ÉTENDUE ET CONSÈQUENCES DES ERREURS DE MESURE DANS LES DONNÉES D'ENQUÊTE

Cyrille HAGNERÉ[(*)], Arnaud LEFRANC[(**)]

[(*)] THEMA et OFCE
[(**)] THEMA et Université de Cergy-Pontoise

Introduction

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 retrospectives, ... Beaucoup d'études économetriques 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 auxiliaires fournissant la vraie valeur des variables enquêtées. Pour procéder à un tel examen, deux voies de recherche ont été empruntées dans les travaux existants.

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 dans 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. 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.

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 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

[1] Voir par exemple sur ce point Grilliches [6].

limité 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.

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 del' enquête Revenus Fiscaux. 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 section 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 section présente les données utilisées. Les troisièmes et quatrièmes sections examinent la qualité des niveaux de salaires et des taux de croissance des salaires déclarés. Deux résultats principaux émergent alors 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 enquête 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-jacente 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.

1. 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 variable X et Y , avec :

$$Y = Y^* + v$$
$$X = X^* + u$$

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

L'estimateur de β par moindre 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.

1.1. 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 :

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

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 :

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

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.

1.2. 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 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. Le terme d'erreur v peut en effet s'écrire :

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

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

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

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

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 :

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

Le biais est donc plus faible si l'erreur de mesure est négativement corrélée à X^*,

c'est à dire si $\rho < 0$ [2].

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

$$b = \beta \mathrm{cov}(X, X^*)/V(X)$$

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 autre variables de la régression.

1.3. 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 $1/(1+V(u)/V(X^*))$ (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 $\mathrm{cov}(X, X^*)/V(X)$.

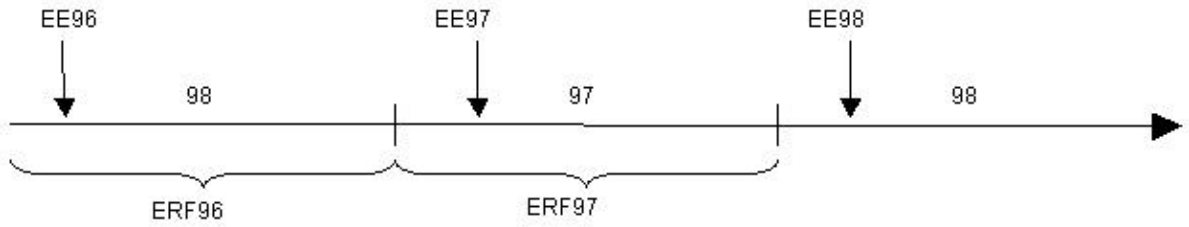
Dans la suite de l'article, ces différents indicateurs sont estimés sur les données de l'échantillon apparié enquête Emploi – Revenus Fiscaux.

2. Données utilisées

Les données utilisées proviennent des vagues 96, 97 et 98 de l'enquête Emploi (EE) et des vagues 96 et 97 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 des ménages enquêtés sont interrogés sur leur situation sur le marché du travail ainsi que sur les revenus salariaux d'activité. L'enquête Emploi est réalisée au mois de mars de chaque année. Par ailleurs, l'échantillon de l'enquête Emploi est renouvelé annuellement par tiers, ce qui permet de suivre les individus enquêtés au cours de trois années consécutives.

Les enquêtes Revenus Fiscaux de 96 et 97 sont, quant à elles, issues de l'appariement des fichiers de déclarations fiscales de la Direction Générale des Impôts[3] et d'une partie de l'échantillon de l'enquête Emploi de l'année correspondante. Les informations recueillies dans les enquêtes 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. Le graphique ci-dessous résume le calendrier des 5 vagues d'enquêtes utilisées dans cette étude.

[2] On peut en fait montrer analytiquement que le biais d'atténuation sera plus faible si et seulement si –1/3 < ? < 0.
[3] 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.



