayant déménagé récemment et les ménages dont la personne de référence est étudiante, pour lesquels le taux d'appariement est plus faible. Notons toutefois que les critères de sélection d'échantillon imposés dans notre étude (voir *infra*) conduisent vraisemblablement à exclure la plupart de ces types de ménage de notre analyse.
- (10) Pour les ménages non-imposables, on peut *a priori* s'interroger sur la validité des déclarations fournies. En outre, la déclaration de revenu peut être exacte au niveau de l'ensemble du foyer fiscal mais inexacte au niveau individuel : permutation de cases dans la déclaration fiscale, cumul des revenus de plusieurs personnes sur une seule case. Ceci serait sans incidence sur le calcul du montant global d'impôt du foyer fiscal mais affecterait nos résultats. On notera toutefois que la qualité des déclarations fournies par les célibataires ne semble pas supérieure à celle du reste de notre échantillon.
- (11) L'examen des valeurs imputées permettrait seulement de juger de la qualité de la procédure d'imputation et non de la qualité des déclarations.
- (12) On divise le revenu fiscal d'activité par le nombre de mois d'activité au cours de l'année fiscale.
- (13) Pour une discussion de l'incidence des erreurs de mesure dans l'analyse de la pauvreté on pourra consulter Lollivier et Verger (2002).
- (14) On exploite ici les variables PRIMFR et PRIMM donnant les valeurs en francs et en mois de salaire des compléments salariaux.
- (15) Le rabais excédentaire est égal à la différence entre la valeur de l'action au moment de la souscription de l'option et le prix d'exercice de l'option consenti au salarié. Les plus-values tirées de l'exercice de l'option ne sont pas, quant à elles, incluses dans la déclaration fiscale de revenu *d'activité*.
- (16) Il est cependant possible que des individus employés durant tous les mois de l'année fiscale considérée aient aussi connu des épisodes de chômage partiel. Par ailleurs, cette restriction n'exclut pas la perception d'indemnités de maladie ou de pensions de préretraite.
- (17) Il existe très peu d'informations dans l'enquête Emploi permettant de mesurer l'influence de ce facteur sur nos résultats. Une des rares informations disponibles concerne le fait d'occuper un logement à titre gratuit. Il s'avère (résultats non reproduits dans l'article) que la qualité des déclarations de salaire des individus hébergés à titre gratuit ne diffère pas de celle du reste de l'échantillon. On soulignera toutefois que la mise à disposition gratuite du logement, mesurée dans l'enquête, n'est pas nécessairement le seul fait de l'employeur mais peut aussi être le fait de proches ou de parents. Elle ne signale donc pas de façon non ambiguë la perception d'un avantage en nature dans l'emploi occupé.
- (18) Pour les seuls individus présents deux années consécutives.
- (19) Ont été exclus les individus dont les ratios des salaires de deux années consécutives est compris entre 9 et 11 et entre 90 et 110. Il s'agit d'individus pour lesquels on peut suspecter l'omission ou l'ajout d'un 0 dans l'enregistrement du salaire.
- (20) Nous retenons la nomenclature des professions à quatre chiffres.
- (21) Voir annexe.

