L'appariement statistique des bases de données SILC, HBS et HFCS : aspects méthodologiques et applications à l'étude de la pauvreté et des inégalités

Huyen TRAN[1]
STATEC Luxembourg et Luxembourg Income Study (LIS)

Guillaume OSIER[2]
STATEC Luxembourg et Luxembourg Income Study (LIS)

TITLE

Statistical matching between the SILC, HBS and HFCS data: methodological aspects and application to the analysis of poverty and social inequalities

RÉSUMÉ

Les techniques d'appariement statistique connaissent depuis plusieurs années un regain d'intérêt parmi les producteurs de statistiques sociales. Cet article présente les résultats d'un exercice d'appariement réalisé à partir des bases de données sur les revenus (Statistics on Income and Living Conditions – SILC), les dépenses (Household Budget Survey – HBS) et le patrimoine des ménages (Household Finance and Consumption Survey – HFCS) au Luxembourg. Les aspects méthodologiques sont abordés et trois exemples d'indicateurs obtenus à partir des données appariées sont présentés : le taux de pauvreté multidimensionnelle basé sur le revenu, la consommation et le patrimoine, le taux d'épargne et le taux de précarité énergétique. Si l'appariement statistique offre une solution à bas coût pour produire des indicateurs sophistiqués à partir d'informations disponibles dans plusieurs sources de données, ces méthodes restent cependant basées sur de la modélisation et doivent donc être employées avec discernement, en tenant compte des hypothèses sous-jacentes.

Mots-clés *: appariement statistique, revenus, consommation, patrimoine, pauvreté, épargne, énergie.*

ABSTRACT

Statistical matching techniques have been gaining interest for several years among the producers of social statistics. This article presents the outcome from a statistical matching exercise between household income (Statistics on Income and Living Conditions - SILC), household expenditure (Household Budget Survey - HBS) and household wealth (Household Finance and Consumption Survey - HFCS) data in Luxembourg. Methodological aspects are addressed, and three examples of indicators based on the fused dataset are presented: the multidimensional poverty rate based on income, consumption and wealth, the saving rate, and the energy poverty rate. Although statistical matching offers a cheap and convenient solution to yield sophisticated indicators using information available from several databases, these methods are model-based and must yet be used with caution, taking into account the main underlying assumptions.

Keywords: *statistical matching, income, consumption, wealth, poverty, saving, energy.*

1. thi.tran@eib.org
2. Guillaume.Osier@statec.etat.lu

1. Introduction

L'appariement statistique (*Statistical matching* en anglais) est une technique connue depuis longtemps, mais qui connait un regain d'intérêt depuis une vingtaine d'années dans le contexte de la modernisation des statistiques sociales au niveau européen. Des initiatives internationales, comme le rapport Stiglitz-Sen-Fitoussi (2009) sur la mesure de la performance économique et du progrès social, ont notamment mis en avant l'importance d'exploiter davantage des micro-données issues des enquêtes sur les ménages et les personnes en complément des statistiques macroéconomiques plus classiques comme le Produit Intérieur Brut. La technique de l'appariement statistique offre une solution à bas coût pour l'intégration des différentes sources de données disponibles et la construction d'indicateurs complexes prenant en compte la multi-dimensionnalité des phénomènes sociaux.

Malgré ses avantages, l'appariement reste néanmoins une technique qui repose essentiellement sur la modélisation des relations entre les variables dans les différentes bases de données. La validité de la méthode repose sur des hypothèses sous-jacentes qu'il est souvent compliqué de valider, en particulier l'hypothèse d'indépendance conditionnelle (*Conditional Independence Assumption* – CIA). Cet aspect ne doit surtout pas être oublié lorsque l'on travaille sur une base de données appariées, qui ne sont pas des données réelles collectées sur le terrain mais plutôt des données reconstruites *ex post*.

Cet article présente un cas concret d'appariement de données statistiques sur la population résidente du Luxembourg, dans lequel les micro-données des enquêtes SILC (*Statistics on Income and Living Conditions*) sur les revenus des ménages, HBS (*Household Budget Survey*) sur la consommation et HFCS (*Household Finance and Consumption Survey*) sur le patrimoine sont fusionnées en une unique source de données. Des indicateurs nouveaux, qu'il n'aurait pas été possible de construire auparavant, peuvent alors être calculés. Cet article présente trois exemples d'indicateurs ainsi obtenus à partir de la distribution jointe du revenu, de la consommation et du patrimoine :

- le taux de pauvreté multidimensionnelle,
- le taux d'épargne des ménages et
- le taux de pauvreté énergétique.

Des estimations sont présentées pour l'ensemble de la population ainsi que pour certaines sous-populations d'intérêt définies d'après des caractéristiques comme l'âge, le genre, le niveau de vie ou encore le statut d'occupation du logement. La validité de l'appariement sera également évoquée.

