CONSIDÉRATIONS THÉORIQUES ET PRATIQUES CONCERNANT LES TECHNIQUES D'APPARIEMENT

THEORETICAL AND PRACTICAL CONSIDERATIONS CONCERNING STATISTICAL MATCHING

Roxana Adam[1]

[1] *16, Institut National de la Statistique, Blvd. Libertatii., 5e arrond., Bucarest, Roumanie, roxana.adam@insse.ro*

Résumé. L'étude présente un aperçu sur la perspective théorique concernant l'appariement statistique, combiné simultanément avec des considérations pratiques. Le sujet est traité au niveau "micro" et traite aussi l'ensemble des problèmes concernant l'hypothèse implicite sur l'indépendance conditionnelle dans la pratique. Les niveaux de validation de Rässler concernant l'appariement statistique sont discutés et proposés pour l'analyse de la qualité de la méthode d'appariement. Finalement, une situation typique d'appariement est décrite en quatre étapes, avec des exemples concrets liés aux enquêtes EMB et EU-SILC.

Mots-clés. l'appariement statistique, les distributions communes, l'indépendance conditionnelle, EMB, EU-SILC

Abstract. The paper presents an overview from a theoretical perspective on statistical matching, simultaneously combined with practical considerations. The subject is approached at the "micro" level and addresses the issue of the conditional independence assumption in practice. Rässler's levels of validity for statistical matching are discussed and proposed for the quality analysis of the matching method. Finally, a typical matching situation is described in four steps with concrete examples related to HBS and EU-SILC survey.

Keywords. statistical matching, joint distributions, conditional independence assumption, HBS, EU-SILC

Introduction

La globalisation, le développement de la société civile, le progrès technologique notamment, constituent des quelques raisons qui déterminent les chercheurs à trouver plus de données statistiques à utiliser dans des analyses complexes pour comparer deux ou plusieurs unités d'observation. Que l'on se réfère à la situation de la population d'un pays ou à des compagnies ou des groupes défavorisés socialement, il est nécessaire d'avoir des données pertinentes, exactes et compatibles afin de déterminer un modèle de comportement. Cet article traite la nécessité de combiner les fichiers statistiques qui ne contiennent pas les mêmes unités ou qui peuvent contenir des unités communes, mais sans un code unique d'identification ou d'autres caractéristiques capables de les identifier.

En l'absence des codes d'identification uniques des unités statistiques ou des variables auxiliaires comme, par exemple, le nom et la date de la naissance ou l'adresse de contact, la connexion des pièces d'informations reste impossible à réaliser par la Connexion des Données (Data Linkage). Dans ce cas, les techniques de la méthode d'appariement, aussi connue sous le nom de Statistical Matching (SM) ou Data Fusion, peuvent constituer la meilleure alternative pour combiner les données. Les données statistiques sont très importantes pour élaborer des politiques, pour planifier et évaluer les services, autrement dit les données sont vitales pour prendre des décisions tant au niveau micro que macro. Il est évident que la personne responsable de la procédure d'appariement a l'obligation de ne pas altérer la qualité des données. Par conséquent, se pose inévitablement la question de la façon dont on peut réaliser la procédure d'appariement.

1. Fondements de l'appariement statistique

Les chercheurs (Rubin, 1987; Gelman et al., 1995; Tanner, 1996; Carlin and Louis, 2000) ont étudié les techniques pour l'imputation des données manquantes se référant à plusieurs méthodes, à savoir: l'imputation multiple, Markov chain Monte Carlo (MCMC), l'inférence Bayesienne, et lorsque les variables d'intérêt ne sont pas observées conjointement, on a proposé de résoudre le problème de l'estimation des paramètres par l'appariement statistique (Rässler, 2002; D'Orazio et al., 2006).