- (22) Sur ce point, on pourra notamment consulter Magnac et Visser (1998) qui notent l'existence d'erreurs importantes dans les déclarations mensuelles rétrospectives de statut sur le marché du travail, dans l'enquête Emploi.
- (23) On gardera toutefois à l'esprit que les résultats de l'annexe s'appliquent aux ménages présent à deux enquêtes Emploi consécutives. Ceci conduit notamment à exclure des individus dont l'insertion au marché du travail est particulièrement précaire : ainsi, 20% des travailleurs pauvres de l'échantillon apparié Emploi97-Revenus Fiscaux 96 n'ont pas répondu à l'enquête Emploi 98 (Lagarenne et Legendre, 2000).
- (24) Dans le cas de l'enquête Emploi, les distributions présentées sont donc plus lisses que la distribution des données brutes. La figure 3 présente une distribution moins lissée qui fait apparaître des pics importants. Voir *infra*.
- (25) On rappelle qu'une valeur de ρ négative conduit à une valeur de γ plus faible que la valeur de λ .
- (26) Nous avons également étudié les résultats par numéro d'enquêteur. Certains enquêteurs présentent des erreurs de mesure très élevées (en valeur absolue) mais aucun n'est assez influent pour modifier l'analyse. On se reportera à l'annexe pour la présentation du critère utilisé pour mesurer l'influence d'un groupe d'observations sur les résultats.
- (27) Notons que cette singularité n'existe plus depuis le 1^{er} janvier 2005.
- (28) Si au lieu d'utiliser le salaire moyen de deux enquêtes Emploi adjacente on compare directement les variations de salaire entre EE96 et EE97 ou EE97 et EE98 aux variations observées entre ERF96 et ERF98, la qualité des déclarations apparaît, pour tous les indicateurs, plus faible que celle mise en évidence dans le tableau 4.
- (29) Pour une analyse sur données américaines, voir Pischke (1995).
- (30) La largeur de la fenêtre est de 25.
- (31) La multiplicité des sources de déclarations du salaire (salaire mensuel, montant annuel des primes, valeur relative des primes en mois de salaire complémentaires) pour les individus déclarant des primes non mensuelles rend plus délicate l'étude des comportements d'arrondi.
- (32) Compte tenu des pas retenus, toute observation considérée comme un arrondi pour un pas donné sera aussi considérée comme un arrondi pour tout pas supérieur.
- (33) La profession est ici définie à partir du code profession à 4 chiffres.
- (34) La statistique de Cook a été calculée après suppression des assistantes maternelles, pour toutes les autres professions de l'échantillon.

Bibliographie

- Bound J. et Krueger A. B. (1991).** “The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?”, *Journal of Labor Economics* 9(1), pp. 1-24.
- Bound J., Brown C., Duncan G. J. et Rodgers W. L. (1994).** “Evidence on the Validity of Cross-Sectional and Longitudinal Labor Market Data”, *Journal of Labor Economics* 12(3), pp. 345-368.
- Cook D. (1979).** “Influential Observations in Linear Regression”, *Journal of the American Statistical Association* 74(365), pp.169-74.
- Duncan G. J. et Hill D. H. (1985).** “An Investigation of the Extent and Consequences of Measurement Error in Labor-Economic Survey Data”, *Journal of Labor Economics* 3(4), pp. 508-532.
- Griliches Z. (1986).** “Economic Data Issues”, in Z. Griliches et M. Intriligator (eds), *Handbook of Econometrics*, vol. 3, North-Holland, Amsterdam, chapter 25, pp. 1465-1514.
- Lagarenne C. et Legendre N. (2000).** “Les travailleurs pauvres en France : facteurs individuels et familiaux”, *Économie et statistiques*.
- Lollivier S. et Verger D. (2002).** “Erreurs de mesure et entrées-sorties de pauvreté”, *Journées de Méthodologie Statistique*, INSEE.
- Magnac T. et Visser, M. (1998).** “Transition Models with Measurement Error”, *CREST working paper*.
- Mellow W. et Sider H. (1983).** “Accuracy of Response in Labor Market Surveys : Evidence and Implications”, *Journal of Labor Economics* 1(4), pp. 331-344.
- Nordberg L., Penttilä I. et Sandström, S. (2001).** “A Study of the Effects of Using Interview Versus Register Data in Income Distribution Analysis with an Application to the Finnish ECHP-Survey in 1996”, *Chintex working paper*.
- Pischke J.-S. (1995).** “Measurement Error and Earnings Dynamics: Some Estimates from the PSID Validation Study”, *Journal of Business and Economics Statistics* 13(3), pp. 305-314.

Effet des critères de sélection de l'échantillon

Compte tenu des différents critères de restriction de l'échantillon imposés dans la présente étude, il convient de s'interroger sur la représentativité des résultats présentés au regard de l'ensemble de la population salariée. L'effet des différents critères de sélection de l'échantillon sur le nombre d'individus et sur les mesures de la qualité des déclarations à l'enquête Emploi est documenté dans le tableau 8. Pour chaque appariement d'enquête, on applique séquentiellement les différents critères de sélection. À chaque étape, on évalue la qualité des déclarations pour les individus satisfaisant le critère considéré (valeur 0) et pour les individus ne le satisfaisant pas (valeur 1). L'échantillon initial est constitué des individus présents, pour l'année fiscale t , aux vagues t et $t+1$ de l'enquête Emploi, pour lesquels la déclaration de revenu fiscal d'activité est disponible et déclarant un salaire à au moins une des vagues de l'enquête Emploi.