2. Rappels sur l'appariement statistique

L'appariement est une technique statistique plutôt ancienne, dont on trouve déjà des traces dans les années 70 (Ruggles, 1974 ; Kadane, 1978). Depuis les années 2000, une littérature abondante a de nouveau émergé autour de l'utilisation de cette technique dans le contexte des statistiques sociales au niveau européen (D'Orazio, 2010). Un cadre théorique rigoureux a également été décrit dans d'Orazio (2006).

Dans ses fondements, l'appariement statistique est une approche assez intuitive qui permet de fusionner deux bases de données à partir de critères communs à ces deux bases. Si l'on considère deux bases de données A et B, A étant la « donneuse » et B la « receveuse », et un ensemble de variables $\boldsymbol{X}$ de nature qualitative ou quantitative et communes à A et B, l'appariement aboutit à la construction d'une base synthétique qui correspond en fait à la base « receveuse » B enrichie de l'information disponible dans la base « donneuse » A. Comme A et B ne couvrent généralement pas les mêmes unités, les variables $\boldsymbol{Y}$ qui se trouvent dans la base A mais pas dans la base B se

retrouvent dans la base finale synthétique sous la forme d'une imputation $\hat{\boldsymbol{Y}}$ de la valeur réelle que l'on obtiendrait sur les unités de la base *B* si l'information était réellement collectée. Comme pour les variables communes ***X***, le cadre des méthodes d'appariement est suffisamment large pour permettre de traiter aussi bien des variables quantitatives que qualitatives.

Tableau 1 – *Principe de l'appariement de deux bases de données A et B*

Base *A* « donneuse »	Base *B* « receveuse »	Base synthétique
Y		
X	***X***	***X***
	Z	***Z***, $\hat{\boldsymbol{Y}}$

La validité de l'approche précédente repose sur l'hypothèse dite d'indépendance conditionnelle (*Conditional Independence Assumption* – CIA), qui postule que la corrélation entre les variables ***Y*** et ***Z*** s'explique uniquement au travers d'un ensemble de variables ***X*** observables et communes aux deux bases de données (D'Orazio, 2010). Autrement dit, il n'y a pas de lien causal direct entre les variables ***Y*** et ***Z***.

Cette hypothèse CIA est cependant rarement vérifiée dans la pratique et, de toute façon, ne peut pas être testée à partir des seules bases de données qui entrent en considération. Néanmoins, des techniques ont été développées (Rässler, 2004) pour tester la sensibilité des résultats de l'appariement au relâchement de cette hypothèse. En outre, des approches alternatives (Paass, 1986) intègrent la possibilité d'utiliser des informations auxiliaires lors de l'appariement afin de contourner l'hypothèse d'indépendance conditionnelle.

La mise en œuvre d'un appariement statistique requiert un travail préalable de réconciliation des bases de données. Cela signifie qu'elles doivent couvrir les mêmes populations d'intérêt et se référer aux mêmes périodes de temps. Si besoin, des ajustements sont réalisés pour permettre la mise en conformité des données, par exemple *via* la suppression d'unités « hors-champ » ou encore l'ajustement de certaines valeurs. Pour une description plus détaillée des étapes préalables de réconciliation des bases de données à partir d'exemples concrets, on pourra consulter Eurostat (2013).

3. Détermination des variables de l'appariement statistique

L'hypothèse fondamentale d'indépendance conditionnelle montre que le choix des variables partagées ***X*** est déterminant pour garantir un bon niveau de précision aux résultats de l'appariement. Ces variables doivent être disponibles dans les deux bases de données *A* et *B*, et collectées d'après les mêmes concepts et définitions. Par exemple, si l'on utilise l'âge de la personne, celui-ci doit être calculé au même point[3] dans chacune des deux bases. De la même façon, on ne pourra pas utiliser le statut d'activité d'une personne si d'un côté celui-ci a été obtenu à partir de réponses auto-déclarées tandis que de l'autre il a été déterminé à partir de la nomenclature utilisée par l'Organisation Internationale du Travail (OIT) et qui est reprise dans l'enquête européenne sur les forces de travail. Il faudra aussi veiller à ce que les variables impliquées dans l'appariement ne souffrent pas d'erreurs de mesure ou encore de la présence d'un nombre élevé de valeurs manquantes ou de valeurs imputées.

3. Généralement l'âge peut être calculé en nombre d'années révolues (c'est-à-dire au 31 décembre de l'année dernière), au moment de l'enquête ou à la fin de l'année de l'enquête.

Les variables communes ***X*** doivent en outre avoir des distributions proches dans les deux bases de données (Eurostat, 2013). Cette condition doit permettre d'assurer une meilleure robustesse aux résultats de l'appariement. La comparaison des distributions entre *A* et *B* peut se faire en utilisant des fonctions de distance comme la distance de Hellinger :

$$HD(V, V') = \sqrt{\frac{1}{2}\sum_{i=1}^{K}\left(\sqrt{P(V=i)} - \sqrt{P(V'=i)}\right)^2},$$

où V et V' désignent deux vecteurs de covariables. Une formule analogue existe dans le cas d'une variable continue X avec deux fonctions de densité f_X et g_X :
On peut également recourir à des tests statistiques pour évaluer la similarité entre les distributions, comme les tests du Chi-deux ou de Kolmogorov-Smirnov. Dans le cas de variables

$$HD(f_X, g_X) = \sqrt{\frac{1}{2}\int\left(\sqrt{f_X(x)} - \sqrt{g_X(x)}\right)^2 dx}.$$

continues, on peut aussi réaliser des tests de Student.