En ce qui concerne le procédé d'appariement statistique (dénommé par la suite SM), on va dénommer les deux sources de données A et B. Elles partagent un ensemble de variables $X_{1...n}$, pendant que la variable Y est observée seulement en A et la variable Z seulement en B. Les variables $X_{1...n}$ sont communes pour les deux sources de données, pendant que les variables Y et Z sont uniques (à voir Figure 1). L'objectif de SM est de rechercher la relation entre Y et Z au niveau "micro" ou "macro" (D'Orazio et al., 2006). Dans cet article, on va se référer seulement au niveau "micro".

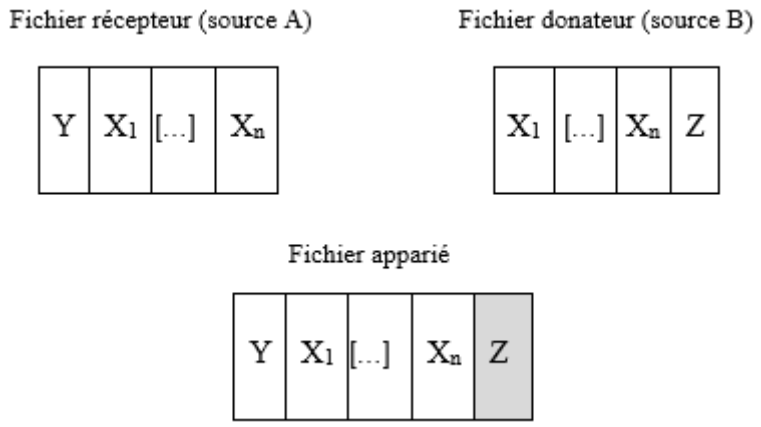

Figure 1: Appariement statistique

On comprend par SM, le transfert de données statistiques entre plusieurs fichiers basés sur des variables communes. Les méthodes les plus utilisées sont *means* of *nearest neighbor* ou *hot deck* à l'aide desquelles, en fonction des variables communes entre deux fichiers, les variables spécifiques qu'on observe seulement dans le fichier donateur sont attribuées aux unités récepteurs du deuxième fichier, créant ainsi la base de données d'intérêt. Sans tenir compte de la méthode utilisée, on recommande la technique SM lorsque le nombre d'unités statistiques identiques dans deux échantillons est zéro ou très petit. Dans ce cas, l'imputation des valeurs d'une variable est possible

grâce à la "similitude" des unités statistiques des deux bases de données, étant et il est possible que survienne l'hypothèse implicite sur l'indépendance conditionnelle (CIA) (Rässler, 2002).

Le problème principal de l'appariement statistique est l'association conditionnelle elle même, à savoir l'association des variables pas observées conjointement, prenant en compte des variables communes des deux fichiers utilisés, qui ne peut pas être évaluée à partir des données observées (Rubin, 1974; Rässler, 2002). Prenant comme point de départ Rubin (1987), l'imputation multiple représente une solution possible qui peut mener au relâchement de CIA ou tout autre choix spécifique pour les paramètres d'association conditionnelle. Rässler (2002) présente l'imputation multiple en tant que l'instrument nécessaire pour atteindre les limites supérieures et inférieures de l'association non-conditionnelle des variables jamais observée conjointement. L'auteur affirme que cela est possible en utilisant des variables communes adéquates qui déterminent, autant que possible, l'association non-conditionnelle. Dans son document, Rässler montre que le principe de l'imputation multiple nous permet d'établir l'association conditionnelle et, par conséquent, il ne serait plus nécessaire de se rapporter à CIA dans le procédé d'appariement statistique.

2. Niveaux de validation selon Rässler sur l'appariement statistique

Lorsqu'on utilise SM, il faut établir le cadre théorique sur le manque des données – un vrai problème – ainsi que la validation du modèle. Sans se concentrer sur un algorithme d'appariement spécifique, Rässler (2002) distingue quatre niveaux de validation qu'une procédure d'appariement peut avoir.

a. Conservation des valeurs individuelles

C'est le niveau de validation le plus difficile de l'appariement statistique se référant à la capacité de la procédure SM de reproduire les valeurs individuelles réelles, mais inconnues, des données de l'échantillon.