A 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, 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. 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 :

$$u = \log(\text{salaire EE}) - \log(\text{salaire ERF})$$

En pratique, le calcul des erreurs de mesure se heurte cependant à plusieurs difficultés. La première tient aux différences dans le champ des variables de salaire enquêtée dans l'enquête Emploi et déclarée dans les sources fiscales. La principale variable de salaire disponible dans l'enquête Emploi est le salaire courant ou habituel dans l'emploi principal. A contrario, les données fiscales enregistrent l'ensemble des revenus salariaux perçus, qu'ils soient issus de l'activité principale ou d'éventuelles activité secondaires [4]. Deux sources d'information complémentaires permettent cependant, au sein de l'enquête Emploi, de se rapprocher du champ de la variable issue de Revenu Fiscaux. D'une part, les individus de l'enquête Emploi sont interrogés sur la perception éventuelle de primes et compléments salariaux non-mensuels et sur le montant de ces primes. Il est donc possible de calculer dans l'enquête Emploi un salaire incluant les primes et correspondant mieux au concept de Revenus Fiscaux. En outre, les individus de l'enquête Emploi sont aussi interrogés sur l'exercice éventuel d'activités secondaires. Notre étude est donc restreinte aux individus déclarant ne pas exercer d'activité secondaire. Ces deux amendements aux données de l'enquête Emploi permettent alors d'assurer la comparabilité a priori des champs de la variable de salaire disponible dans nos deux sources de données.

La seconde limite à la comparabilité stricte des données issues des deux sources tient aux différences dans la période de référence utilisée pour les déclarations de salaire. Les déclarations à l'enquête Revenus Fiscaux se rapportent aux revenus perçus au cours de l'année fiscale considérée. Dans l'enquête Emploi, seuls les compléments de rémunération non-mensuels et les primes sont déclarés sur une base annuelle. A contrario, dans l'enquête Emploi, le salaire principal 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, au salaire du mois de février de l'année d'enquête.

Du fait de cette différence dans la période de référence, un écart entre les deux sources de données est susceptible d'apparaître même en l'absence d'erreurs de déclarations dans l'enquête Emploi. Plusieurs restrictions de l'échantillon utilisé dans l'évaluation permettent cependant de limiter l'incidence des différences dans la période de référence.

On se restreint tout d'abord aux individus ayant été employés continûment au cours de l'année fiscale considérée.[5] Par ailleurs, afin de se restreindre à un ensemble d'individus dont le salaire mensuel en

[4] La variable utilisée dans Revenu Fiscaux est la somme des traitements et salaires annuels déclarés. On n'utilise pas ici les imputations de salaire effectuées pour les ménages de l'enquête Emploi non appariés aux sources fiscales. On notera par ailleurs que le montant de salaire des déclaration fiscales inclut la CSG non-déductible alors que cette dernière n'est pas incluse dans les déclarations à l'enquête Emploi. On a donc soustrait la CSG non-déductible de la déclarations fiscale afin de rendre comparable les données des deux sources.

[5] Notons que la variable utilisée dans Revenus Fiscaux inclut d'éventuelles indemnités chômage. La restriction de l'échantillon aux individus qui ont connu une activité toute l'année permet d'assurer que les revenus déclarés correspondent uniquement à des salaires.

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 au cours de cette même période ainsi que les individus employés sous une forme contractuelle autre que le contrat à durée indéterminée. Ces deux restrictions permettent alors de réduire la variabilité infra-annuelle des rémunérations mensuelles et de rendre plus comparables les deux sources. Il demeure toutefois possible que la durée du travail se modifie au cours de l'année. De ce fait, le salaire perçu au mois de février ne sera pas nécessairement représentatif du salaire perçu au cours des autres mois de l'année. Nous examinerons donc l'incidence de ce facteur en étudiant séparément la qualité des déclarations des individus ne déclarant pas une même durée de travail aux deux vagues d'enquête Emploi encadrant l'enquête Revenus Fiscaux et pour les individus déclarant avoir des horaires de travail irréguliers.