Parmi l'ensemble des variables communes aux deux bases de données et collectées suivant les mêmes définitions, il faut ensuite déterminer celles qui sont les plus corrélées avec les variables d'intérêt de l'appariement, c'est-à-dire les variables ***Y*** et ***Z*** présentes dans les bases *A* et *B* et que l'on cherche à réunir dans une seule et même base. Pour cela, on peut recourir à des méthodes de sélection automatique du type régression pénalisée (par exemple, Ridge, LASSO ou Elastic Net) ou à des algorithmes basés sur des arbres de décision (par exemple, Random Forest). À ce sujet on pourra consulter Schork (2018). Préalablement, il faudra veiller à ne retenir pour l'appariement que des variables avec un faible pourcentage de valeurs manquantes (Van Buuren, 2018).

Le package R *StatMatch*[4] permet de mettre en œuvre de façon concrète l'appariement statistique en utilisant pour l'imputation des données dans la base « receveuse » des méthodes de type hot-deck. La méthode du hot-deck est l'archétype des méthodes d'imputation par donneur. Elle consiste à imputer les valeurs dans la base « receveuse » en prenant les valeurs observées dans la base « donneuse » sur des individus jugés suffisamment « proches ». Cette proximité entre observations se mesure à partir d'une fonction de distance, par exemple la distance euclidienne, définie sur la base des valeurs prises par les variables communes aux deux bases de données. Pour une observation donnée, le « donneur » est en fait choisi de façon aléatoire parmi un ensemble de candidats possibles. Cela permet d'obtenir plus de variabilité dans la distribution des valeurs imputées. Pour une revue des méthodes de hot-deck on pourra consulter (Andridge & Little, 2010). La grande force de ces méthodes est qu'elles conduisent à imputer des valeurs à partir de données réellement observées, ce qui permet de préserver les relations entre les variables et d'éviter de possibles incohérences qui pourraient survenir entre elles lorsqu'on utilise des méthodes d'imputation basées sur des modèles.

4. https://cran.r-project.org/web/packages/StatMatch/StatMatch.pdf

4. Les limites de la mesure de la pauvreté

Chaque année, le STATEC calcule des indicateurs clés sur la pauvreté et les inégalités au Luxembourg à partir de son enquête SILC sur les revenus et les conditions de vie des ménages. Cette enquête est conduite chaque année par le STATEC auprès d'un échantillon de 4 000 ménages et 10 000 individus représentatif de la population résidente au Grand-Duché[5]. Des enquêteurs réalisent les entretiens en face-à-face au domicile des ménages. À partir de 2022, un protocole multimodal associant le face-à-face, le téléphone et l'internet a remplacé le protocole précédemment en vigueur.

Les méthodes actuelles de calcul du taux de pauvreté au Luxembourg mais aussi en Europe se basent principalement sur le revenu disponible des ménages, c'est-à-dire leur revenu calculé après la prise en compte des cotisations sociales et des impôts directs. Cependant, le revenu n'est pas le seul facteur affectant le bien-être d'un ménage. La mesure de la pauvreté basée uniquement sur le revenu présente en fait trois faiblesses principales. Premièrement, le revenu peut fluctuer dans le temps. Ceci est particulièrement vrai pour les travailleurs indépendants ou les chômeurs de courte durée. Meyer & Sullivan (2012) et Brewer & O'Dea (2012) constatent qu'il est préférable d'utiliser la consommation des ménages plutôt que leur revenu si l'on souhaite mieux appréhender le niveau de vie des personnes défavorisées aux États-Unis et au Royaume-Uni. Par ailleurs, la théorie du « revenu permanent » de Friedman suggère que les comportements décisionnels des ménages sont fondés sur les attentes de revenu à long terme plutôt que sur le niveau de revenu actuel. Deuxièmement, le revenu reflète mal la consommation de biens durables comme les voitures ou les logements. Troisièmement, les ménages peuvent compter sur leur épargne et leur patrimoine financier pour lisser leur consommation au cours d'une année difficile et ainsi conserver le même niveau de vie qu'auparavant. Ceux des ménages qui disposent d'un niveau élevé de patrimoine et d'épargne accumulés sont donc moins susceptibles d'être exposés au risque de pauvreté que les ménages disposant de peu de ressources. Ces limites soulignent l'importance d'un cadre multidimensionnel tenant compte à la fois du revenu, de la consommation et du patrimoine afin d'obtenir une meilleure évaluation du bien-être économique des ménages.

