

Introduction

Fondements  
de  
l'appariement  
statistique

Niveaux de  
validation  
selon Rässler  
sur  
l'appariement  
statistique

Une situation  
typique  
d'appariement

Harmonisation et  
réconciliation des  
sources multiples

Analyse du pouvoir  
explicatif pour les  
variables communes

Méthodes  
d'appariement

Evaluation de la  
qualité des résultats

Conclusions

# CONSIDÉRATIONS THÉORIQUES ET PRATIQUES CONCERNANT LES TECHNIQUES D'APPARIEMENT

Roxana ADAM, Institut National de la Statistique (Roumanie)

10ème COLLOQUE FRANCOPHONE SUR LES SONDAGES  
LYON, 2018

Introduction

Fondements  
de  
l'appariement  
statistique

Niveaux de  
validation  
selon Rässler  
sur  
l'appariement  
statistique

Une situation  
typique  
d'appariement

Harmonisation et  
réconciliation des  
sources multiples

Analyse du pouvoir  
explicatif pour les  
variables communes

Méthodes  
d'appariement

Évaluation de la  
qualité des résultats

Conclusions



## CENTENAIRE 2018

L'année où les Roumains marquent les 100 ans depuis ce qu'on appelle "la Grande union", c'est-à-dire l'union de la Transylvanie, de la Bessarabie, de la Bucovine et du Cadrilatere à ce qu'était alors le royaume de Roumanie.

# Le contenu

- 1 Introduction
- 2 Fondements de l'appariement statistique
- 3 Niveaux de validation selon Rässler sur l'appariement statistique
- 4 Une situation typique d'appariement
  - Harmonisation et réconciliation des sources multiples
  - Analyse du pouvoir explicatif pour les variables communes
  - Méthodes d'appariement
  - Evaluation de la qualité des résultats
- 5 Conclusions

Introduction

Fondements de l'appariement statistique

Niveaux de validation selon Rässler sur l'appariement statistique

Une situation typique d'appariement

Harmonisation et réconciliation des sources multiples

Analyse du pouvoir explicatif pour les variables communes

Méthodes d'appariement

Evaluation de la qualité des résultats

Conclusions

# Introduction

## Introduction

Fondements  
de  
l'appariement  
statistique

Niveaux de  
validation  
selon Rässler  
sur  
l'appariement  
statistique

Une situation  
typique  
d'appariement

Harmonisation et  
réconciliation des  
sources multiples

Analyse du pouvoir  
explicatif pour les  
variables communes

Méthodes  
d'appariement

Evaluation de la  
qualité des résultats

## Conclusions

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.

La méthode d'appariement connue sous le nom de Statistical Matching (SM) ou Data Fusion peut constituer la meilleure alternative pour combiner les données.

# Fondements de l'appariement statistique

Introduction

Fondements de l'appariement statistique

Niveaux de validation selon Rässler sur l'appariement statistique

Une situation typique d'appariement

Harmonisation et réconciliation des sources multiples

Analyse du pouvoir explicatif pour les variables communes

Méthodes d'appariement

Évaluation de la qualité des résultats

Conclusions

Fichier récepteur (source A)

Y	X <sub>1</sub>	[...]	X <sub>n</sub>
---	----------------	-------	----------------

Fichier donateur (source B)

X <sub>1</sub>	[...]	X <sub>n</sub>	Z
----------------	-------	----------------	---

Fichier apparié

Y	X <sub>1</sub>	[...]	X <sub>n</sub>	Z
---	----------------	-------	----------------	---

Figure 1: Appariement statistique

- la méthode **d'appariement statistique** (dénommé par la suite SM) pour deux sources de données A et B;
- 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".

# Fondements de l'appariement statistique

Introduction

Fondements  
de  
l'appariement  
statistique

Niveaux de  
validation  
selon Rässler  
sur  
l'appariement  
statistique

Une situation  
typique  
d'appariement

Harmonisation et  
réconciliation des  
sources multiples

Analyse du pouvoir  
explicatif pour les  
variables communes

Méthodes  
d'appariement

Évaluation de la  
qualité des résultats

Conclusions

- les techniques les plus utilisées sont **means of nearest neighbor** ou **hot deck**;
- sans tenir compte de la technique utilisée, on recommande la méthode SM lorsque le nombre d'unités statistiques identiques dans deux échantillons est zéro ou très petit;
- l'imputation des valeurs d'une variable est possible grâce à la "similitude" des unités statistiques des deux bases de données, basé sur l'**hypothèse implicite sur l'indépendance conditionnelle (CIA)** (Rässler, 2002);
- 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 (Rubin, 1987).