Si l'on compare, dans chaque sous-partie de la table, la première et l'avant-dernière ligne, il apparaît que les différents critères de sélection de l'échantillon conduisent à exclure environ 50% des observations. On constate par ailleurs que la qualité mesurée des déclarations de l'enquête Emploi est plus élevée pour l'échantillon final que pour l'échantillon initial : la valeur de λ passe d'environ 0,81 à environ 0,86 et la valeur de γ passe d'environ 0,91 à environ 0,96. Ce résultat est conforme aux attentes, dans la mesure où ces critères ont pour but d'assurer une plus grande comparabilité des variables de revenu des deux sources. L'effet des différents critères sur la qualité mesurée est toutefois relativement limité, ceci d'autant plus que les différences de concept de revenu entre les enquêtes Emploi et Revenu Fiscaux conduisent de toute évidence, pour l'ensemble de l'échantillon, à sous-estimer la qualité des déclarations à l'enquête Emploi.

L'effet de chaque critère de sélection, considéré isolément, va généralement dans le sens attendu. La qualité des déclarations est particulièrement mauvaise pour les individus qui n'ont pas été employés continûment au cours de l'année fiscale. La valeur de l'erreur pour ces individus indique des revenus fiscaux en moyenne plus élevés que les revenus déclarés à l'enquête Emploi, ce qui provient vraisemblablement du fait que les indemnités de chômage, lorsqu'elles sont perçues, sont incluses dans la déclaration fiscale de revenu d'activité. Pour des raisons similaires, la coïncidence entre déclaration fiscale et déclaration de salaire à l'enquête Emploi est assez mauvaise pour les individus exerçant des activités secondaires rémunérées. Là encore, les revenus déclarés à l'enquête Revenus Fiscaux sont en général plus élevés que les revenus de l'emploi principal enregistrés dans l'enquête Emploi. À l'inverse, les rémunérations déclarées à l'enquête Emploi par les assistantes maternelles sont en moyenne plus élevées que celles de la déclaration fiscale, du fait des possibilités de déduction fiscale spécifiques à cette profession (voir *infra*). Le critère d'absence de changement de profession conduit quant à lui à exclure un pourcentage étonnamment élevé d'individus (plus d'un quart de l'échantillon), alors qu'on ne se fonde ici que sur les trois premiers chiffres de la nomenclature des professions. Il convient de s'interroger sur d'éventuelles imprécisions dans l'enregistrement des professions qui pourraient conduire à surestimer les changements de profession effectifs. On notera toutefois que les résultats varient peu lorsqu'on exclut les individus pour lesquels l'enquête enregistre un changement de profession. Pour finir, le cas des individus employés sur une forme contractuelle autre qu'un CDI ou un emploi public titulaire est plus ambigu. Pour ces individus, la corrélation entre l'erreur de mesure et la valeur déclarée à l'enquête Emploi est plus forte (en valeur absolue) que pour le reste de l'échantillon, ce qui aggrave les biais d'estimation lorsque le salaire figure comme variable expliquée de l'estimation économétrique. Par contre, le biais d'atténuation est lui aussi plus faible, ce qui conduit à une valeur plus élevée de l'indicateur γ et donc à un biais plus faible lorsque le salaire est utilisé comme variable explicative.

Analyse de l'influence des différentes professions sur les résultats obtenus

Il est possible que les résultats soient influencés par certaines professions particulières, qu'il s'agisse de professions dont les revenus sont difficiles à mesurer dans l'enquête Emploi ou de professions pour lesquelles le concept de revenu enquêté dans l'enquête Emploi diffère fortement de celui retenu pour la déclaration fiscale.