C'est pourquoi il est souhaitable de disposer d'un taux de pauvreté multidimensionnelle combinant à la fois le revenu, la consommation et le patrimoine financier des ménages (STATEC, 2022). Cependant, une telle analyse nécessiterait d'avoir une source de données unique fournissant des informations conjointes sur toutes ces dimensions. Actuellement, il n'existe aucune source de données de ce type au Luxembourg. Par conséquent, nous utilisons des techniques d'appariement statistique pour faire correspondre les informations sur la consommation tirée de l'enquête sur le budget des ménages (*Household Budget Survey* – HBS) avec le niveau de patrimoine obtenu dans l'enquête de la Banque Centrale du Luxembourg sur la consommation et le patrimoine financier des ménages (*Household Finance and Consumption Survey* – HFCS) et avec le revenu tel que collecté dans l'enquête SILC sur les revenus et les conditions de vie. Bien que ces trois sources concernent des périodes différentes (2021 pour HBS, 2020 pour les revenus de l'enquête SILC et 2018 pour le HFCS), elles ont néanmoins été rapprochées pour les besoins de cet exercice d'appariement.

Dans cet exercice, SILC a été considérée comme la base « receveuse », HBS et HFCS étant des bases « donneuses ». La méthodologie présentée dans le chapitre précédent a été déroulée afin de déterminer la liste des variables retenues pour l'appariement, avec des listes différentes selon que l'appariement concernait SILC et HBS ou SILC et HFCS :

- **Pour l'appariement SILC/HBS :** déciles du revenu du ménage (disponible et par équivalent-adulte) ; taille du ménage ; nombre d'adultes et d'enfants dans le ménage ;

5. Voir à ce sujet : https://statistiques.public.lu/fr/enquetes/enquetes-particuliers/silc-conditions-vie.html

nombre de pièces du logement ; nombre de membres du ménage en emploi ; nombre d'étudiants dans le ménage ; nombre d'étrangers dans le ménage ; âge, niveau d'éducation, statut marital et pays de naissance du chef de ménage ; région de résidence et statut d'occupation du logement.

- **Pour l'appariement SILC/HFCS :** déciles du revenu du ménage (disponible et par équivalent-adulte) ; taille du ménage ; nombre d'adultes et d'enfants dans le ménage ; nombre de voitures que possède le ménage ; nombre de membres du ménage en emploi ; nombre d'étrangers dans le ménage ; nombre d'individus de 65 ans ou plus dans le ménage ; âge, niveau d'éducation, statut d'activité, statut dans l'emploi et nombre d'heures travaillées par le chef de ménage ; statut d'occupation du logement.

5. Résultats : taux de pauvreté multidimensionnelle

Le taux de pauvreté multidimensionnelle est défini au croisement de la pauvreté de revenu, la pauvreté de consommation et la pauvreté de patrimoine :

- **Pauvreté de revenu** : le revenu de l'individu[6] est inférieur à 60% du revenu médian.
- **Pauvreté de consommation** : la consommation de l'individu[7] est inférieure à 60% de la consommation médiane.
- **Pauvreté de patrimoine** : le patrimoine financier de l'individu[8] est inférieur à 3 mois de ressources monétaires[9]. Ici, nous utilisons uniquement le patrimoine financier pour mesurer la pauvreté car celui-ci peut être facilement liquidé afin de lisser la consommation du ménage.

Les résultats sont présentés dans la Figure 1, où les calculs ont été réalisés au niveau individuel et au niveau ménage. Au niveau individuel, nous avons constaté que le taux de pauvreté selon le revenu était de 18,1% en 2021, tandis que les taux de pauvreté de consommation et de patrimoine étaient respectivement de 19,8% et 32,7%. Calculés au niveau des ménages, ces taux étaient respectivement de 15,6%, 18,5% et 29,9% pour les dimensions du revenu, de la consommation et du patrimoine.

6. Le revenu d'un individu est défini en partant du revenu disponible de son ménage, c'est-à-dire le revenu après impôts directs et cotisations sociales, et en le divisant par le nombre d'équivalents-adulte calculé selon l'échelle d'équivalence dite « OCDE modifiée ». L'ajustement par les échelles d'équivalence permet de comparer le niveau de vie de ménages avec des tailles et des compositions différentes.
7. Même approche que pour la pauvreté de revenu.
8. Là aussi, le patrimoine financier d'un individu correspond au patrimoine financier de son ménage ajusté par le nombre d'équivalents-adulte dans le ménage.
9. En 2021, le seuil de pauvreté monétaire par personne au Luxembourg est de 2 124 euros par mois (source : SILC), ce qui représente 60% du revenu médian des ménages ajustés par le nombre d'unités de consommation. Ainsi, trois mois de ressources correspondent à 6 372 euros par nombre d'unités de consommation. Il est à noter qu'en 2020, le seuil de pauvreté est de 1 892 euros par mois et par unité de consommation, donc le seuil en 2021 augmente de 12%.

