Rapport de méthodes

Fusion de données

0

Statistische Grundlagen und Übersichten
Bases statistiques et produits généraux
Basi statistiche e presentazioni generali

Etude de faisabilité

Office fédéral de la statistique
Bundesamt für Statistik
Ufficio federale di statistica
Uffizi federal da statistica
Swiss Federal Statistical Office

OFS BFS UST

Neuchâtel, 2004

Statistik der Schweiz

Die vom Bundesamt für Statistik (BFS) herausgegebene Reihe «Statistik der Schweiz» gliedert sich in folgende Fachbereiche:

0 Statistische Grundlagen und Übersichten
1 Bevölkerung
2 Raum und Umwelt
3 Arbeit und Erwerb
4 Volkswirtschaft
5 Preise
6 Industrie und Dienstleistungen
7 Land- und Forstwirtschaft
8 Energie
9 Bau- und Wohnungswesen
10 Tourismus
11 Verkehr und Nachrichtenwesen
12 Geld, Banken, Versicherungen
13 Soziale Sicherheit
14 Gesundheit
15 Bildung und Wissenschaft
16 Kultur, Medien, Zeitverwendung
17 Politik
18 Öffentliche Verwaltung und Finanzen
19 Rechtspflege
20 Einkommen und Lebensqualität der Bevölkerung
21 Nachhaltige Entwicklung und regionale Disparitäten

Statistique de la Suisse

La série «Statistique de la Suisse» publiée par l'Office fédéral de la statistique (OFS) couvre les domaines suivants:

0 Bases statistiques et produits généraux
1 Population
2 Espace et environnement
3 Vie active et rémunération du travail
4 Economie nationale
5 Prix
6 Industrie et services
7 Agriculture et sylviculture
8 Energie
9 Construction et logement
10 Tourisme
11 Transports et communications
12 Monnaie, banques, assurances
13 Protection sociale
14 Santé
15 Education et science
16 Culture, médias, emploi du temps
17 Politique
18 Administration et finances publiques
19 Droit et justice
20 Revenus et qualité de vie de la population
21 Développement durable et disparités régionales

Statistique de la Suisse

Rapport de méthodes

Fusion de données

Etude de faisabilité

Auteur **Monique Graf**
Office fédéral de la statistique

Editeur **Office fédéral de la statistique**

Office fédéral de la statistique
Bundesamt für Statistik
Ufficio federale di statistica
Uffizi federal da statistica
Swiss Federal Statistical Office

OFS BFS UST

Neuchâtel, 2004

Préambule

La loi sur la statistique fédérale, dans les principes de la collecte des données, prescrit que la Confédération renonce à organiser des relevés pour la statistique fédérale lorsqu'elle dispose des données requises. De plus, le nombre des enquêtes et des modes de collecte des données doit être limité au strict nécessaire. Dans ce cadre, l'office fédéral de la statistique cherche à harmoniser et à rassembler les informations contenues dans des sources diverses (registres administratifs ou enquêtes). Le présent rapport présente le résultat d'un mandat donné par le conseil de direction au service de méthode statistique concernant une évaluation de la faisabilité de différents types de fusion de données. L'auteur remercie Ph. Eichenberger (METH) pour de nombreuses et intéressantes discussions, ainsi que Claire Jobin (KULT) pour ses remarques fort utiles.

Résumé

Il existe deux types de fusion de données: le couplage d'enregistrements et l'appariement statistique. On procède à un couplage lorsque la même unité statistique est recherchée dans deux sources de données différentes, et à un appariement lorsque la fusion des deux sources est basée sur la création d'unités synthétiques. Ces deux modes de fusion comportent quelques analogies techniques, mais sont des procédés de nature totalement différente, aboutissant à des fichiers de résultats qui n'ont pas les mêmes conditions d'application. Le rapport commence par présenter plus en détail ces deux modes de fusion, puis se termine par quelques recommandations. En Annexe, on trouve quelques applications tirées de la littérature internationale, ensuite les principales notions et difficultés liées à l'appariement statistique sont illustrées par des exemples.

Mots-clé

rapport de méthodes; record linkage; statistical matching; couplage; appariement; indépendance conditionnelle.

Complément d'information: Monique Graf, tél. 032 713 66 15
monique.graf@bfs.admin.ch

Réalisation: Service de méthodes statistiques, OFS

Diffusion: Office fédéral de la statistique
CH-2010 Neuchâtel
Tél. 032 713 60 60 / Fax 032 713 60 61
Order@bfs.admin.ch

Internet: http://www.statistik.admin.ch

Numéro de commande: 338-0022

Prix: gratuit

Série: Statistique de la Suisse

Domaine: 0 Bases statistiques et produits généraux

Langue originale: Français

Graphisme/Layout: OFS

Copyright: OFS, Neuchâtel 2004
La reproduction est autorisée, sauf à des fins commerciales, si la source est mentionnée.

ISBN: 3-303-00275-4

Table des matières

1 Introduction **5**

2 Types de fusion **5**
2.1 Couplage d'enregistrements 5
2.2 Couplage et protection des données 5
2.3 Appariement statistique 6
2.4 Comparaison entre couplage, appariement et extrapolation 7

3 Recommandations et conclusion **8**
3.1 Problèmes logistiques dans le cas de couplage entre registres 8
3.2 Recommandations spéciales au sujet des appariements 8
3.3 Recommandation générale 9

Annexes **10**
A Exemples de couplages d'enregistrement 10
B Exemples d'appariement statistique 10
C Indépendance conditionnelle 12
D Harmonisation des univers 14
E Distribution conditionnelle 19

Bibliographie **23**

1 Introduction

Dans une fusion de données, on réunit deux ou plusieurs fichiers de données en un seul, dans le but d'harmoniser et de compléter des sources d'information disjointes au départ. Il existe deux types de fusion de données : le couplage d'enregistrements (anglais : *record linkage*) et l'appariement statistique (anglais : *statistical matching*). On procède à un couplage lorsque la même unité statistique est recherchée dans deux sources de données différentes, et à un appariement lorsque la fusion des deux sources est basée sur la création d'unités synthétiques. Ces deux modes de fusion comportent quelques analogies techniques, mais sont des procédés de nature totalement différente, aboutissant à des fichiers de résultats qui n'ont pas les mêmes conditions d'application. Le rapport commence par présenter plus en détail ces deux modes de fusion, puis se termine par quelques recommandations. En Annexe, on trouve quelques applications tirées de la littérature internationale, ensuite les principales notions et difficultés liées à l'appariement statistique sont illustrées par des exemples.