Pour examiner ce problème, on étudie l'influence de chaque profession sur la valeur estimée du coefficient γ . Rappelons que cet indicateur est égal au coefficient de régression de la vraie valeur de la variable sur la valeur déclarée à l'enquête. On peut donc utiliser les outils statistiques permettant de détecter, dans les modèles de régression, les observations dites "influentes". On retient ici la statistique de Cook (Cook 1979), que l'on calcule pour chaque profession présente dans notre échantillon⁽³³⁾. La statistique de Cook est une mesure normalisée du changement dans la valeur du vecteur de coefficients estimé lorsqu'on exclut de l'échantillon une ou plusieurs observations. Sous l'hypothèse que les observations exclues suivent le même modèle statistique que le reste de l'échantillon, la statistique de Cook suit asymptotiquement une loi de Fisher $F(p, n-p)$ où p est le nombre de coefficients estimés et n le nombre d'observations. Pour les tailles d'échantillon dont on dispose, on sera amené à rejeter l'hypothèse nulle que les observations exclues suivent le même modèle statistique dès lors que la valeur de la statistique de Cook est supérieure ou égale à 2,3 (respectivement 3) pour un risque de première espèce de 10% (respectivement 5%).

Tableau 8 : effet des critères de sélection sur la taille de l'échantillon et l'étendue des erreurs de mesure

Critère de sélection	Valeur	obs.	μ	σ	λ	ρ	γ	obs.	μ	σ	λ	ρ	γ
Échantillon initial Changement d'établissement Changement de profession Emploi secondaire CDD Problème RF Problème EE Assistante maternelle	0	10934	-0,06088	0,27165	-0,136	0,80793	0,89466	11058	-0,0355	0,26614	-0,14351	0,81224	0,90716
	1	568	-0,3212	0,47612	-0,30398	0,69339	0,83432	585	-0,25066	0,47636	-0,32961	0,68064	0,82764
	0	9021	-0,04525	0,23534	-0,14048	0,83476	0,93733	9099	-0,02493	0,23427	-0,14378	0,83655	0,94315
	1	1345	-0,05572	0,32039	-0,10395	0,79609	0,85482	1374	-0,01383	0,2988	-0,11153	0,80565	0,87261
	0	7209	-0,0452	0,23808	-0,14515	0,83705	0,94524	7254	-0,0252	0,24265	-0,14579	0,83325	0,94020
	1	1812	-0,04547	0,22416	-0,11767	0,82334	0,90104	1845	-0,02388	0,19792	-0,13433	0,85314	0,95814
	0	6923	-0,0384	0,22481	-0,14981	0,84617	0,96375	6975	-0,01861	0,2258	-0,14983	0,84673	0,96462
	1	286	-0,20981	0,42159	-0,09694	0,7748	0,82338	279	-0,18996	0,47836	-0,11129	0,7267	0,77045
	0	5954	-0,04295	0,19761	-0,11338	0,86433	0,95314	5996	-0,02267	0,19988	-0,11738	0,86396	0,95656
	1	969	-0,01045	0,34691	-0,27354	0,7851	0,99974	979	0,00625	0,34344	-0,26416	0,78633	0,98981
Échantillon initial Activité continue Changement d'établissement Changement de profession Emploi secondaire CDD Problème RF Problème EE Assistante maternelle	0	5810	-0,04269	0,19375	-0,11343	0,86705	0,95692	5847	-0,02311	0,19725	-0,11886	0,86464	0,95899
	1	144	-0,05327	0,31707	-0,11429	0,80258	0,87056	149	-0,00537	0,28487	-0,08469	0,84977	0,90858
	0	5805	-0,04305	0,19177	-0,11443	0,86944	0,96121	5839	-0,02424	0,19116	-0,12007	0,87182	0,97028
	1	5	0,37766	0,9666	2,04732	0,0944	0,20747	8	0,80088	1,11442	-2,42611	0,05805	-0,11524
	0	5723	-0,04492	0,18056	-0,10704	0,86857	0,95276	5755	-0,02607	0,18034	-0,11688	0,8702	0,96473
	1	87	0,10416	0,58683	-0,22169	0,67731	0,75341	92	0,16204	0,63862	-0,08249	0,66591	0,68639
	ERF97-EE97												
	0	23454	-0,0594	0,27666	-0,15477	0,8114	0,91491	23560	-0,0254	0,27249	-0,15702	0,81325	0,92016
	1	22854	-0,05164	0,25588	-0,14941	0,83082	0,93903	22940	-0,01902	0,25335	-0,14645	0,83086	0,93668
	0	600	-0,35496	0,63964	-0,38212	0,58845	0,66074	620	-0,26176	0,62434	-0,48311	0,59416	0,72108
Échantillon initial Activité continue Changement d'établissement Changement de profession Emploi secondaire CDD Problème RF Problème EE Assistante maternelle	0	18779	-0,05019	0,24106	-0,1599	0,83392	0,95387	18913	-0,02201	0,23702	-0,14716	0,83676	0,94598
	1	4075	-0,05833	0,31526	-0,12279	0,81883	0,89908	4027	-0,00494	0,3187	-0,14402	0,81024	0,90469
	0	14469	-0,04979	0,24944	-0,17042	0,8308	0,95959	14556	-0,02094	0,24611	-0,15745	0,83184	0,94859
	1	4310	-0,05152	0,21051	-0,11643	0,84648	0,93154	4357	-0,0256	0,20372	-0,10575	0,85657	0,93546
	0	13933	-0,04309	0,23703	-0,17663	0,83845	0,97872	14021	-0,01388	0,23362	-0,16402	0,83984	0,96787
	1	536	-0,22404	0,43358	-0,12758	0,77806	0,84694	535	-0,20599	0,42713	-0,10841	0,77969	0,83659
	0	11973	-0,04862	0,21000	-0,14623	0,85718	0,974	12024	-0,01866	0,2113	-0,13944	0,8546	0,96406
	1	1960	-0,00929	0,35877	-0,2864	0,77129	0,986	1997	0,01494	0,33683	-0,25605	0,78319	0,97283
	0	11617	-0,0495	0,20816	-0,14662	0,85761	0,97494	11673	-0,01976	0,20808	-0,13526	0,85561	0,96115
	1	356	-0,01978	0,26174	-0,13344	0,84584	0,94668	351	0,01791	0,29737	-0,21439	0,83382	1,01958
Échantillon initial Activité continue Changement d'établissement Changement de profession Emploi secondaire CDD Problème RF Problème EE Assistante maternelle	0	11602	-0,04972	0,20532	-0,14745	0,861	0,981	11660	-0,02073	0,20111	-0,13648	0,86389	0,97458
	1	15	0,12184	0,99335	1,13454	0,09689	0,16954	13	0,85211	1,41146	1,51117	0,05491	0,11827
	0	11383	-0,0536	0,18844	-0,12962	0,85948	0,96255	11441	-0,02412	0,18689	-0,1258	0,85975	0,95905
	1	234	0,14966	0,62045	-0,32624	0,69228	0,83778	232	0,19494	0,64093	-0,18449	0,6622	0,71348