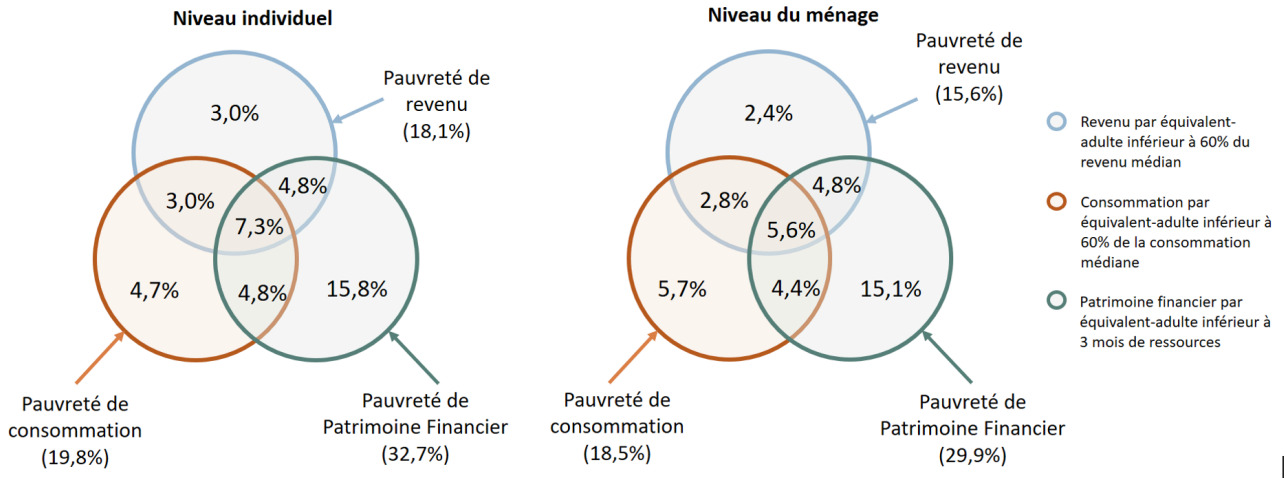

Figure 1 – *Taux de pauvreté multidimensionnelle basée à la fois sur le revenu, la consommation et le patrimoine financier des ménages*

[Source : STATEC, Calcul des auteurs à partir de fichiers synthétiques obtenus par l'appariement de EU-SILC 2021 avec EBM 2021 et HFCS 3ème Vague (2018). Il convient de noter que le revenu collecté dans EU-SILC 2021 provient en fait de l'année 2020.]

Finalement, un résultat important est que la pauvreté au Luxembourg, qui dépasse 18% lorsque sa mesure se base uniquement sur le revenu, chute à 7,3% si l'on tient compte à la fois du revenu, de la consommation et du patrimoine financier.

La proportion de personnes pauvres en patrimoine est nettement plus élevée que pour les deux autres dimensions. Cela reflète le constat selon lequel le patrimoine est beaucoup plus inégalitairement réparti entre les ménages que ne le sont le revenu et la consommation. Le chevauchement entre ces trois dimensions marque le taux de pauvreté multidimensionnelle. Il est de 7,3% au niveau individuel et 5,6% au niveau ménage. Ces deux chiffres représentent une situation de pauvreté « extrême », dans laquelle les ménages touchés ne peuvent s'appuyer ni sur des niveaux de revenus suffisants ni sur des réserves monétaires ou de l'aide financière extérieure pour maintenir un niveau de vie adéquat.

6. Autres applications de l'appariement des bases de données SILC, HBS et HFCS : le taux d'épargne des ménages et le taux de pauvreté énergétique

Sur la base des données appariées, nous arrivons également à produire d'autres indicateurs synthétiques, comme le taux d'épargne des ménages, qui est un indicateur important sur l'accumulation de richesse, et le taux de pauvreté énergétique. Cette approche est conforme aux lignes directrices recommandées dans le rapport Stiglitz, Sen et Fitoussi (2009) et par l'OCDE (2013).

Le taux d'épargne d'un ménage est obtenu en retranchant de son revenu disponible le montant de ses dépenses de consommation finale, et en exprimant cette différence en pourcentage du revenu disponible. Le taux d'épargne moyen par ménage a ainsi été estimé à 28%. Sur la Figure 2, on voit cependant des différences marquées dans les taux moyens d'épargne selon le quintile du revenu : plus le revenu du ménage est élevé, plus le taux d'épargne est grand. On observe même une désépargne, c'est-à-dire une épargne négative, parmi les ménages du premier quintile.