2 Types de fusion

Il existe deux grands types de méthodes que l'on peut ranger sous le terme de fusion de microdonnées : les *couplages d'enregistrements* (*record linkage*), où l'on cherche à identifier la même unité statistique dans deux sources de données (enquêtes ou registres), et les *appariements statistiques* ou greffes de données, basées sur l'appariement entre sosies définis par des valeurs proches sur un ensemble de variables communes à deux sources distinctes ne contenant pas nécessairement les mêmes unités.

2.1 Couplage d'enregistrements

Gill (2001) représente la référence la plus complète sur le sujet du couplage. Winkler (2001) donne une liste de logiciels de couplage. On distingue deux types de couplage : les *couplages exacts* et les *couplages probabilistes*. Dans les premiers, chaque unité est repérée dans chaque source de données par un numéro d'identification unique, et le couplage est alors simple et rapide. Les couplages probabilistes sont basés sur des variables identifiantes qui sont soit entachées d'erreur, soit codées diversement. La première tâche consiste donc à les harmoniser. Le couplage proprement dit est vu comme une décision : couplage certain, probable, improbable, impossible. A chaque décision est attachée une probabilité d'erreur qu'on essaie de minimiser. Quelques exemples internationaux sont présentés à l'Annexe A.

2.2 Couplage et protection des données

Le couplage d'enregistrements peut poser d'importants problèmes de protection de données. L'article 18, alinéa 1 de la loi sur la protection des données précise que «L'organe fédéral qui collecte systématiquement des données est tenu de préciser le but et la base juridique du traitement, les catégories de participants au fichier et de destinataires des données.» La principale question

à se poser est la suivante : *est-ce que le couplage implique l'usage d'informations personnelles - nouvelles ou anciennes - dans un but différent du but original ?*
En Grande Bretagne le Cabinet Office (2002) propose un cadre analytique pour les activités de couplage sous forme de 6 questions que voici :

1. Quel est l'objectif général de la procédure ?
2. Quels sont les bénéfices du couplage proposé ?
3. Quelle méthode de couplage est envisagée, et quelles sont les alternatives ?
4. Quels sont les coûts et les risques du couplage ?
5. Quel est le rapport risque - bénéfice ?
6. Quelles sont les mesures prises pour maximiser les bénéfices et minimiser les risques ?

Ce cadre analytique est développé pour 3 applications.

2.3 Appariement statistique

Dans l'appariement statistique, on fabrique un individu synthétique en combinant l'information sur deux individus "proches" (dans un sens à définir) provenant de deux sources différentes.
Considérons deux fichiers de données A et B provenant de deux enquêtes. Les enquêtés ne sont donc pas nécessairement les mêmes dans chaque enquête. Supposons que A et B possèdent un certain nombre de variables en commun, auxquelles on donne le nom collectif de X. Les variables spécifiques de A sont désignées par Y, et celles de B par Z. Ainsi A = (X ; Y) et B = (X ; Z). Le but de la fusion de données est de former un fichier combiné (X ;Y ; Z), de façon à pouvoir reconstituer la distribution conjointe de Y et Z. Ce procédé s'appelle "*statistical matching*" aux USA et au Canada, et "*Data fusion*" au Royaume-Uni (Heady et al., 1997). On parle en France de "*greffe de données*" (Saporta, 2002), et au Canada francophone d'"*appariement statistique*" (Singh et al., 1993), enfin en Allemagne de "*Datenfusion*" (Rässler, 2002). Voir à l'Annexe B quelques exemples et domaines d'application.

Le bon sens dicte qu'il ne peut pas être possible de reconstituer en toute circonstance la distribution conjointe de Y et de Z. Les questions que l'on peut se poser alors sont (Radner, 1981) :

1. Quelle est la précision des appariements statistiques ?
2. Pour quel but et sous quelles conditions les résultats d'une fusion sont-ils d'une précision suffisante ?
3. Quels sont les facteurs déterminants pour la précision des résultats ?
4. Quelles sont les méthodes optimales de fusion ?
5. Comment aborder les problèmes pratiques tels que ceux résultant de différences entre échantillons et entre les variables X ?
6. Quelle est la sensibilité des résultats aux hypothèses faites pour effectuer la fusion ?
7. Quelle est la précision relative de méthodes concurrentes ?

En 1972 déjà, Sims avait formalisé la condition nécessaire pour qu'il soit possible de reconstituer un fichier de données (X ; Y ; Z) à partir de deux fichiers (X ; Y) et (X ; Z) sans perte d'information et sans distorsion des relations entre Y et Z. Cette condition est celle de l'indépendance conditionnelle entre Y et Z sachant X (pour la définition et un exemple, voir Annexe C, et pour des considérations plus théoriques, voir Annexe E). Les études de qualité des fusions ont donc porté sur

1. la plausibilité de l'hypothèse d'indépendance conditionnelle dans des situations réelles
2. la sensibilité des résultats à cette hypothèse, à l'aide de simulations
3. les moyens d'élargir cette condition très restrictive

2.4 Comparaison entre couplage, appariement et extrapolation

Le couplage d'enregistrements et l'appariement statistique ont quelques analogies techniques, mais sont réellement des activités de nature totalement différentes.
Selon la loi sur la statistique, article 4, 3^{e} alinéa l'utilisation des données administratives doit prévaloir sur le prélèvement d'informations par enquête, lorsque ces données sont disponibles. Cette loi encourage donc formellement le couplage. Ce serait à notre avis une interprétation abusive de la loi que de prétendre qu'elle exige de la statistique fédérale qu'elle procède à des appariements statistiques, c'est-à-dire à la création d'individus synthétiques comme substituts à des personnes réelles. L'appariement statistique est en réalité une micro simulation, une création de données sur la base d'un modèle a priori qui peut être utile, mais dans un but précis, fixé à l'avance. *On crée des données, mais pas d'information.*

Il y a conflit entre les deux objectifs suivants :

1. limiter au strict nécessaire la collecte des données à la source,
2. organiser les enquêtes de telle sorte que de nombreuses informations soient disponibles en même temps.