Au total, les variables de revenu utilisées pour évaluer les erreurs de mesure sont : dans le cas de l'enquête Revenus Fiscaux, la valeur des salaire traitements annuels nets de la CSG non-déductible en équivalent mensuel ; pour l'enquête Emploi, la somme du salaire mensuel déclaré à l'enquête Emploi et de l'équivalent mensuel des compléments salariaux et primes perçues sur une base non-mensuelle. L'échantillon est restreint aux individus n'exerçant pas d'activité secondaire, ayant été employés continûment en contrat à durée indéterminée au cours de l'année fiscale considérée et n'ayant pas changé d'établissement. En outre, 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.

Il conviendra enfin 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 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 nos mesures du salaire pourraient subsister. En outre, notre procédure doit aussi être interprétée comme 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 déclarations dans les montants salariaux, des erreurs de déclarations dans les variables de sélection de l'échantillon pourraient conduire à un écart entre les deux enregistrements des niveaux de salaire[6].

3. 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 la figure 1 et les tableaux 1 à 4.

L'examen des distributions de salaires obtenues dans Emploi et Revenus Fiscaux indique que les valeurs faibles du salaire mensuel (entre 5 000 et 10 000 francs) sont plus représentées dans l'enquête Emploi que dans 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 quelque 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.

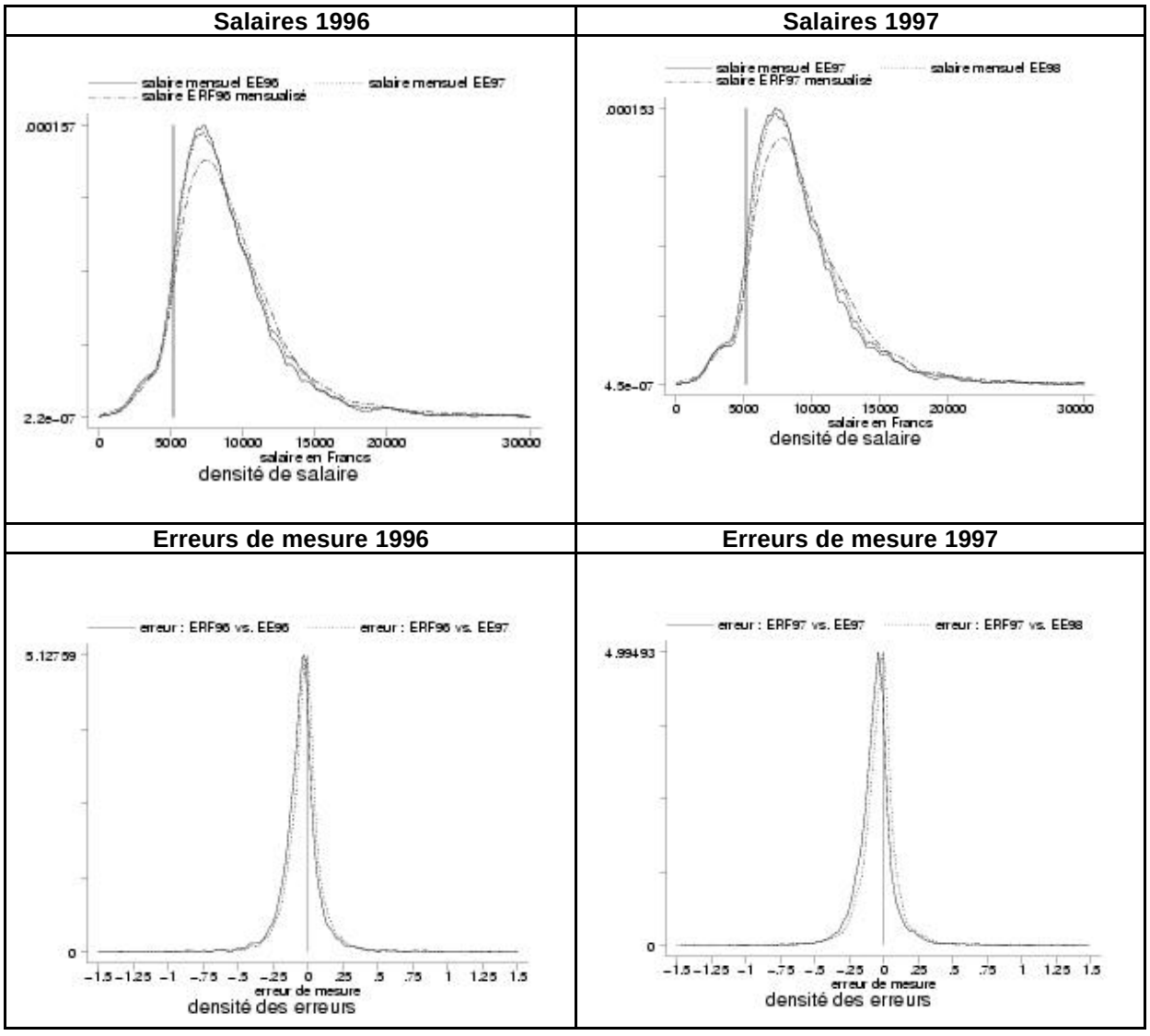
En moyenne, les déclarations à l'enquête Emploi tendent à sous-estimer la valeur du salaire mensualisé. Lorsqu'on compare l'enquête Revenus Fiscaux à l'enquête Emploi de la même année la valeur moyenne de l'erreur de l'ordre de -0.05 points de logarithme. Sur ce point nos résultats contrastent avec ceux obtenus dans des études américaines similaires. Bound et Krueger [4] et Bound et al. [3] trouvent en général une valeur moyenne de l'erreur positive et inférieure à 0.007 points de logarithme. Il est cependant possible que nos résultats proviennent du fait que l'on compare le salaire