La reproduction exacte des valeurs est possible seulement dans le cas où les variables communes X1...n déterminent déjà exactement la variable Y.

Ce niveau est évidemment très difficile à réaliser et, dans la plupart des cas, il n'est pas pratique.

b. Conservation des distributions communes

Cela suppose que les unités des deux sources de données sont indépendantes tant à l'intérieur qu'entre les deux échantillons et que le fichier synthétique (qui résulte de l'appariement) est un échantillon choisi aléatoirement d'une distribution synthétique. Ce niveau est possible seulement si les variables uniques de la base de données A et les variables uniques de la base de données B sont indépendantes conditionnellement par rapport aux variables communes des deux enquêtes.

Il peut être considéré comme le niveau le plus intéressant car il se réfère au fichier synthétique comme un échantillon réel à une seule source. Par conséquent, une procédure d'appariement correcte garde les distributions des variables communes et permet toute inférence statistique individuelle sur la base du fichier synthétique.

c. Conservation des structures de la co-variation/corrélation

Comme dans le cas du niveau de validation précédent, on doit considérer le fichier synthétique comme généré aléatoirement d'une population artificielle, qui a au moins les mêmes moments et la

même structure de corrélation en tant que population actuelle d'intérêt. Autrement dit, la structure de la co-variation des données après et avant l'appariement devrait être la même. On souligne que les variables qui sont indépendantes conditionnellement ne se trouvent pas, également, en corrélation conditionnellement, pourtant la situation vice versa n'est généralement pas valable.

Pour que le troisième niveau soit suffisant, on doit assurer CIA pour les variables qui ne sont pas communes dans les deux bases de données.

d. Conservation des distributions marginales

Ce niveau de validation représente ce qu'on souhaite être une exigence minimale de l'appariement statistique et traite la conservation des distributions qui peuvent être observées, dès le début, séparément dans les deux bases de données. Autrement dit, l'exigence minimale pour l'appariement statistique est que les distributions marginales des variables individuelles de l'enquête initiale (originelle) soient aussi conservées après la procédure d'appariement.

Le quatrième niveau de validation est difficile à atteindre lorsqu'on rencontre un plan de sondage complexe. Ce qu'on peut tester est si les distributions empiriques du fichier donateur sont similaires à celles du fichier synthétique. Pourtant, ces tests peuvent être réalisés par des expérimentations, simulations ou en utilisant une troisième base de données complète. On apprend dans ce cas que, dans une situation typique d'appariement, on peut contrôler seulement le quatrième niveau.

3. Une situation typique d'appariement

Tout ce qu'on a remarqué jusqu'à présent est le fait que la procédure SM est très importante lorsqu'il y a une base de données principale, ayant des informations d'intérêt, mais il nous manque une variable qui se retrouve dans une autre base de données, les deux étant des échantillons de la même population. Son importance réside, premièrement, dans des aspects liés aux coûts, en corrélation avec la nécessité d'utiliser les données déjà existantes pour réaliser les analyses nécessaires au niveau décisionnel. Autrement dit, pour mieux connaitre les données déjà existantes et les utiliser dans plusieurs buts.

Dans cette section, on poursuit la description au niveau général de la procédure d'appariement pour deux enquêtes. Bien que les exemples appartiennent à la statistique officielle, on considère que sa structure peut être aussi appliquée dans le domaine des affaires ou académique lorsque les données proviennent de deux sources de la même population statistique.

En partant des aspects théoriques, les exemples pratiques se réfèrent aux enquêtes sociales européennes*: Enquête sur le budget des ménages* (EMB) et *Statistiques de l'Union Européenne sur le revenu et les conditions de vie* (EU-SILC). L'intérêt à l'égard de SM, dans le cas des deux enquêtes, peut être aussi rencontré dans d'autres travaux récents (à voir Donatielle et al.,2014 et Lamarche, 2017). L'élément de nouveauté de cet article réside dans la structuration succincte du procédé d'appariement à pas principaux, dans un cadre général, de sorte qu'il soit plus facile à comprendre par les pionniers du domaine.