# Niveaux de validation selon Rässler sur l'appariement statistique

Introduction

Fondements de l'appariement statistique

Niveaux de validation selon Rässler sur l'appariement statistique

Une situation typique d'appariement

Harmonisation et réconciliation des sources multiples

Analyse du pouvoir explicatif pour les variables communes

Méthodes d'appariement

Évaluation de la qualité des résultats

Conclusions

## a. Conservation des valeurs individuelles

- La reproduction exacte des valeurs est possible seulement dans le cas où les variables communes  $X_1 \dots X_n$  déterminent déjà exactement la variable  $Y$ .

## b. Conservation des distributions communes

- 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.

# Niveaux de validation selon Rässler sur l'appariement statistique

Introduction

Fondements de l'appariement statistique

Niveaux de validation selon Rässler sur l'appariement statistique

Une situation typique d'appariement

Harmonisation et réconciliation des sources multiples

Analyse du pouvoir explicatif pour les variables communes

Méthodes d'appariement

Évaluation de la qualité des résultats

Conclusions

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

- Les variables qui sont indépendantes conditionnellement ne se trouvent pas, également, en corrélation conditionnellement. Pourtant la situation vice versa n'est généralement pas valable.

## d. Conservation des distributions marginales

- Représente ce qu'on souhaite être une exigence minimale de l'appariement statistique.
- 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.



# Une situation typique d'appariement

Introduction

Fondements  
de  
l'appariement  
statistique

Niveaux de  
validation  
selon Rässler  
sur  
l'appariement  
statistique

**Une situation  
typique  
d'appariement**

Harmonisation et  
réconciliation des  
sources multiples

Analyse du pouvoir  
explicatif pour les  
variables communes

Méthodes  
d'appariement

Évaluation de la  
qualité des résultats

Conclusions

## Simulation SM entre EBM 2014 et EU-SILC 2014

# Une situation typique d'appariement

Introduction

Fondements de l'appariement statistique

Niveaux de validation selon Rässler sur l'appariement statistique

Une situation typique d'appariement

Harmonisation et réconciliation des sources multiples

Analyse du pouvoir explicatif pour les variables communes

Méthodes d'appariement

Évaluation de la qualité des résultats

Conclusions

	EU-SILC 2014	EBM 2014
	<i>Statistiques de l'Union Européenne sur le revenu et les conditions de vie</i>	<i>Enquête sur le budget des ménages</i>
<b>Périodicité:</b>	annuel	trimestriel
<b>Unité d'observation:</b>	ménage	ménage
<b>Période de référence:</b>	Plusieurs périodes de référence (selon la nature des questions): - la semaine avant l'entrevue - au cours des 12 derniers mois - l'année calendaire précédente ( <b>le revenu</b> )	- mois calendaire
<b>Échantillon:</b>	7995 logements occupés de façon permanente/ annuel	9360 logements occupés de façon permanente/ trimestriel

**Objectif:** créer une base de données intégrée contenant des informations détaillées sur les revenus des ménages au niveau régional en appliquant la méthode d'appariement statistique.

# I. Harmonisation et réconciliation des sources multiples

Introduction

Fondements de l'appariement statistique

Niveaux de validation selon Rässler sur l'appariement statistique

Une situation typique d'appariement

**Harmonisation et réconciliation des sources multiples**

Analyse du pouvoir explicatif pour les variables communes

Méthodes d'appariement

Evaluation de la qualité des résultats

Conclusions

- l'harmonisation de la définition des unités;
- l'harmonisation de la période de référence;
- le complètement de la population;
- l'harmonisation des variables;
- l'harmonisation des classifications;
- l'ajustement des erreurs de mesurage (précision);
- l'ajustement des données manquantes;
- la dérivation des variables;