L'appariement statistique voudrait concilier ces deux objectifs, mais n'y parvient que très partiellement, dans la mesure où le fichier de données résultant n'est utilisable que sous certaines hypothèses et pour certaines analyses plus ou moins prévues à l'avance.

Ces remarques ne sont pas faites pour décourager une organisation concertée et harmonisée des enquêtes. Les extrapolations sur la base de grandes enquêtes comme l'ESPA sont appliquées depuis longtemps à l'OFS et ont prouvé leur efficacité pour garantir la cohérence des résultats entre enquêtes plus détaillées, mais menées sur un plus petit nombre de personnes. Les extrapolations sont d'ailleurs aussi basées sur des hypothèses et ne sont pas non plus de portée universelle. Cependant, dans le cas des extrapolations, ces hypothèses de base sont largement dépendantes de l'organisation de l'enquête, en particulier du plan d'échantillonnage, et peuvent donc être contrôlées. Les poids d'extrapolation permettent de publier des tables jusqu'à un certain niveau de détail, et pas au-delà. Les poids ont certes une influence sur les corrélations estimées des variables, mais

l'appariement statistique n'est pas de même nature, car les corrélations entre les variables Y et Z sont totalement construites et dépendantes du modèle[1].

3 Recommandations et conclusion

3.1 Problèmes logistiques dans le cas de couplage entre registres

Sur le plan logistique, le couplage de registres peut poser de nombreux problèmes techniques et organisationnels. Il serait souhaitable de répondre à l'avance aux questions suivantes :

- Si les registres proviennent d'administrations différentes, qui est le maître du fichier résultant du couplage ?
- Qui gère les mises à jour ?
- Qui s'occupe des problèmes de protection des données ?

Cependant tout effort d'harmonisation serait une plus-value pour les registres.

3.2 Recommandations spéciales au sujet des appariements

- Définir précisément dans quel but l'appariement entre enquêtes est envisagé et faire une étude de faisabilité. Le but est par exemple d'appliquer un modèle de micro simulation.
- Ne pas sous-estimer le travail nécessaire dans la phase d'harmonisation des univers et des variables. Ce travail est à la fois analytique et technique. Les liens entre variables créés par la fusion dépendent de manière cruciale de cette étape.
- Prévoir des études de sensibilité aux hypothèses de modèle[2].

Citons quelques remarques de Heady et Evers (1997) et commentaires de Holt (1997) au sujet des appariements :

- Les données construites par appariement statistique ne peuvent jamais être utilisées pour mener des recherches sur les relations entre les variables spécifiques provenant de sources différentes, car ces relations sont des conséquences des hypothèses à la base de la fusion.
- Il existe des méthodes pour approximer ces relations à l'aide d'informations provenant d'une source externe, mais on ne peut pas conduire de tests empiriques sur la validité des hypothèses ainsi posées.
- Il y a forcément quelque défaut à cette source externe, sinon l'étude des variables Y et Z pourrait se faire à l'aide de cette source.

[1] Voir en Annexe D l'exemple (artificiel) sur les salaires, où les variables Y (salaire en 1995) et Z (salaire en 2000) ont une corrélation théorique de 1 en cas de couplage, mais de 0.49 en cas d'appariement.

[2] En Annexe D, sur l'exemple déjà cité, on combine deux hypothèses d'harmonisation des univers avec deux méthodes de fusion pour obtenir 4 résultats différents.

3.3 Recommandation générale

Faire dans chaque cas une évaluation du rapport bénéfices / risques.

Bénéfices d'un couplage :

- regroupement d'informations sur une même unité (personne, ménage, entreprise) ;
- meilleure mise en valeur de l'information existante.

Risques d'un couplage :

- activité pouvant violer la loi sur la protection des données ;
- risques d'erreur lorsque l'unité est repérée par des variables identifiantes (et non un identificateur unique) ;
- problèmes avec les univers des deux enquêtes (ou de l'enquête et du registre) à fusionner qui ne sont pas nécessairement les mêmes, d'où découlent des difficultés au niveau des poids d'extrapolation.

Bénéfices d'un appariement :

- meilleur accès à l'information à l'origine dispersée dans des sources différentes ;
- permet d'éviter des enquêtes trop longues aboutissant à un taux de réponse médiocre ;
- harmonisation facilitée entre pays ;
- utile dans un but précis, par exemple un modèle de micro simulation.

Risques d'un appariement :

- l'appariement statistique ne remplace pas une enquête ;
- risque d'induire l'utilisateur en erreur sur la généralité des relations entre variables : il ne peut pas y avoir création d'information ;
- problèmes avec les univers des deux enquêtes à fusionner qui ne sont nécessairement les mêmes, d'où découlent des difficultés au niveau des poids d'extrapolation ;
- difficultés dans l'harmonisation des variables de fusion entre les deux enquêtes ;
- si l'appariement est fait sans précautions, il peut même y avoir perte d'information.

Annexes

A Exemples de couplages d'enregistrement

Les couplages exacts sont praticables pour les enquêtes auprès des personnes et des ménages, si l'on dispose, comme aux Pays-Bas, d'un identificateur personnel pour toute la population. Cet identificateur permet (à terme) d'établir un système cohérent entre les enquêtes, et de faire des liens entre variables basés sur les personnes (voir le volume spécial de *Netherlands Official Statistics*). Dans ce volume, l'article de Arts et al. (2000) montre que tous les problèmes méthodologiques ne sont pas résolus pour autant : citons par exemple, le problème du choix de la pondération lorsque le même individu n'a pas le même poids dans les deux enquêtes.