Premier pas: Harmonisation et réconciliation des sources multiples.

Conformément à D'Orazio et al (2006), il faut prendre en considération:

(a) l'harmonisation de la définition des unités

(b) l'harmonisation de la période de référence

(c) le complètement de la population

(d) l'harmonisation des variables

(e) l'harmonisation des classifications

(f) l'ajustement des erreurs de mesurage (précision)

(g) l'ajustement des données manquantes

(h) la dérivation des variables

Revenons par exemple aux enquêtes EMB et EU-SILC et précisons que les définitions des unités statistiques (par exemple: le ménage) doivent coïncider dans les deux sources. Il faut aussi surveiller pour les autres variables d'intérêt que les définitions soient similaires (par exemple: le statut d'occupation professionnelle, le niveau d'éducation, le statut civil, le revenu etc). Lorsque les variables sont définies similairement, mais sont classifiées différemment, il est nécessaire que les données soient reclassifiées dans une variante commune pour les deux sources de données (par exemple: la classification des occupations professionnelles en Roumanie formées d'un code de 4 digit dans une enquête et d'un code de 2 digit dans une autre, sera reclassifiée par le regroupement des occupations à 4 digit en groupes de 2 digit).

Deuxième pas: Analyse du pouvoir explicatif pour les variables communes

Cette étape a une importance particulière car il est crucial que SM se réalise sur la base des variables communes, qui prennent la forme de bons prédicteurs des informations spécifiques qui doivent être transférées du donateur au destinataire. Leulescu et Agafitei (2013) soutiennent que, pour la plupart des cas, le point de référence est CIA. Quant à l'appariement statistique, D'Orazio et al (2006) constate que cette hypothèse est si forte qu'elle est, malheureusement, rarement vérifiée en pratique. L'absence de l'indépendance conditionnelle peut mener à des interférences incorrectes lorsqu'on analyse les données obtenues par SM.

Quant à l'appariement, il est préférable d'utiliser les variables à un haut niveau de qualité sans erreurs et données manquantes (Cibella, 2010), évitant d'utiliser des variables traitées comme variables communes pour l'appariement (Scanum 2010).

Dans l'exemple d'appariement statistique entre EMB et EU-SILC discuté par Donatiello et al. (2014), on présente l'impossibilité de réaliser l'appariement statistique sous l'hypothèse de l'indépendance conditionnelle (CIA), et afin d'éviter CIA, on utilise des informations auxiliaires disponibles (par exemple: le revenu mensuel des ménages). Par conséquent, les auteurs prennent en considération comme alternative, l'appariement statistique basé sur l'exploration de l'incertitude due à l'absence des informations communes sur les variables Y et Z. Il est à mentionner que la distance Hellinger (HD) a été utilisée, avant tout, comme une mesure de la cohérence des variables communes, et aussi pour analyser la ressemblance/dissimilarité des distributions des variables des deux ensembles de données.

Troisième pas: Méthodes d'appariement

On a utilisé plusieurs méthodes d'appariement. Rässler (2002) présente la manière d'aborder traditionnellement la méthode d'appariement par rapport à: l'appariement non-contraint, l'appariement contraint, l'appariement contraint catégoriquement, le concept topologique et l'attribution multiple aux variables spécifiques. Les différentes méthodes alternativement présentées par l'auteur sont: l'imputation multiple, l'imputation de la régression avec des résidus aléatoires, la procédure d'imputation multivariée non-itérative, l'augmentation des données, imputations univariées itératives par équations chaînées et des études de simulation basées sur des données

normales multivariées.

D'Orazio et al. (2006) classifie les méthodes d'appariement dans le cas des populations limitées comme il suit: des méthodes paramétriques conformément à CIA, des méthodes paramétriques ou les informations auxiliaires sont disponibles et des méthodes non-paramétriques. Ultérieurement, D'Orazio (2011) présente la méthode mixte, possible seulement lorsque SM a lieu au niveau micro et consiste dans la réalisation d'une association entre les méthodes paramétriques et non-paramétriques.