[6] Sur ce point, on pourra consulter Magnac et Visser [7] qui notent l'existence d'erreurs importantes dans les déclarations mensuelles rétrospectives de statut sur le marché du travail de l'enquête Emploi.

de février au salaire moyen perçu au cours de l'année considérée. 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 points 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 à 4 présentent aussi la valeur moyenne de l'erreur de mesure pour différentes sous-populations. La valeur moyenne de l'erreur de mesure varie assez peu avec le sous-groupe considéré sauf éventuellement dans le cas des agents des administrations nationales pour lesquels l'écart entre les salaires déclarés à l'enquête Emploi et ceux issus de Revenus Fiscaux est en moyenne plus important que pour les autres groupes.

Figure 1 - densité des salaires et des erreurs de mesure



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 par 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 0. 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 section 1. Ces mesures sont présentées dans les tableaux 1 à 4.

Le ratio de fiabilité λ est pour l'ensemble des années et pour les échantillons complets de l'ordre de 0.83. 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 (de l'ordre de -0.15) est toujours significativement différent de zéro. Les erreurs de mesure ne semblent donc pas vérifier l'hypothèse d'indépendance généralement faite 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 du salaire Revenus Fiscaux sur le salaire enquête Emploi (dernière colonne) suggère un biais particulièrement faible du coefficient estimé lorsque le salaire enquête Emploi est utilisé comme variable explicative, de l'ordre d'environ 5%.

Les résultats obtenus pour l'ensemble de notre échantillon sont assez proches de ceux obtenus dans des évaluation similaires menées dans le cas américain. Bound et Krueger [4] font état d'une valeur de λ de l'ordre de 0.85 à partir de données annuelles appariant enquête sur la population active et registres de Sécurité Sociale. Bound et al. [3] à partir des données du PSID-Validation Study aboutissent quant à eux à une valeur de ? 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 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 $\mathrm{cov}(X, X^*)/V(X)$ 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.