- Au Canada, Krewski et al. (2001) rapportent que la Base Canadienne de Données sur la mortalité BCDM a été mise en couplage avec les recensements de la population et de l'agriculture de 1971 et qu'en 1997, le registre national des doses (RND) a été couplé avec la BCDM pour examiner les liens entre la mortalité et l'exposition à de faibles doses de rayonnements ionisants. Ces auteurs s'intéressent à l'impact des erreurs de couplage sur l'estimation de la mortalité par cohorte.
- Au Royaume-Uni, depuis 1971, la ONS Longitudinal Study (LS) est basée sur un couplage anonymisé entre des données individuelles du recensement et le National Health Service Central Register (NHSCR), permettant d'ajouter des éléments de biographie tels que naissance, décès, veuvage, occurrence d'un cancer, etc. Au départ, 4 jours calendaires ont été choisis, et tous les individus ayant leur anniversaire en ces jours sont inclus dans l'étude. La LS contient donc environ 1% de la population recensée. Pour des raisons de confidentialité, chaque personne de la LS possède un numéro d'identification unique qui figure dans le NHCR. Chaque nouvel événement du NHCR est introduit régulièrement dans la LS en utilisant le numéro d'identification. Smith et al. (2002) donnent une présentation concise de la LS, avec des recommandations au sujet de la qualité.
- En France, l'échantillon démographique permanent (EDP) existe depuis 1968. Il est construit sur le même principe que la LS (4 jours d'anniversaire). L'EDP est basé sur les recensements et les registres d'état civil. Les biographies sont donc constituées par les bulletins du recensement et les actes d'état civil des mêmes individus au cours du temps : naissances, mariages, naissance d'enfants, etc.

B Exemples d'appariement statistique

B.1 Domaine du marketing

Dans ce domaine, la fusion de données est très utilisée. L'image qui en est donnée (à l'extérieur en tout cas) est très peu critique. Voyons deux exemples :
L'institut Médiamétrie en France définit ainsi la fusion d'enquêtes : "Traitement statistique consistant à utiliser les réponses de deux enquêtes portant sur des échantillons et des sujets différents, afin de créer des individus virtuels qui auraient répondu à une seule enquête fusion des deux. Cette technique repose sur la recherche des sosies, c'est à dire d'individus considérés comme proches

sur un ensemble de variables commun aux deux enquêtes". Un panel ainsi constitué est le panel AUDE, produit commun à Médiamétrie et SECODIP, basé sur une fusion statistique entre les 2000 ménagères du panel Médiamat et les 9000 ménagères des différents panels de consommateurs SECODIP (Agence Média JFC, France).

En Allemagne, l'institut GfK (Gesellschaft für Konsum-, Markt- und Absatzforschung) introduit des informations du panel de consommateurs ConsumerScan dans le panel de télévision AGF (Arbeitsgemeischaft Fernsehforschung) par la même technique (Paarbildungsprozess), "mariant" un consommateur avec un téléspectateur sur la base de caractéristiques dépendant du but recherché.

B.2 Modèles de micro simulation

Le projet EUROMOD, financé par le *Programme communautaire en matière de recherche socio-économique spécifique (CT97-3060)*, concerne la réalisation d'un micro modèle de simulation à l'échelle européenne destiné à étudier les effets des politiques fiscales et parafiscales, notamment du point de vue de l'étude de la distribution des revenus. Ce projet est basé, pour ce qui concerne le Royaume-Uni, sur le projet POLIMOD, développé à l'Université de Cambridge. Dans ce projet les données sont obtenues par fusion entre la Family Resources Survey (FRS) et la Family Expenditure Survey (FES). Le procédé de fusion est décrit dans plusieurs notes, voir Dayal et al. (2000). Les points importants sont : a) comme les univers des deux enquêtes ne sont pas les mêmes, il n'a pas été possible de trouver des pondérations des données qui soient telles que la distribution des variables communes soit identique (même s'il y a calage sur des variables démographiques ou autres, il n'y a aucune raison de penser que la distribution des revenus dans les deux enquêtes sera similaire pour d'autres caractéristiques de la population) ; b) le stade d'harmonisation des définitions des variables communes entre les deux enquêtes a clairement demandé un très grand investissement et a un impact sur le résultat final qui est très difficile à quantifier ; c) on peut définir plusieurs objectifs à la fusion qui impliquent des procédures différentes, par exemple : on peut essayer de reproduire dans la FRS la répartition des dépenses de la FES, ou alors on peut préférer imputer la répartition attendue des dépenses de la FRS. La première option sera difficile à réaliser à un niveau de détail suffisant, et pour la seconde, il n'existera pas d'information indépendante permettant d'évaluer le succès de l'opération. *Cela prouve qu'il est impossible d'enrichir un jeu de données par fusion de deux enquêtes, de façon à répondre à tous les besoins.* Dans le cas du projet POLIMOD, les auteurs ont cherché une méthode de fusion telle que les résultats du modèle de simulation des effets des politiques fiscales et parafiscales obtenus à l'aide de la seule FES soient reproduits de façon satisfaisante avec les dépenses imputées dans la FRS.

B.3 Micro simulation au Canada

Statistique Canada dispose de plusieurs modèles de micro simulation : BD/MSPS, modèle de micro simulation transversal d'individus et de familles pour l'analyse de politiques fiscales ; LifePaths, modèle longitudinal dynamique d'individus et de familles créant à l'aide d'équations sur le comportement des cheminements de vie individuels complets ; POHEM, modèle longitudinal de micro simulation de la santé et des maladies ; Xecon, modèle expérimental non empirique d'entreprises et de consommateurs. StatCan a aussi développé Modgen, outil d'aide à la construction et

l'utilisation de modèles longitudinaux[3].
Il importe de voir que les données obtenues par fusion reposent dans tous ces cas sur un modèle complexe élaboré dans un but précis. Il serait donc délicat de les utiliser pour d'autres études.

B.4 Fusions de données en Italie

D'Orazio et al (2003) citent plusieurs exemples de fusion réalisés à ISTAT. Ils distinguent les objectifs *micro* et *macro*. Dans l'objectif micro, le but est de reconstituer des unités statistiques synthétiques à partir de deux sources portant sur des unités différentes de la même population. Les auteurs citent quelques évaluations d'objectifs micro. L'objectif macro consiste à obtenir des tablex agrégés présentant l'information croisée sur des variables provenant de sources différentes. Ce dernier cas peut être abordé par des modèles de pondération. Un exemple d'un objectif macro à ISTAT est la construction d'une matrice de comptabilité sociale, basée sur une enquête de la Banque d'Italie et deux enquêtes de ISTAT. Nous nous intéressons ici aux objectifs micro.