Beaucoup des techniques proposées pour SM au niveau micro sont basées sur des méthodes développées pour l'imputation des valeurs manquantes: paramétriques (par exemple: l'imputation régressive); des méthodes non-paramétriques (par exemple: l'imputation hot deck) ou mixtes (par exemple: des méthodes basées sur l'appariement prévisible) (Donatiello et al., 2014).

D'Orazio (2013) a réalisé SM entre EM et EU-SILC par l'entremise du logiciel R utilisant le paquet StatMatch. L'auteur a utilisé la méthode random hot deck sous les conditions de CIA et, ultérieurement, il a utilisé la variable du revenu du ménage comme information auxiliaire pour relâcher CIA.

De la même manière, Lamarche (2017) a adopté SM entre EMB et EU-SILC par les méthodes: *random hot deck* et l'approche mixte. On considère que, dans ce cas, le modèle *random hot deck* est à préférer car la sélection systématique des variables permet la réalisation de l'appariement pour tous les pays de l'EU qui collecte les deux enquêtes. Le point faible de cette méthode réside dans le fait que les variables d'appariement peuvent être différentes au niveau des pays, en fonction de la spécificité de chaque pays et pas seulement, et dans ce cas, on recommande d'étudier l'harmonisation du processus de sélection des variables.

Quatrième pas: Evaluation de la qualité des résultats

La qualité et la cohérence des sources de données, le pouvoir explicatif des variables communes, les méthodes d'appariement adéquates aux sources de données constituent des exigences préliminaires pour évaluer le processus d'appariement. Une fois que ceux-ci sont accomplis, on peut évaluer les autres critères de qualité. Dans ce sens, il est préférable qu'au moins un des niveaux de Rässler sur le SM soit validé. Même si la reproduction des valeurs au niveau individuel représente le niveau le plus désirable, il est le plus difficile à atteindre en pratique. Dans un but pratique, il devrait exister un jeu de données qui contienne, en même temps, les variables $X_{1\ldots n}$, Y et Z, pouvant être ultérieurement divisé en deux fichiers séparés, un jeu de données fondamentales et un set de données pour les tests. Du jeu de données pour les tests, on doit éliminer la variable Z et on doit ensuite réaliser la procédure d'appariement pour comparer les nouvelles valeurs traitées à celles originales.

Au deuxième niveau, la reproduction de la distribution des variables communes est difficile à atteindre, mais utilisant la distance Hellinger peut constituer une solution acceptable dans le cas où il n'y a pas d'informations auxiliaires qui peuvent être directement incorporées dans la procédure d'appariement. Au troisième niveau, la reproduction de la structure de corrélation entre les variables d'intérêt suppose la reproduction de la distribution des variables communes seulement dans de certaines conditions spécifiques. Dans ce sens, le coefficient de corrélation de Pearson devrait être suffisant lorsqu'il y a une information auxiliaire disponible.

Quant au dernier niveau, la reproduction de la distribution marginale pour chaque variable d'intérêt lorsque le plan de sondage n'est pas très complexe, une solution possible serait de comparer la probabilité de la distribution des deux sources par la représentation graphique des quantiles, à l'aide d'un crampon Q-Q. On peut aussi utiliser le *test* x^2 (Chi-squared test), le *test t* ou d'autres tests non-paramétriques ainsi que des régressions multiples. Une concordance réussie devrait mener à des relations similaires entre les variables communes et celles spécifiques entre le donateur et le fichier

conforme à l'appariement statistique. On tire la conclusion que, pour la plupart des cas, la procédure d'appariement statistique a du succès si les distributions empiriques marginales et communes de $X_{1...n}$, et Y, ainsi qu'observées dans le fichier du donateur, sont presque les mêmes dans le fichier obtenu suite au SM (Baker et al., 1989; Rässler, 2002).