*Variables:*

- *Variables au niveau du ménage;*
- *Variables au niveau individuel.*

# I. Harmonisation et réconciliation des sources multiples

## ex. Niveau de formation

Définition des variables	Variable	EBM2014		EU-SILC 2014	
		Fréquence absolue	Fréquence relative (%)	Fréquence absolue	Fréquence relative (%)
<b>Variables au niveau individuel</b>					
<b>Niveau de formation</b>	Edlev				
Personnes âgées 0 - 15 ans (hors champ d'application d'EU-SILC)	0	5414	8.83	1732	9.99
Pas d'éducation	1	368	0.6	170	0.98
Enseignement primaire et secondaire (ISCED1+ISCED2+ISCED3)	2	21175	34.53	5836	33.68
Enseignement post-secondaire non-supérieur + Enseignement supérieur de cycle court (ISCED4+ISCED5)	3	28669	46.76	7850	45.3
Enseignement tertiaire (ISCED 6+7+8)	4	4996	8.15	1741	10.05

Introduction

Fondements de l'appariement statistique

Niveaux de validation selon Rässler sur l'appariement statistique

Une situation typique d'appariement

Harmonisation et réconciliation des sources multiples

Analyse du pouvoir explicatif pour les variables communes

Méthodes d'appariement

Évaluation de la qualité des résultats

Conclusions

## II. Analyse du pouvoir explicatif pour les variables communes

Introduction

Fondements de l'appariement statistique

Niveaux de validation selon Rässler sur l'appariement statistique

Une situation typique d'appariement

Harmonisation et réconciliation des sources multiples

Analyse du pouvoir explicatif pour les variables communes

Méthodes d'appariement

Évaluation de la qualité des résultats

Conclusions

- Donatiello et al.(2014) discutaient l'impossibilité de réaliser l'appariement statistique entre EBM et EU-SILC sous l'hypothèse de l'indépendance conditionnelle (CIA).
- Alternative: l'appariement statistique basé sur l'exploration de l'incertitude due à l'absence des informations communes sur les variables Y et Z.
- La distance Hellinger (HD) était utilisée, comme une mesure de la cohérence des variables communes, et aussi pour analyser la ressemblance/ différence des distributions des variables des deux ensembles de données.

## II. Analyse du pouvoir explicatif pour les variables communes

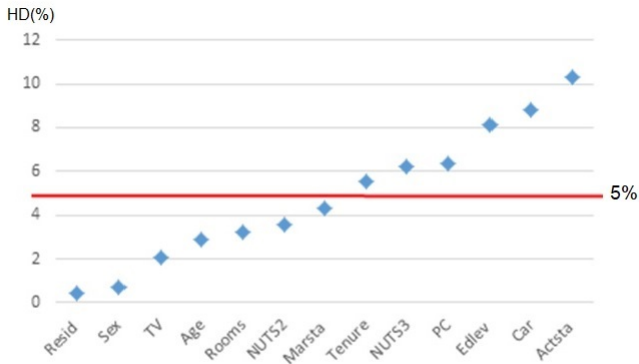


Figure 2: La distance de Hellinger (%) par les variables communes principales de EBM et EU-SILC, 2014

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

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

# III. Méthodes d'appariement

Introduction

Fondements  
de  
l'appariement  
statistique

Niveaux de  
validation  
selon Rässler  
sur  
l'appariement  
statistique

Une situation  
typique  
d'appariement

Harmonisation et  
réconciliation des  
sources multiples

Analyse du pouvoir  
explicatif pour les  
variables communes

**Méthodes  
d'appariement**

Evaluation de la  
qualité des résultats

Conclusions

## Techniques proposées pour SM au niveau micro:

- paramétriques (par exemple: l'imputation régressive);
- non-paramétriques (par exemple: l'imputation hot deck);
- mixtes (par exemple: des méthodes basées sur l'appariement prévisible).

(Donatiello et al., 2014)

# III. Méthodes d'appariement

Introduction

Fondements  
de  
l'appariement  
statistique

Niveaux de  
validation  
selon Rässler  
sur  
l'appariement  
statistique

Une situation  
typique  
d'appariement


Harmonisation et  
réconciliation des  
sources multiples

Analyse du pouvoir  
explicatif pour les  
variables communes

Méthodes  
d'appariement

Évaluation de la  
qualité des résultats

Conclusions

- l'aide de  et du paquet [StatMatch](#), par la fonction [NND.hotdeck](#);
- première intention était d'utiliser la fonction [NND.hotdeck](#) en incluant toutes les variables avec  $HD \leq 5\%$  pour attribuer le revenu de EBM dans l'UE-SILC;
- 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 réceptrice;
- pour le SM final, on a choisi comme variables d'appariement: [NUTS2](#) et [Age](#).