C Indépendance conditionnelle

Sous l'hypothèse d'indépendance conditionnelle, l'appariement statistique peut se faire sans biais ni perte d'information. Il est donc très important de déterminer si cette condition est remplie. Comme nous allons le voir, ce n'est malheureusement que rarement le cas.

Définition

Il y a *indépendance conditionnelle* entre des variables Y et Z étant donné une variable X, lorsque dans tout sous-groupe d'unités statistiques défini par une valeur fixée de X, Y et Z sont indépendantes.

Exemple : Survie au naufrage du Titanic

Le célèbre naufrage du Titanic donne un excellent exemple de la signification de l'hypothèse d'indépendance conditionnelle.

La table 1 montre les données complètes sur la survie au naufrage en fonction du genre et de la classe du passager, ou de l'appartenance à l'équipage. On voit que seulement 2.8% des passagères de $1^{ère}$ classe se sont noyées (l'histoire dit que ces 4 romantiques n'ont pas voulu abandonner leur mari). Celles voyageant en 3^{e} classe ont eu moins de chance, puisque 54% d'entre elles sont décédées. Quant aux hommes, la différence entre classes est moins criante, mais la survie terriblement plus faible que pour les femmes, puisque 66% des voyageurs de $1^{ère}$ classe et 83% de ceux de 3^{e} ont disparu. Cela signifie que le taux de survie dépend à la fois du genre et de la classe.
Supposons que nous disposions d'une part du fichier des passagers et membres d'équipage (table 2) et d'autre part de la statistique de décès par genre telle qu'elle aurait pu être donnée juste après le naufrage (table 3) :

[3] Voir http ://www.statcan.ca/francais/spsd/

TAB. 1 – Titanic données complètes

Genre	F				M			
Survie	0		1		0		1	
Classe	Total	%	Total	%	Total	%	Total	%
Crew	3	13.04	20	86.96	670	77.73	192	22.27
First	4	2.76	141	97.24	118	65.56	62	34.44
Second	13	12.26	93	87.74	154	86.03	25	13.97
Third	106	54.08	90	45.92	422	82.75	88	17.25

TAB. 2 – Fichier des personnes à bord

Genre	F		M	
Classe	Total	%	Total	%
Crew	23	2.60	862	97.40
First	145	44.62	180	55.38
Second	106	37.19	179	62.81
Third	196	27.76	510	72.24

La fusion des deux fichiers ne peut se faire que sur le genre, seule variable commune : on applique alors aux voyageurs de chaque classe le taux de survie selon le genre. C'est l'hypothèse "d'indépendance conditionnelle du taux de survie et de la classe, sachant le genre", autrement dit, l'adage "les femmes et les enfants d'abord" a la même validité dans toutes les classes, ou encore, connaissant le genre d'une personne, l'information sur sa classe ne précise rien de plus sur sa survie. On obtient alors la table 4 où le lien entre le taux de survie et la position sociale a disparu. Cette table n'a donc que l'apparence de données complètes.

Le taux de survie par sexe, constant dans toutes les classes, a quelque chose d'artificiel : la fusion a été effectuée *sans contrainte* (unconstrained match) et le résultat est que la variance conditionnelle (étant donné le sexe) du taux de survie estimé est nulle. Pour rendre le fichier résultant plus réaliste, on peut effectuer un tirage aléatoire dans chaque cellule de la table 2, avec probabilité de succès égale au taux de survie donné dans la table 3, et contraintes sur les marges : c'est alors une fusion "*avec contraintes*". Le nombre de survivants estimé par cellule est, dans ce cas, donné par le nombre de "succès" (tirage de la survie pour une personne de la cellule).

Cet exemple montre qu'on ne peut pas créer d'information par appariement statistique sous l'hypothèse d'indépendance conditionnelle. Ainsi l'expression "enrichissement d'un jeu de données par appariement statistique" est totalement abusive.

TAB. 3 – Taux de survie selon le genre

Survie	0		1	
Genre	Total	%	Total	%
F	126	26.81	344	73.19
M	1364	78.80	367	21.20

TAB. 4 – Survie estimée par fusion sans contrainte sur le genre

Genre	F				M			
Survie	0		1		0		1	
Classe	Total	%	Total	%	Total	%	Total	%
Crew	6	26.81	17	73.19	679	78.80	183	21.20
First	39	26.81	106	73.19	142	78.80	38	21.20
Second	28	26.81	78	73.19	141	78.80	38	21.20
Third	53	26.81	143	73.19	402	78.80	108	21.20

D Harmonisation des univers

Lors de la fusion des données provenant de sources différentes, la question de l'harmonisation des univers est extrêmement délicate et a une importance décisive pour les résultats. Dans certains cas, suivant la manière dont les variables communes sont harmonisées, même le sens de la relation entre variables spécifiques aux deux sources est modifié. Le célèbre paradoxe de Simpson permet d'illustrer ce fait.

D.1 Caveat au sujet des regroupements : le paradoxe de Simpson

Le paradoxe de Simpson se réfère à l'observation qu'il est possible de changer le sens d'une association, ou le signe d'une corrélation par regroupement en classes. Il existe de nombreux exemples de ce paradoxe. Rinott & Tam en donnent une illustration dont nous nous inspirons ici. Il s'agit d'un exemple simulé, ce qui permet d'en expliquer complètement le mécanisme et aussi d'éclairer le lien entre le paradoxe et les difficultés liées à la fusion de données.