Le tableau 9 présente les valeurs de la statistique de Cook par profession et par ordre décroissant de la statistique de Cook⁽³⁴⁾. Les résultats montrent une valeur particulièrement élevée de la statistique de Cook pour les assistantes maternelles, ce qui nous amène à exclure cette profession de l'analyse. La singularité des assistantes maternelles s'explique par un régime de déclaration fiscale particulier. La législation fiscale autorise en effet les assistantes maternelles agréées à déduire de leur rémunération, dans leur déclaration fiscale, une somme forfaitaire égale à 3 fois le montant horaire du Smic, par jour et pour chacun des enfants qui leur sont confiés. Ceci explique donc vraisemblablement les écarts importants entre les déclarations aux deux enquêtes et le fait, mis en évidence dans le tableau 8, que les déclarations à l'enquête Emploi sont en moyenne nettement plus élevées que celles faites à l'administration fiscale.

Deux autres professions ont une valeur de la statistique de Cook sensiblement plus fortes que les autres, même si les valeurs observées ne conduisent pas à rejeter l'hypothèse nulle pour les seuils usuels. Il s'agit des "Professeurs agrégés et Certifiés" et des "Employés de maison". Nous avons comparé les résultats présentés dans l'article à ceux obtenus lorsque l'on supprime ces deux professions de nos échantillons. L'effet est négligeable : les mesures de qualité des déclarations varient au plus de 1,5%. Les résultats discutés dans cet article ne dépendent donc pas de l'inclusion de quelques professions particulières dont les déclarations de revenu seraient nettement différentes dans les deux sources ou dont les revenus seraient difficiles à appréhender dans l'enquête Emploi.