Le fait de tenir compte de l'origine des déclarations, en distinguant les déclarations faites à l'enquête par l'individu percevant le salaire de celles faites par un tiers, suggère que les réponses faites par un tiers sont moins précises que les réponses faites directement par la personne recevant le salaire déclaré. Si l'étendue de la sous-estimation dans l'enquête Emploi est comparable pour les deux types de déclaration, le ratio de fiabilité est en général (sauf pour 1996) plus élevé pour les auto-déclarations (de l'ordre de 0.85) que pour les déclarations par des tiers (de l'ordre de 0.82). Ce résultat diffère quelque peu de ceux obtenus par Mellow et Sider [8] et Bound et Krueger [4] qui ne trouvent pas de différences notables dans la qualité des deux types de déclaration.

Tableau 1 - erreurs de mesure ERF96-EE96

	valeur	n	Mean	sd	lambda	rho	cov(x,x*)/v(x)
Tous		4882	-0.0447	0.1933	0.8423	-0.1385	0.9465
Rdq	0	2804	-0.0446	0.1902	0.8495	-0.1339	0.9524
	1	2078	-0.049	0.1975	0.8315	-0.1459	0.9378
Sexe	H	2603	-0.0409	0.1931	0.8108	-0.1629	0.9224
	F	2279	-0.0491	0.1934	0.8551	-0.1379	0.9646
Prime	0	1434	-0.044	0.2075	0.8584	-0.1275	0.9589
	2	2640	-0.0457	0.1719	0.8308	-0.138	0.9293
fonction publique	0	4395	-0.0412	0.1992	0.8373	-0.1405	0.9411
	1	382	-0.0782	0.1113	0.8887	-0.0511	0.9276
temps complet	0	654	-0.0567	0.2151	0.8657	-0.0533	0.9029
	1	4228	-0.0429	0.1897	0.8095	-0.1987	0.9563
heures stables	0	1657	-0.0477	0.2171	0.8363	-0.1147	0.9161
	1	3225	-0.0432	0.1799	0.8444	-0.1568	0.9685
horaires irréguliers	0	2838	-0.0455	0.1939	0.8351	-0.1347	0.9324
	1	2033	-0.0439	0.193	0.8494	-0.1464	0.9651

Note : **rdq** 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 ; **prime** vaut 0 pour les individus déclarant ne pas toucher des compléments salariaux non-mensuels dans les deux enquêtes Emploi utilisées et 2 pour les individus déclarant percevoir des compléments salariaux non-mensuels dans les deux enquêtes ; **fonction publique** vaut 1 pour les personnes employées dans une administration nationale ; **temps complet** vaut 1 pour les personnes déclarant un horaire hebdomadaire habituel égal à 39 heures dans les deux enquêtes ; **heures stables** vaut 1 pour les individus dont l' horaire hebdomadaire habituel ne varie pas d'une enquête à l'autre ; **horaires irréguliers** vaut 1 pour les individus déclarant ne pas avoir d'horaires de travail réguliers.

Tableau 2 - erreurs de mesure ERF96-EE97

	valeur	n	mean	sd	lambda	rho	cov(x,x*)/v(x)
tous		4882	-0.0234	0.1921	0.8439	-0.1515	0.9621
rdq	0	2804	-0.0236	0.1845	0.8571	-0.1389	0.9687
	1	2078	-0.023	0.202	0.825	-0.1705	0.9522
sexe	H	2603	-0.0223	0.1876	0.8195	-0.1745	0.9476
	F	2279	-0.0246	0.1972	0.8503	-0.1502	0.9704
prime	0	1434	-0.0204	0.2091	0.8565	-0.1481	0.9777
	2	2640	-0.0249	0.1701	0.8339	-0.1447	0.9401
fonction publique	0	4395	-0.0198	0.1978	0.8393	-0.1545	0.9580
	1	382	-0.0547	0.1155	0.8812	-0.0455	0.9145
temps complet	0	654	-0.0416	0.2224	0.8577	-0.1101	0.9409
	1	4228	-0.0205	0.1869	0.814	-0.2042	0.9704
heures stables	0	1657	-0.0323	0.2081	0.8475	-0.1369	0.9526
	1	3225	-0.0188	0.1833	0.8395	-0.1618	0.9661
horaires irréguliers	0	2838	-0.0229	0.1944	0.8344	-0.1436	0.9397
	1	2033	-0.0244	0.1895	0.8541	-0.164	0.9919