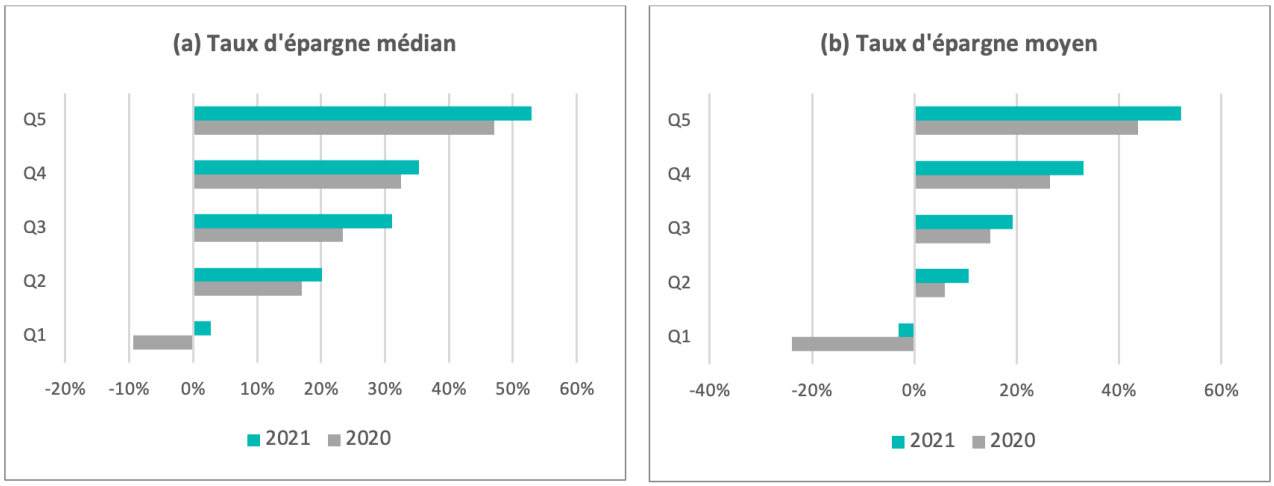

Figure 2 – *Variation du taux d'épargne moyen et médian par quintile entre 2020 et 2021*
[Sources : STATEC, Calcul des auteurs à partir de fichiers synthétiques obtenus à partir de l'appariement d'EU-SILC 2021, d'EBM 2021 et de HFCS 3e vague (2018) pour les taux d'épargne en 2021 ; et d'EU-SILC 2020, d'EBM 2020 et de HFCS 3e vague (2018) pour les taux d'épargne en 2020. Il convient de noter que les revenus collectés dans EU-SILC 2021 se rapportent à l'année 2020 et que ceux collectés dans EU-SILC 2020 se rapportent à l'année 2019.]

Un autre sujet d'importance de cette étude est celui de la pauvreté énergétique. Dans le contexte actuel de crise énergétique, la pauvreté énergétique est devenue un sujet central pour les statistiques sociales. Celle-ci est cependant définie différemment entre les pays en développement et les pays développés. Dans le premier cas, elle est généralement comprise comme un manque d'accès aux services énergétiques, alors que dans le second cas elle est attribuée au poids excessif des dépenses énergétiques par rapport aux revenus des ménages. C'est cette seconde approche que nous avons retenue dans le cas du Luxembourg (Di Falco *et al.*, 2021). Pour mesurer quantitativement la pauvreté énergétique, nous avons utilisé les deux indicateurs suivants :

i. Taux d'effort énergétique (TEE) élevé :

$$\text{TEE} = \frac{\text{Dépenses énergétiques du ménage}}{\text{Revenu du ménage}} > 2 \times \text{Valeur médiane nationale en 2012 ;}$$

ii. Bas Revenus, Dépenses Élevées (BRDE) :

$$\text{BRDE} = \begin{cases} \dfrac{\text{Dépenses énergétiques du ménage}}{\text{par unité de consommation}} > \text{Valeur médiane nationale en 2012} \\ (\text{Revenu net du ménage} - \text{Charges du logement}) < 60\% \times \\ \text{Médiane du (Revenu net du ménage} - \text{Charges du logement) en 2012.} \end{cases}$$

Les mesures TEE sont fréquentes dans la littérature (Di Falco *et al.*, 2021) car elles sont faciles à calculer et à expliquer. Cependant, elles ne tiennent pas compte des niveaux de revenu et

pourraient alors inclure des ménages ayant un niveau de revenu élevé mais faisant un possible gaspillage d'énergie, ce qui entrainerait des dépenses énergétiques importantes. L'indicateur BRDE, quant à lui, est plus compliqué car il inclut deux conditions simultanées : pour être considéré en précarité énergétique, un ménage doit cumuler à la fois un faible revenu et des dépenses énergétiques élevées. Les ménages qui se situent en dessous du seuil conventionnel de 60% du revenu médian et dont les dépenses d'énergie sont supérieures au niveau médian sont considérés comme étant en situation de pauvreté énergétique. Suivant Di Falco *et al.* (2021), nous avons choisi l'année 2012 comme point de référence pour notre mesure de la pauvreté énergétique. Pour l'indicateur TEE, la valeur médiane nationale en 2012 était de 7,2%, donc tous les ménages dont la part des dépenses énergétiques par rapport au revenu est supérieure au double de ce seuil (soit 14,4%) sont considérés en situation de pauvreté énergétique. Concernant l'indicateur BRDE, la dépense énergétique médiane ajustée par le nombre d'unités de consommation dans le ménage s'élevait à 1 194 euros par an en 2012. Quant au seuil défini par 60% du revenu net médian après la prise en compte des charges de logement, sa valeur était de 17 263 euros par an et par ménage en 2012, soit 1 438 euros par mois.