Tableau 9 : analyse de l'influence des professions : valeurs les plus élevées du ratio de Cook

Code	Intitulé	Observation	Ratio de Cook
ERF96-EE96			
5631	Assistantes maternelles, gardiennes d'enfants	44	2,30785
3411	Professeurs agrégés et certifiés	139	0,86744
3734	Cadres commerciaux des PME	33	0,47724
6411	Conducteurs routiers et grands routiers (salariés)	105	0,28713
6415	Conducteurs-livreurs, coursiers (salariés)	94	0,19435
5411	Secrétaires	281	0,15057
6891	Nettoyeurs	66	0,11760
4211	Instituteurs	121	0,11670
5512	Vendeurs en alimentation	85	0,10825
5633	Concierges, gardiens d'immeubles	26	0,07353
5311	Agents de police	50	0,06945
5632	Employés de maison et femmes de ménage chez des particuliers	38	0,05169
5212	Préposés des PTT	38	0,04694
6515	Magasiniers	60	0,04290
6299	Ouvriers qualifiés divers de type industriel	14	0,03892
ERF96-EE97			
5631	Assistantes maternelles, gardiennes d'enfants	42	2,15359
3411	Professeurs agrégés et certifiés	142	0,87136
5632	Employés de maison et femmes de ménage chez des particuliers	41	0,49786
6411	Conducteurs routiers et grands routiers (salariés)	112	0,3183
5512	Vendeurs en alimentation	87	0,16913
6891	Nettoyeurs	67	0,15091
5411	Secrétaires	275	0,14243
5611	Serveurs et commis de restaurant ou	46	0,11235
6415	Conducteurs-livreurs, coursiers (salariés)	96	0,10888
3731	Cadres de l'exploitation des magasins de vente	16	0,09169
3727	Cadres administratifs ou financiers	56	0,06284
4621	Chefs de petite surface de vente (salariés ou mandataires)	9	0,05179
5311	Agents de police	51	0,04805
4625	Représentants en biens de consommation	14	0,04344
6299	Ouvriers qualifiés divers de type industriel	16	0,03049
ERF97-EE97			
5631	Assistantes maternelles, gardiennes d'enfants	112	8,03872
5632	Employés de maison et femmes de ménage auprès des particuliers	107	1,88378
3411	Professeurs agrégés et certifiés	332	0,9322
6411	Conducteurs routiers et grands routiers (salariés)	248	0,63327
5411	Secrétaires	497	0,23566
5633	Concierges, gardiens d'immeubles	48	0,22714
4628	Acheteurs non classés cadres, aide-acheteurs	15	0,17333
4211	Instituteurs	249	0,16926
5512	Vendeurs en alimentation	174	0,11846
6891	Nettoyeurs	148	0,10770
6415	Conducteurs-livreurs, coursiers (salariés)	154	0,08847
4633	Assistants techniques de la réalisation des spectacles	3	0,05151
5311	Agents de police	93	0,04404
5221	Aide-soignants (du public ou du privé)	250	0,04324
3415	Enseignants de l'enseignement supérieur	25	0,03721
ERF97-EE98			
5631	Assistantes maternelles, gardiennes d'enfants	109	4,79013
5632	Employés de maison et femmes de ménage auprès des particuliers	110	1,84618
3411	Professeurs agrégés et certifiés	321	1,58720
6411	Conducteurs routiers et grands routiers (salariés)	246	0,55285
5633	Concierges, gardiens d'immeubles	45	0,21354
5411	Secrétaires	490	0,19349
3822	Ingénieurs et cadres de bureau d'étude	27	0,18303
4514	Personnels administratifs de catégorie B de l'État	75	0,18228
4211	Instituteurs	254	0,14279
6415	Conducteurs-livreurs, coursiers (salarié)	160	0,07729
5221	Aide-soignants (du public ou du privé)	249	0,06013
5611	Serveurs et commis de restaurant ou de café	104	0,05833
5519	Caissiers de magasin	118	0,05480
5514	Vendeurs en équipement de la personne, articles de sport	42	0,05414
5311	Agents de police	93	0,05296