Tableau 3 - erreurs de mesure ERF97-EE97

	valeur	n	mean	sd	lambda	rho	cov(x,x*)/v(x)
tous		9441	-0.0522	0.2043	0.831	-0.1718	0.9633
rdq	0	5622	-0.0495	0.2001	0.8432	-0.1559	0.9656
	1	3819	-0.0561	0.2104	0.8102	-0.1991	0.9578
sexe	H	5017	-0.0555	0.1968	0.8156	-0.1928	0.9605
	F	4424	-0.0485	0.2125	0.8304	-0.1675	0.9578
prime	0	2932	-0.0533	0.2171	0.8518	-0.1637	0.9878
	2	4934	-0.0484	0.1891	0.8139	-0.1929	0.9575
fonction publique	0	8620	-0.0497	0.2092	0.8278	-0.1751	0.9615
	1	679	-0.0786	0.1359	0.852	-0.0998	0.9241
temps complet	0	1286	-0.0519	0.2757	0.8008	-0.1644	0.9083
	1	8155	-0.0522	0.1907	0.8079	-0.2151	0.9720
heures stables	0	3396	-0.0521	0.2211	0.8447	-0.1563	0.9684
	1	6045	-0.0522	0.1943	0.8173	-0.188	0.9581
horaires irréguliers	0	5678	-0.0577	0.1787	0.854	-0.1198	0.9451
	1	3745	-0.0443	0.2374	0.8034	-0.2405	0.9945

Tableau 4 - erreurs de mesure ERF97-EE98

	valeur	n	mean	sd	lambda	rho	cov(x,x*)/v(x)
tous		9441	-0.0223	0.2061	0.8286	-0.1725	0.9601
rdq	0	5622	-0.0186	0.2022	0.8404	-0.1595	0.9651
	1	3819	-0.0279	0.2116	0.8084	-0.1944	0.9497
sexe	H	5017	-0.0249	0.1975	0.8146	-0.1933	0.9593
	F	4424	-0.0195	0.2155	0.8265	-0.1693	0.9534
prime	0	2932	-0.0256	0.2179	0.8509	-0.1727	0.9969
	2	4934	-0.0184	0.1905	0.8116	-0.1846	0.9450
fonction publique	0	8620	-0.0191	0.2109	0.8255	-0.1749	0.9576
	1	679	-0.0542	0.1396	0.8451	-0.1091	0.9231
temps complet	0	1286	-0.0333	0.2799	0.7959	-0.1907	0.9249
	1	8155	-0.0206	0.1918	0.806	-0.214	0.9671
heures stables	0	3396	-0.0226	0.2312	0.8326	-0.1596	0.9530
	1	6045	-0.0222	0.1906	0.823	-0.1863	0.9660
horaires irréguliers	0	5678	-0.0287	0.1797	0.8527	-0.1294	0.9526
	1	3745	-0.0132	0.2403	0.7995	-0.2308	0.9747

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 pour ces derniers est en général plus faible de 2 à 3 points de pourcentage que celui calculé sur l'échantillon des femmes. Cette différence résulte à la fois d'une part moins importante de l'erreur de mesure dans la variance totale des salaires déclarés par les femmes et d'une corrélation plus faible entre l'erreur de mesure et la vraie valeur du salaire perçu.

Par ailleurs, alors que l'examen de la moyenne des erreurs semblait indiquer une sous-déclaration dans l'enquête Emploi plus prononcée pour les salariés de la fonction publique que pour les autres individus, les ratios de fiabilité indiquent que les salaires déclarés dans l'enquête Emploi par les fonctionnaires sont en général plus fortement corrélés aux valeurs déclarées dans Revenu Fiscaux que pour les autres salariés. Les valeurs du ratio de fiabilité sont pour ce sous-échantillon les plus élevées, de l'ordre de 0.9 pour ERF96 et 0.95 pour ERF97.