Considérons un échantillon simulé de 100 employés dont on connaît l'âge et le salaire (en milliers de francs) en 1995 (fig. 1). Supposons que la même information soit obtenue en 2000 sur les mêmes 100 employés. En comparant les deux colonnes de la table 5 sur le salaire moyen par groupe, on constate une diminution du salaire entre 1995 et 2000 dans chaque classe d'âge. Il semble que le salaire de ces 100 employés aie *diminué* dans la période de 5 ans. Cependant en construisant ces données, on a rajouté 1000.- à chaque salaire, si bien que le salaire moyen global a *augmenté* de la même quantité. Comment cette diminution apparemment paradoxale dans chaque

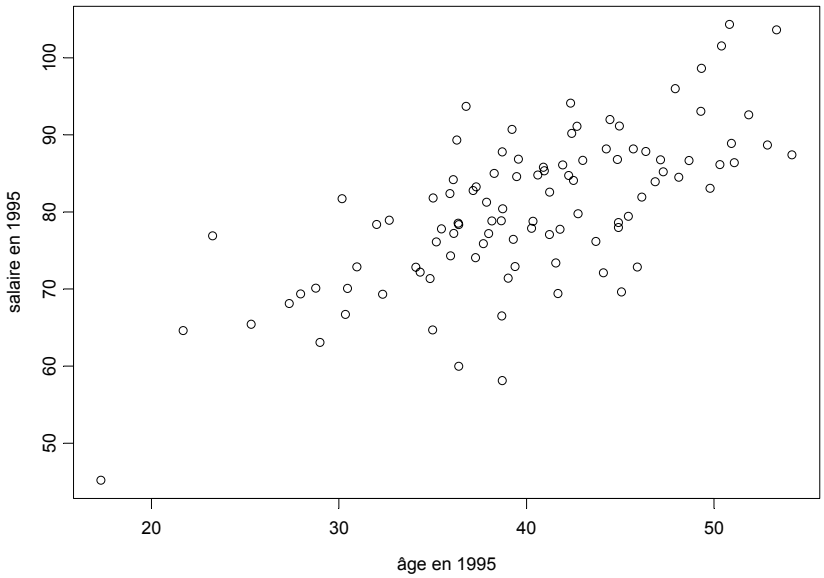


FIG. 1 – Simulation du salaire en fonction de l'âge

classe d'âge peut-elle se produire ?

TAB. 5 – Effectif et salaire moyen (en millier de CHF) des 100 employés

Classe d'âge	Effectif		Salaire	
	1995	2000	1995	2000
âge ≤ 40	50	19	75.4	70.6
$40 <$ âge ≤ 50	41	57	83.7	81.7
$50 <$ âge	9	24	93.3	89.2
Total	100	100	80.4	81.4

Les données de 1995 ont été obtenues par simulation de l'âge et du salaire $(X; S)$ selon une loi normale bivariée, avec moyenne et écart-type de l'âge = (40, 7), moyenne et écart-type du salaire = (80, 10) et coefficient de corrélation de 0.7 entre les deux variables.

Les données de 2000 $(X'; S')$ sont simplement obtenues par : $X' = X + 5$ et $S' = S + 1000$. Le coefficient de corrélation est donc toujours de 0.7 en 2000.

L'explication du paradoxe est simple : les employés qui avaient entre 35 et 40 ans, ou entre 45 et 50 ans en 1995, sont passés en 2000 dans la classe d'âge supérieure. Nous avons construit un exemple présentant une corrélation positive entre l'âge et le salaire. La diminution du salaire observée dans chaque classe d'âge provient du fait que les employés les plus âgés de la classe en

1995, qui tendent à avoir un salaire plus élevé, sont remplacés par des plus jeunes en 2000, qui sont moins bien payés. L'accroissement attendu, donné par la régression du salaire sur l'âge en 1995, est de 5000.-, alors que l'accroissement réel n'a été que de 1000.-. Le salaire a donc diminué d'environ 4000.- par rapport aux espoirs fondés sur la relation âge - salaire en 1995.

En quoi cet exemple nous éclaire-t-il sur la fusion de données ?

Il montre l'importance de l'harmonisation des univers. La population étant globalement la même, on voit le même accroissement de 1000.- au niveau global qu'au niveau individuel. Ce n'est pourtant pas le cas à l'intérieur des classes d'âge, car elles ne correspondent plus aux mêmes groupes d'individus. Dans chaque classe d'âge, le salaire moyen diminue, ce qui ne correspond à la réalité, ni au niveau individuel, ni au niveau global. Pour harmoniser les univers, il faudrait changer la définition des limites de classe, en rajoutant 5 ans en 2000, puisque la population a vieilli uniformément. L'harmonisation des variables de classe d'âge doit suivre l'évolution de la population. Comme celle-ci est formée des mêmes individus, l'harmonisation amène à modifier les limites.

Si au contraire, on était en présence d'une population qui se renouvelle, de telle sorte que la répartition en classes d'âge soit stable, on serait amené à ne pas modifier les limites de classe pour harmoniser les classes d'âge.

D.2 Deux modèles d'harmonisation – fusion par le rang

Considérons le problème de la fusion et imaginons deux fichiers A et B, où A contient l'âge et le salaire de 1995, et B les mêmes variables en 2000, mais pour des individus différents. Pour la fusion à l'aide de l'âge, faut-il utiliser les valeurs les plus proches entre X et X' ou faut-il soustraire 5 ans à X' ? Le fichier résultant de la fusion entre A et B sera une conséquence des hypothèses qui sont posées.

Supposons que 50 des 100 employés de la fig. 1 soient tirés au hasard et attribués à A, et les autres à B. On simule de cette façon le cas d'une même population dont on voit une partie en 1995 et une autre en 2000. Cette population a donc vieilli de 5 ans, et la fusion la plus naturelle doit se faire entre X et $X' - 5$ (sosies par la date de naissance). On verra alors une augmentation de 1000.- aussi pour les moyennes par classe.

Pour simuler une population qui se renouvelle, avec les mêmes relations âge - salaire qu'avant, il faut trouver des sosies tels que : $X' = X$ (sosies par l'âge) et alors $S' = S - 4000$.-.