# III. Méthodes d'appariement

Introduction

Fondements de l'appariement statistique

Niveaux de validation selon Rässler sur l'appariement statistique

Une situation typique d'appariement

Harmonisation et réconciliation des sources multiples

Analyse du pouvoir explicatif pour les variables communes

Méthodes d'appariement

Évaluation de la qualité des résultats

Conclusions

```
#SM
names(samp.A) #EU-SILC

require(StatMatch)
group.v <- c("NUTS2")

X.mtc <- "age"
out.nnd <- NND.hotdeck(data.rec=samp.A, data.don=samp.B,
                       match.vars=X.mtc, don.class=group.v)
#donation classes are dedfined according to one or more categorical common variables
#donation classes are formed using large geographical areas ("NUTS2")
# while distances are computed on age ("age")

summary(out.nnd$dist.rd) # summary distances rec-don
summary(out.nnd$noad) # summary available donors at min. dist.

head(out.nnd$mtc.ids)
names(samp.B)

fA.nnd <- create.fused(data.rec=samp.A, data.don=samp.B,
                      mtc.ids=out.nnd$mtc.ids,
                      z.vars="inc_hbs")
head(fA.nnd) #first 6 obs.
```

Utilisant *mtc.ids* - identifié les donneurs correspondants du jeu de données EBM à EU-SILC.

	rec.id	don.id
[1,]	"588"	"30609"
[2,]	"589"	"36868"
[3,]	"610"	"51712"
[4,]	"613"	"31858"
[5,]	"620"	"45821"
[6,]	"626"	"31594"

# IV. Evaluation de la qualité des résultats

Introduction

Fondements de l'appariement statistique

Niveaux de validation selon Rässler sur l'appariement statistique

Une situation typique d'appariement

Harmonisation et réconciliation des sources multiples

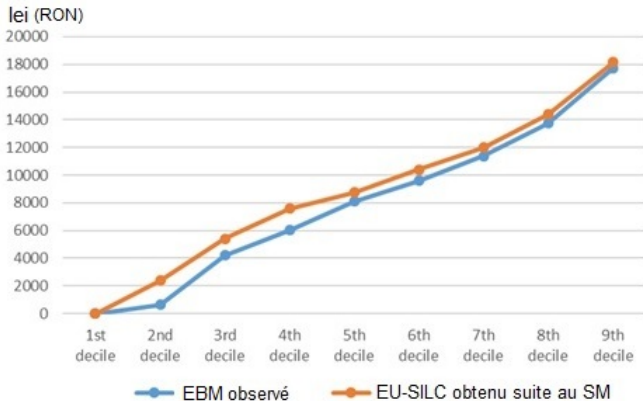
Analyse du pouvoir explicatif pour les variables communes

Méthodes d'appariement

Evaluation de la qualité des résultats

Conclusions

### Le seuil du décile de revenu



La procédure d'appariement statistique a du succès si les distributions empiriques marginales et communes de  $X_1 \dots X_n$ , 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).

# Conclusions

Introduction

Fondements de l'appariement statistique

Niveaux de validation selon Rässler sur l'appariement statistique

Une situation typique d'appariement

Harmonisation et réconciliation des sources multiples

Analyse du pouvoir explicatif pour les variables communes

Méthodes d'appariement

Évaluation de la qualité des résultats

Conclusions

- la littérature spécialisée traite ce sujet le plus souvent dans une perspective théorique. En pratique, basées sur des données réelles, les techniques d'appariement présentent certaines limitations;
- la pratique nous montre que CIA est très difficile à respecter;
- l'évaluation de la base de données synthétique obtenue suite à la procédure d'appariement est importante;
- la recherche et le développement des techniques d'appariement doivent toujours tenir compte de la conservation de la qualité des données résultantes;
- 3 risques pour la qualité des données:
  - calcul du poids nouveaux;
  - lointaines du contraintes méthodologiques;
  - diminuer la précision.

# Bibliographie courtes:



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



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



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



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



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

10<sup>e</sup> COLLOQUE FRANCOPHONE  
SUR LES SONDAGES

24 au 26 octobre 2018

UNIVERSITE DE LYON - FRANCE



SFDS

Lyon 1

UNIVERSITE  
LAURENCE  
LYON 2

INSTITUT  
DES SCIENCES  
ET DE LA  
STATISTIQUE

Institut  
Cantile  
Jordan



Institutul National de Statistica



**Merci pour votre attention!**

[roxana.adam@insse.ro](mailto:roxana.adam@insse.ro)

► [www.insse.ro](http://www.insse.ro)