Enfin 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. La différence entre les deux groupes ne tient d'ailleurs 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. Par contre, notre procédure de mensualisation des déclarations fiscales annuelles ne donne pas des résultats plus proches des déclarations mensuelles dans le cas des individus déclarant la même durée de travail hebdomadaire à deux enquêtes consécutives que dans le cas des autres individus. De même, les déclarations faites par les individus déclarant travailler à temps complet dans deux enquêtes consécutives ont en général un ratio de fiabilité un peu plus faible que pour les autres individus.

4. 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.

Tableau 5 - erreurs de mesure dans les variations de salaire ERF9697 - EE9697

	valeur	n	mean	sd	lambda	rho	cov(x,x*)/v(x)
tous		3176	-0.0028	0.1921	0.4816	-0.9724	0.2100
rdq	0	1714	-0.0002	0.2068	0.4833	-0.9808	0.1785
	1	1462	-0.0059	0.1734	0.4788	-0.9581	0.2434
sexe	H	1663	-0.0042	0.1619	0.4808	-0.9692	0.2175
	F	1513	-0.0012	0.2207	0.4821	-0.9742	0.2052
prime	0	821	0.0015	0.2206	0.4910	-0.9760	0.2839
	2	1560	0.0001	0.1774	0.4769	-0.9741	0.1742
fonction publique	0	2862	-0.0030	0.1974	0.4817	-0.9727	0.2090
	1	223	0.0067	0.1388	0.4802	-0.9635	0.2350
temps complet	0	454	-0.0055	0.2801	0.4816	-0.9519	0.2786
	1	2722	-0.0024	0.1732	0.4816	-0.9815	0.1634
heures stables	0	1487	-0.0013	0.2111	0.4805	-0.9557	0.2608
	1	1689	-0.0042	0.1737	0.4831	-0.9937	0.0758
horaires irréguliers	0	1513	-0.0080	0.1214	0.4555	-0.9427	0.1849
	1	1657	0.0019	0.2393	0.4874	-0.9785	0.2272

Les différentes statistiques de fiabilité des variations de salaire calculées à partir de l'enquête Emploi sont présentées dans le tableau 5. 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 -0.0028 et n'est pas significativement différente de 0.

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 ? 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 déclarations 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 variation de salaire.

L'hypothèse d'erreurs de mesure classiques ne semble par ailleurs pas vérifiée dans nos données. Le coefficient de régression 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 Revenus Fiscaux est très faible. Ceci suggère la possibilité d'un important biais vers 0 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 de régression de la variation du log du salaire de Revenus Fiscaux sur la variation du log du salaire dans l'enquête Emploi 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. A partir de déclarations annuelles de salaire dans le PSID-Validation Study, Bound et al. [3] 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. (-.2 contre -.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.

A partir de déclarations annuelles de salaire en différence première issues des Current Population Surveys et des registres de Sécurité Sociale, Bound et Krueger [4] aboutissent pour la différence première du salaire, à une valeur de λ de l'ordre de 0.65 et à une valeur de $cov(X,X^*)/V(X)$ 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.

Tableau 6 - matrice de mobilité salariale enquêtes Emploi

	décile EE 98									
décile EE 97	1	2	3	4	5	6	7	8	9	10
1	89.56	10.44	0	0	0	0	0	0	0	0
2	8.36	73.07	15.48	1.55	0.31	0.62	0.31	0.31	0	0
3	2.19	14.37	66.88	14.69	1.88	0	0	0	0	0
4	0	0.94	13.13	65.31	18.75	1.25	0.31	0.31	0	0
5	0.31	0.31	2.82	18.18	61.44	15.67	0.63	0.31	0	0.31
6	0	0.31	0.63	0.63	15.94	64.69	16.56	0.94	0	0.31
7	0	0	0	0.31	0.63	17.24	66.14	15.05	0.63	0
8	0	0	0.31	0.31	0.63	0.63	16.3	65.83	15.67	0.31
9	0	0	0	0	0.31	0	0	17.39	72.67	9.63
10	0	0	0	0	0	0	0	0	8.18	91.82