Simulation SM entre EMB 2014 et EU-SILC 2014

Pour illustration, on a souhaité de réaliser une brève simulation sur les enquêtes EMB 2014 et EU-SILC 2014 en Roumanie. Ainsi que signalé par Donatiello et al. (2014) et Lamarche (2017), le revenu des deux enquêtes est collecté dans des périodes différentes. Si pour EMB 2014, on peut distinguer la variable revenu pour l'année 2014, dans le cadre EU-SILC 2014, la variable revenu se réfère aux sommes obtenues en 2013. Pour réaliser l'exercice d'appariement dans cette situation, on a analysé le revenu d'EMB 2013 et EMB 2014. Les différences par groupes de revenu on été très faibles, et c'est la raison pour laquelle on a choisi de se rapporter à une procédure SM entre EMB 2014 et EU-SILC 2014.

Ainsi que mentionné ci-dessus, il est improbable de réaliser l'appariement statistique sous l'hypothèse CIA (Donatiello et al., 2014), par conséquent, pour l'étape ultérieure, on a sélectionné les variables communes déjà existantes dans les deux enquêtes, on les a harmonisées et on a évalué le pouvoir explicatif de celles-ci. A l'aide de la distance de Hellinger (HD), on a identifié les variables les plus adéquates (à voir Figure 2).

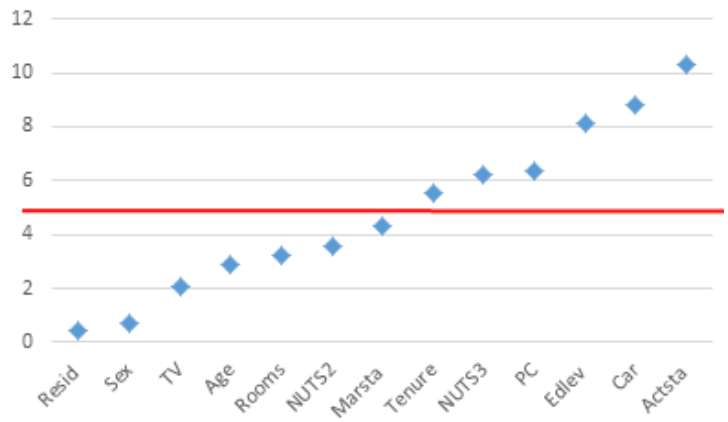

Figure 2: La distance de Hellinger (%) par les variables communes principales de EMB et EU-SILC, 2014

Les variables qui sont sous le seuil de 5% sont Resid, Sex, TV, Age, NUTS2 et Marsta (à voir la définition des variables dans l'Annexe 1). Dans le but de l'exercice, l'intention de la procédure SM a été d'inclure dans le nouveau fichier synthétique, au niveau individuel, la variable binaire pauvre – non-pauvre (les personnes situées sous le seuil de la pauvreté) de EU-SILC à l'aide de R et du paquet *StatMatch*, par la fonction NND.hotdeck. Le premier essai a été celui d'inclure toutes les variables issues de l'analyse de la distance de Hellinger ayant une valeur plus petite de 5%. Cela n'a pas été possible car la fonction NND.*hotdeck* est limitée par le fait que toutes les combinaisons des variables de la base de données donatrice doivent se retrouver dans la base de données bénéficiaire. Par conséquent, on a éliminé l'une après l'autre les variables ayant la plus grande distance de Hellinger, jusqu'au point où les combinaisons des variables se sont retrouvées dans les deux bases de données. On a considéré pourtant que cette procédure ne respecte pas le plan des échantillons, et, pour le SM final, on a choisi comme variables d'appariement: *NUTS2* et *Age*.

Il va de soi que cet exemple est discutable, mais il est pertinent dans le contexte théorique car il attire l'attention sur les problèmes qui peuvent apparaître en pratique. Si les modèles proposés pour l'appariement statistique sont conditionnés par des hypothèses difficiles à respecter en pratique, la question que tout chercheur doit se poser est la suivante: *quelle serait la meilleure méthode d'appariement de sorte que les hypothèses du modèle ne soient pas violées et que les résultats reflétent la réalite le plus exactement possible*.