La Figure 3 montre un TEE de 4,9% et un BRDE de 3,9% pour l'année 2021, tandis que 1,8% des ménages au Luxembourg sont en situation de précarité énergétique d'après les deux définitions. Ces chiffres appellent cependant une mise à jour pour tenir compte des poussées inflationnistes récentes que l'on enregistre sur les prix de l'énergie. Pour plus d'éléments sur cette question, on pourra consulter STATEC (2022).

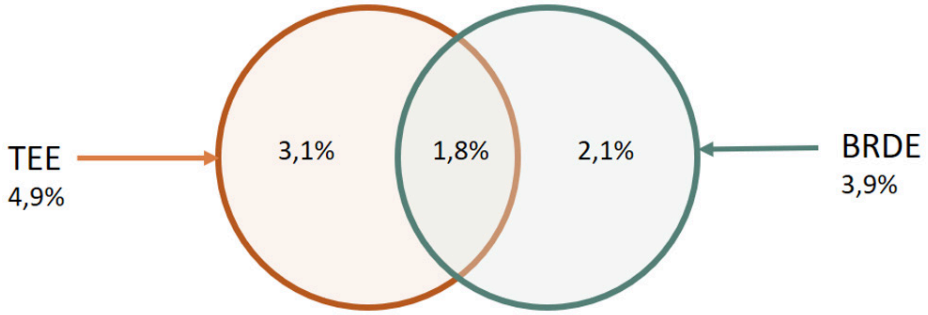

Figure 3 – *Indicateurs de précarité énergétique au Luxembourg (TEE et BRDE)*
[Sources : STATEC, Calcul des auteurs à partir de fichiers synthétiques obtenus par l'appariement de EU-SILC 2021 avec EBM 2021. Il convient de noter que le revenu collecté dans EU-SILC 2021 provient de l'année 2020.]

Si les indicateurs TEE et BRDE sont aujourd'hui couramment utilisés, ils doivent être complétés par des indicateurs de nature plus subjective sur le ressenti des ménages (Charlier *et al.*, 2015). C'est ce qui est illustré dans le Tableau 2 où l'on voit que les ménages en situation de précarité énergétique sont les plus modestes en termes de revenu et de mesures subjectives. Ces mesures subjectives proviennent des réponses données par les ménages aux questions de savoir si : (i) le ménage ne peut pas se chauffer suffisamment en hiver, (ii) le ménage rencontre des difficultés pour les paiements des factures énergétiques. On voit qu'il y a une corrélation entre le fait pour un ménage d'être en situation de précarité énergétique selon les indicateurs TEE et BRDE, et les difficultés à payer son loyer, rembourser ses emprunts ou encore payer ses factures courantes en lien avec son logement. De ce point de vue, on peut conclure que les mesures subjectives « valident » les indicateurs objectifs de pauvreté énergétique (TEE et BRDE).

Tableau 2 – *Indicateurs objectifs et subjectifs de précarité énergétique*

	TEE		BRDE		TEE & BRDE	
	Oui	*Non*	*Oui*	*Non*	*Oui*	*Non*
Statistiques de base						
- Revenu net moyen par unité de consommation (EUR/an)	20 934	51 566	20 015	51 435	15 545	50 884
- Dépenses énergétiques moyennes (EUR/an)	2 957	1 623	2 258	1 657	2 668	1 662
Mesures subjectives						
- Ne pas pouvoir chauffer suffisamment son logement	0,8%	2,4%	4,6%	2,3%	2,2%	2,3%
- Arriérés sur le paiement des factures énergétiques	7,6%	2,6%	9,6%	2,6%	15,9%	2,6%
- Arriérés sur le paiement des prêts hypothécaires ou des loyers	4,1%	2,0%	6,2%	1,9%	9,9%	1,9%
- Arriérés sur le paiement des autres emprunts	13,8%	4,2%	13,6%	4,4%	22,6%	4,4%

[Sources : STATEC, Calcul des auteurs à partir de fichiers synthétiques obtenus par l'appariement de EU-SILC 2021 avec EBM 2021. Il convient de noter que les revenus collectés dans EU-SILC 2021 proviennent de l'année 2020.]

7. Conclusion

L'appariement statistique est une technique puissante et rentable qui permet d'exploiter tout le potentiel des bases de données disponibles afin de construire des indicateurs beaucoup plus riches et bien plus intéressants que ceux établis à partir d'une unique source de données. Le taux de pauvreté multidimensionnelle en est un parfait exemple, dans la mesure où il fournit une image de la pauvreté qui est beaucoup plus fine que celle définie seulement à partir du niveau de revenu des ménages. Les deux autres exemples présentés dans cet article sur l'épargne et la précarité énergétique des ménages sont tout aussi pertinents à l'aune des crises que traversent actuellement les pays européens dans le domaine économique et dans celui de l'énergie.

Toutefois, nous conseillons de réaliser une analyse de sensibilité pour toutes les variables appariées. De façon générale, il convient d'être prudent lors de l'utilisation de mesures synthétiques basées sur des données appariées, car celles-ci reposent sur des estimations qui sont de nature expérimentale et doivent donc faire l'objet de tests et de validations supplémentaires. Des analyses plus poussées sur la qualité et la robustesse de l'appariement sont en cours. Celles-ci consistent notamment à comparer les distributions des principales variables socio-économiques observées sur les bases de données originelles et sur la base de données synthétiques issue de l'appariement. Pour plus de détails on pourra consulter Eurostat (2013).