Tableau 7 - matrice de mobilité salariale enquêtes Revenus Fiscaux

	décile ERF 97									
décile ERF 96	1	2	3	4	5	6	7	8	9	10
1	83.23	11.39	2.53	1.58	0	0.32	0.32	0	0.63	0
2	10.38	70.13	15.72	2.2	0.63	0.63	0.31	0	0	0
3	1.57	13.21	65.72	16.04	2.2	0.63	0.31	0	0.31	0
4	1.58	2.21	14.51	64.04	14.51	3.15	0	0	0	0
5	0.63	2.2	1.26	13.52	63.21	16.67	1.57	0.31	0.31	0.31
6	0.31	0.31	0	1.57	16.35	62.58	17.61	1.26	0	0
7	0.32	0.63	0	1.26	2.21	14.83	65.3	14.51	0.95	0
8	0.31	0	0	0	0.31	0.63	13.21	70.13	14.78	0.63
9	0	0.31	0	0.31	0.31	0	1.26	13.84	77.67	6.29
10	0	0	0	0.32	0.32	0	0.32	0	5.71	93.33

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. A titre illustratif, nous présentons dans les tableaux 6 et 7 des matrices de mobilité interdéciles calculées à partir des deux sources de données de salaire. La comparaison des deux matrices suggère que les biais dans l'évaluation de la mobilité salariale restent limités. Les écarts entre les deux matrices de transition sont surtout marqués aux deux extrêmes de la distribution de salaires. Dans le bas de la distribution de salaires, l'enquête Emploi semble sous-estimer l'étendue de la mobilité salariale. Ceci est particulièrement marqué pour le premier décile. Par contre, la mobilité dans le haut de la distribution de salaires semble surestimée dans l'enquête Emploi, comparativement aux données de l'enquête Revenus Fiscaux.

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 Revenus Fiscaux et mensuelles dans 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.

Bibliographie

[1] Abowd, J. et Card, D., "On the Covariance Structure of Hours and Earnings Changes, *Econometrica*, vol 57, no 2, pp 411-455, 1989

[2] Altonji, Joseph, "Intertemporal Labor Supply : Evidence from Microdata, *Journal of Political Economy*,1986,

[3] Bound, J., Brown, C., Duncan, G. J. et Rodgers, W. L., "Evidence on the Validity of Cross-Sectional and Longitudinal Labor Market Data", *Journal of Labor Economics*, vol 12, no 3, pp 345-368 ,1994

[4] Bound, John et Krueger, Alan B., "The Extent of Measurement Error in Longitudinal Earnings Data: Do Two Wrongs Make a Right?", *Journal of Labor Economics*, vol 9, no 1, pp 1-24, 1991

[5] Duncan, Greg J. et Hill, Daniel H., "An Investigation of the Extent and Consequences of Measurement Error in Labor-Economic Survey Data, *Journal of Labor Economics*, vol 3, no 4, pp 508-532, 1985

[6] Griliches, Zvi, "Economic Data Issues", in Griliches, Zvi et Intrilligator, Michael, *Handbook of Econometrics*, vol 3, chapitre 25, pp 1465-1514, 1986.

[7] Magnac, Thierry et Visser, Michael, "Transition Models with Measurement Errors", *Review of Economics and Statistics*, Vo 81, no 3, August 1999, pp 466-474.

[8] Mellow, Wesley et Sider, Hal, "Accuracy of Response in Labor Market Surveys: Evidence and Implications", *Journal of Labor Economics*, Vol 1, no4, October 1983, pp 331-344.

[9] Pischke, Jörn-Steffen, "Measurement Error and Earnings Dynamics: Some Estimates from the PSID Validation Study", *Journal of Business and Economics Statistics*, vol 13, no 3, pp 305-314, july 1995,