Conclusions

Les techniques d'appariement sont utiles lorsqu'il est nécessaire de combiner deux bases de données, sans avoir un élément d'identification commun, de la même population. La littérature de spécialité traite ce sujet le plus souvent dans perspective théorique et méthodologique, et lorsqu'il s'agit d'un exercice effectif, on observe qu'en pratique, basées sur des données réelles, les techniques d'appariement présentent certaines limitations. Dans cette étude, on peut observer que, bien que du point de vue de la méthodologie, il est important de respecter l'hypothèse implicite sur l'indépendance conditionnelle, la pratique nous montre que CIA est très difficile à respecter et la meilleure alternative serait son relâchement.

Un autre problème important est l'évaluation de la base de données synthétique obtenue suite à la procédure d'appariement. Bien que Rässler ait proposé quatre niveaux de validation, dans la plus optimiste situation, les chercheurs réussissent à conserver la distribution marginale des variables individuelles de l'enquête initiale (originale) dans le fichier synthétique. En dépit de ces obstacles qui prennent la forme de la violation des hypothèses des modèles en pratique, il est nécessaire de développer une procédure qui permette aux chercheurs de combiner deux bases de données de la même population lorsqu'on a besoin d'une information qui existe, mais qui n'existe pas dans la base de données principale pour l'analyse. La recherche et le développement des techniques d'appariement doivent toujours tenir compte de la conservation de la qualité des données résultant de la procédure.

Bibliographie

Baker, K., Harris, P., O'Brien, J. (1989). Data Fusion: An Appraisal and Experimental Evaluation, *Journal of the Market Research Society*, 31, 153-212.

Carlin, B.P., Louis, T.A. (2000), Bayes and Empirical Bayes Methods for Data Analysis, *Chapman and Hall*, London.

Cibella, N (2010). How to choose the matching variables, Report WP2, ESS-net, *Statistical Methodology Project on Integration of Surveys and Administrative Data*, EUROSTAT.

Donatiello, G., D'Orazio, M., Frattarola,D., Rizzi, A., Scanu, M. And Spaziani, M. (2014). Statistical Matching of Income and Consumption Expenditures, *International Journal of Economic Sciences*, Vol. III, No. 3.

D'Orazio, M. (2013). Statistical matching: Metodological issues and practice with R-StatMatch, *Eustat*.

D'Orazio, M. (2011). Statistical Matching and Imputation of Survey Data with the Package "StatMatch" for the R Environment. *Conference Of European Statisticians,* Ljubljana, Slovenia, 9-11 May 2011.

D'Orazio, M. (2011). Statmatch: Statistical matching [Computer software manual]. Available from http://CRAN.R-project.org/package=StatMatch.

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

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (1995), *Bayesian Data Analysis*, Chapman and Hall, London.

Lamarche, P. (2017). Measuring Income, Consumption and Wealth jointly at the micro-level, *Draft methodological note – June 20 2017*, Eurostat, Luxembourg, available from: http://ec.europa.eu/eurostat/documents/7894008/8074103/income_methodological_note.pdf

Leulescu și Agafiței (2013). *Statistical matching: a model based approach for data integration*, EUROSTAT.

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

Rubin, D.B. (1987), *Multiple Imputation for Nonresponse in Surveys*, Wiley, New York.

Scanu, M. (2010), Recommendations on statistical matching, Report WP2, ESS-net, *Statistical Methodology Project on Integration of Surveys and Administrative Data*, EUROSTAT.

Tanner, M.A. (1996). *Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions*. Springer, New York.

Annexe 1: Liste de la définition des variables

Variables au niveau individuel
Sex – Sexe
Age - Groupe d'âge
Actsta - Statut de l'activité
Edlev - Niveau de formation
Marsta - État civil
Resid - Résidence
NUTS2 - Région NUTS2
NUTS3 - Région NUTS3

Variables au niveau du ménage
Rooms - Nombre de chambres disponibles du ménage
Tenure - Statut d'occupation
Car - Voiture
TV - Télé
PC - Ordinateur