L'hypothèse d'indépendance conditionnelle constitue une autre limite. Dans cet exercice d'appariement des bases de données SILC, HBS et HFCS, nous avons supposé que l'hypothèse d'indépendance conditionnelle était vérifiée, ce qui est probablement une simplification de la réalité. On peut toutefois considérer une telle simplification comme raisonnable dans la mesure

où les variables utilisées pour l'appariement incluent le revenu du ménage, qui est disponible dans les trois bases de données sous revue et qui est lié de manière directe aux indicateurs qui nous intéressent. De façon générale, on trouvera dans Rässler (2004) et Lamarche (2017) des éléments de réponse sur la validité ou non de l'hypothèse d'indépendance conditionnelle et de quelle manière la procédure d'appariement statistique doit être adaptée lorsque cette hypothèse ne peut pas être vérifiée.

Il faut enfin mentionner que les années de référence des trois bases de données avec lesquelles nous avons travaillé n'étaient pas les mêmes : 2021 pour HBS, 2020 pour SILC et 2018 pour HFCS. Cette situation est malheureuse et contraire aux lignes de bonnes pratiques, mais il s'agissait là d'une contrainte liée au fait que SILC collecte les revenus détaillés pour l'année précédant l'enquête et que l'enquête HFCS n'est conduite que tous les 3 ou 4 ans.

Références

Andridge R. R. and R. J. A. Little (2010), « A Review of Hot Deck Imputation for Survey Non-Response », *International Statistical Review*, 78(1), pp. 40-64.

Brewer M. and C. O'Dea (2012), « Measuring living stardards with income and consumption: evidence from the UK », *ISER Working Paper Series*, N° 2012-05, Institute for Social and Economic Research (ISER), Essex.

Charlier D., A. Risch et C. Salmon (2015), « Les indicateurs de la précarité énergétique en France », *Revue d'Économie Française*, 4, pp. 187-230.

D'Orazio M., M. Di Zio and M. Scanu (2006), *Statistical Matching: Theory and Practice*, Chichester, Wiley.

D'Orazio M. (2010), « Evaluation of the accuracy of statistical matching, Report WP1 ESS-net », Statistical Methodology Project on Integration of Surveys and Administrative Data.

Di Falco E., O. Thunus et G. Zardet (2021), « Analyse sur la précarité énergétique au Luxembourg », Document de travail du STATEC, https://statistiques.public.lu

Eurostat (2013), « Statistical matching: a model based approach for data integration », *Methodologies and Working Papers*, https://ec.europa.eu/eurostat

Kadane J. B. (1978), « Some statistical problems in merging data files », Department of Treasury, Compendium of Tax Research, Washington, DC: US Government Printing Office, pp. 159-179.

Lamarche P. (2017), « Estimating consumption in the HFCS - Experimental results on the first wave of the HFCS », European Central Bank, Statistics Paper Series, 22, https://www.ecb.europa.eu/

Meyer B. D. and J. Sullivan (2012), « Identifying the Disadvantaged: Official Poverty, Consumption Poverty, and the New Supplemental Poverty Measure », *Journal of Economic Perspective*, 26(3), pp. 111-136.

OCDE (2013), « OECD Framework for Statistics on the distribution of income, consumption and wealth », OECD Publishing, https://www.oecd.org/statistics/framework-for-statistics-on-the-distribution-of-household-income-consumption-and-wealth-9789264194830-en.htm

Paass G. (1986), « Statistical match: evaluation of existing procedures and improvements by using additional information », *in* G. H Orcutt, J. Merz, and H Quinke (eds.), *Microanalitic Simulation Models to Support Social and Financial Policy*, Elsevier, pp. 401-422.

Rässler S. (2002), *Statistical Matching, a Frequentist Theory, Practical Applications and Alternative Bayesian Approach*, Springer.

Rässler S. (2004), « Data fusion: identification problems, validity, and multiple imputation », *Austrian Journal of Statistics*, 33(1-2), pp. 153-171.

Ruggles N. and R. Ruggles (1974), « A strategy for merging and matching microdata sets », *Annals of Economic and Social Measurement*, 1(3), pp. 353-371.

Schork J. (2018), « Automatic Variable Selection for Imputation Models: Common Methods Applied to EU-SILC », *STATEC Working Paper*, N° 98/2018, https://statistiques.public.lu

STATEC (2022), « D'une crise à l'autre : la cohésion sociale sous pression », Rapport travail et cohésion sociale, Analyses, https://statistiques.public.lu/

Stiglitz J., A. Sen et J.-P. Fitoussi (2009), *Rapport de la Commission sur la mesure des performances économiques et du progrès social*, Éditions Odile Jacob.

Van Buuren S. (2018), *Flexible imputation of missing data*, Chapman & Hall/CRC, https://stefvanbuuren.name/publication/vanbuuren-2018/