La fig. 2 montre les deux possibilités. A la fig. 2(a), 50 employés de la fig. 1 sont tirés au hasard et forment le fichier A (le reste est attribué à B). La fig. 2(b) montrent les deux modèles pour le fichier B. Modèle 1 : population sans renouvellement (âge en 2000 = X'), et modèle 2 : distribution stable des âges (âge en 2000 = X), avec dans les deux cas la même relation théorique âge - salaire en 2000 (trait plein : droite des moindres carrés identique pour les deux nuages). Par construction de l'exemple, les salaires en 1995 et 2000 pour le même individu ont une corrélation de 1. Cette relation théorique, correspondant au couplage, est représentée par les pointillés à la fig. 2(c) pour un employé et son sosie par la date de naissance (cas du modèle 1) : il gagne 1000.- de plus en 2000 par rapport à 1995, et à la fig. 2(d) pour des employés sosies par l'âge (cas du modèle 2) : un employé gagne 4000.- de moins en 2000 qu'un collègue du même âge en 1995. On voit

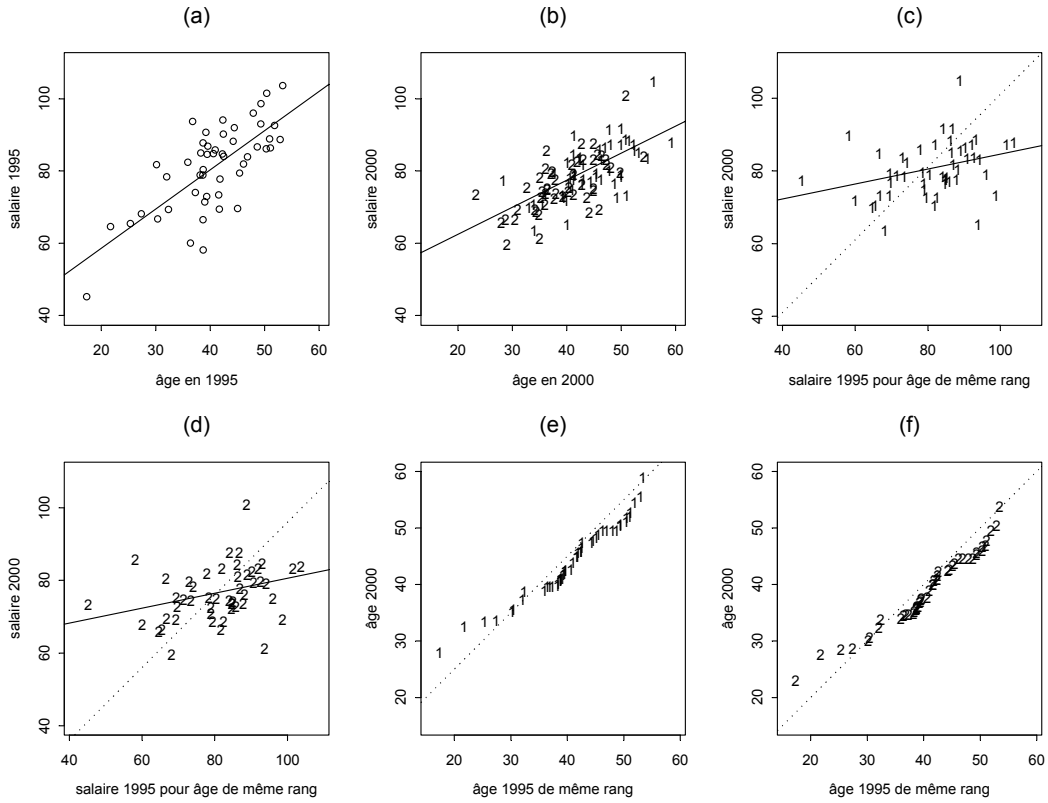


FIG. 2 – Comparaison des deux modèles pour la fusion par le rang de l'âge.
(a) fichier A, couples $(X; S)$;
(b) fichier B, 1 : couples $(X'; S')$ 2 : couples $(X; S - 4000)$;
(c) $1^{ère}$ fusion entre A et B, droite des moindres carrés (trait plein), relation théorique (pointillés) ;
(d) $2^{ème}$ fusion entre A et B, droite des moindres carrés (trait plein), relation théorique (pointillés) ;
(e) et (f) sosies par le rang de l'âge, modèles 1 et 2

une différence appréciable avec la relation obtenue par fusion (trait plein). La corrélation qui est de 1 lorsqu'on possède le lien au niveau de l'individu, n'est plus que de 0.325 pour des sosies du modèle 1. Si notre échantillon était suffisamment grand pour qu'on trouve pour chaque élément de B un sosie très proche dans A, la corrélation entre les deux variables de salaire serait de $(0.7)^2 = 0.49$ qui est la corrélation théorique sous l'hypothèse d'indépendance conditionnelle.

En tirant au hasard les éléments de A *et de* B *dans le fichier complet, on a construit un exemple qui respecte l'hypothèse d'indépendance conditionnelle du salaire 1995 d'un élément de* A *et du salaire 2000 d'un autre élément de* B *pour un même âge X (dans les deux modèles). En effet, par construction, le salaire ne dépend que de X et d'une erreur aléatoire indépendante. Cependant, les éléments de même rang dans les fichiers* A *et* B *ne sont pas toujours d'âge très proche. Une erreur supplémentaire est donc introduite. Le tirage au hasard garantit que cette erreur n'induit pas un biais systématique.*

D.3 Fusion par le plus proche voisin

Plutôt que de fusionner par le rang des âges, on pourrait chercher pour chaque élément de B l'élément d'âge le plus proche dans A. Certains éléments de A peuvent alors être pris deux fois, et d'autres jamais. Une représentation graphique de la fusion par le plus proche voisin pour les deux modèle de population est fournie à la fig. 3.

La corrélation résultante entre l'âge et le salaire reconstitué surestime la vraie valeur (dans notre exemple, elle est de 0.82 au lieu de 0.7).

TAB. 6 – Corrélation entre salaire 2000 et salaire 1995 reconstitué

	Fusion par le rang		Fusion par le plus proche voisin	
	corrélation	probabilité	corrélation	probabilité
Modèle 1	0.325	0.173	0.148	0.008
Modèle 2	0.325	0.173	0.530	0.708

La table 6 résume les résultats des fusions. Sous l'hypothèse d'indépendance conditionnelle, les deux variables de salaire du fichier B, le salaire 2000 et le salaire 1995 obtenu par fusion, ont une corrélation théorique de 0.49. La probabilité donnée est la p-valeur du test bilatéral d'égalité des corrélation empiriques avec cette valeur théorique (test de Fisher). On constate que la corrélation de 0.148 est significativement différente de la valeur théorique au seuil de 1%. La méthode de fusion par le plus proche voisin a mieux fonctionné que la méthode du rang dans le cas du modèle 2, mais moins bien pour le modèle 1.

D.4 Discussion

La méthode de fusion par le plus proche voisin est dite *sans contraintes*, car aucune restriction n'est posée sur le nombre de fois qu'une unité de A est utilisée pour l'appariement. La fusion par le

rang est une fusion *avec contraintes* ; elle garantit que les distributions marginales sont conservées. On voit clairement à la fig. 3 (d) que le plus âgé des employés du fichier A a été fusionné avec 4 employés du fichier B. Il n'a en revanche qu'un seul sosie dans B sous la méthode avec contraintes, voir fig.2 (f).

Résumons pour le modèle 1 (population sans renouvellement). Cet exemple montre que

1. si on fusionne simplement par classe d'âge (table 5), on *change le sens de la relation* entre les variables spécifiques (les 2 salaires) ;
2. si on fusionne des sosies, l'hypothèse d'indépendance conditionnelle est respectée dans cet exemple. La corrélation théorique entre les variables spécifiques n'est plus que de 0.49, alors qu'elle est de 1 si le lien au niveau individuel est connu ;
3. dans la fusion par le rang de l'âge, la corrélation estimée entre les variables spécifiques est de 0.325 ;
4. dans la fusion par le plus proche voisin, elle n'est plus que de 0.148.

Comme le dit Rodgers (1981, 1984, 1988), se basant sur plusieurs études de simulation, "le mieux qu'on puisse espérer avec l'appariement statistique est qu'il n'y ait pas de perte d'information, et en général même cet objectif n'est pas atteint".

E Distribution conditionnelle

Le *coefficient de corrélation partiel* $\rho_{YZ.X}$ est une mesure d'association entre Y et Z quand les valeurs de la variable X sont tenues fixes (voir par exemple Afifi et Azen, 1972). Il varie entre -1 et 1 et est lié aux coefficients de corrélation ordinaires par la relation

$$\rho_{YZ.X} = \frac{\rho_{YZ} - \rho_{YX}\rho_{ZX}}{\sqrt{(1-\rho_{YX}^2)\,(1-\rho_{ZX}^2)}}$$

Il représente la façon la plus simple de spécifier la loi conditionnelle de Y et Z sachant X. L'hypothèse d'indépendance conditionnelle implique que ce coefficient de corrélation partielle est nul, et dans ce cas, le coefficient de corrélation ordinaire ρ_{YZ} est égal au produit des coefficients de corrélation ordinaires de Y et X et de Z et X. Nous avons utilisé cette propriété dans l'exemple des salaires.

Puisque le coefficient de corrélation partiel mesure l'écart à l'indépendance conditionnelle, il ne peut pas être observé à l'aide du résultat d'un appariement statistique. En général la relation

$$-1 \;\leq\; \rho_{YZ.X} \leq 1$$

$$\Longleftrightarrow$$

$$\rho_{YX}\,\rho_{ZX} - \sqrt{(1-\rho_{YX}^2)\,(1-\rho_{ZX}^2)} \;\leq\; \rho_{YZ} \leq \rho_{YX}\,\rho_{ZX} + \sqrt{(1-\rho_{YX}^2)\,(1-\rho_{ZX}^2)}$$

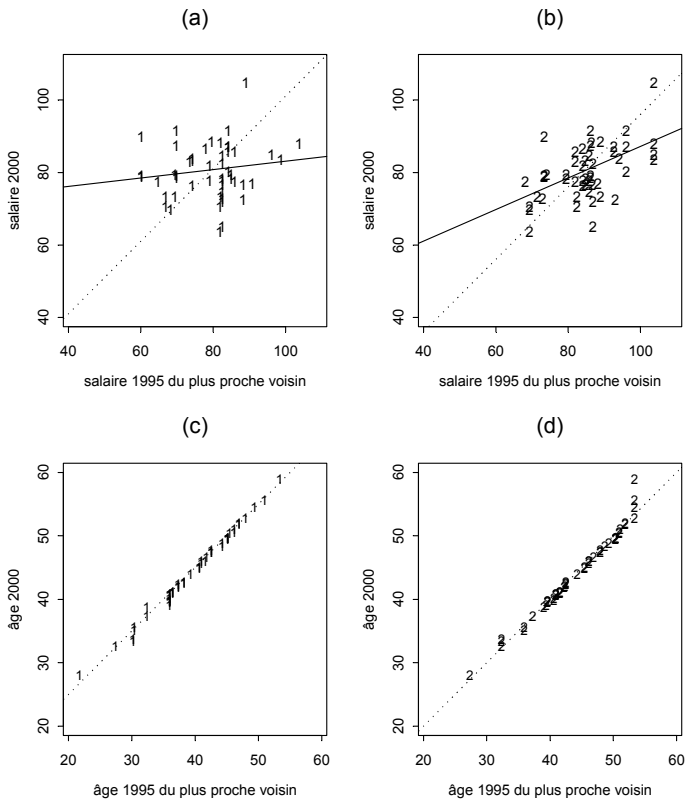


FIG. 3 – Fusion par le voisin le plus proche par l'âge pour les deux modèles.
(a) fusion avec X (fichier A) le plus proche de X'-5 (fichier B) ;
(b) fusion avec X (fichier A) le plus proche de X' (fichier B) ;
(c) âge des sosies pour (a) ;
(d) âge des sosies pour (b).

montre que l'hypothèse d'indépendance conditionnelle correspond à la valeur centrale de l'intervalle de variation de ρ_{YZ}. De plus, l'incertitude sur le coefficient de corrélation ordinaire croît rapidement lorsque les corrélations de Y et X et de Z et X diminuent. Dans l'exemple des salaires, ces corrélations étaient égales à 0.7 et l'intervalle de variation de la corrélation de Y et Z valant alors (-0.02, 1.00) devient donc très peu informatif. Il commence à prendre un sens pour des corrélations au moins égales à 0.9, ce qui est rare dans les enquêtes.

Ces considérations restent vraies lorsqu'il y a plusieurs variables de fusion (variables X). Il faudrait donc que les corrélations (multiples) des variables spécifiques avec les variables de fusion soient très élevées (supérieures à 0.95, par exemple), pour que l'on puisse affirmer en toute généralité quelque chose d'utile sur la corrélation non conditionnelle de Y et Z.

E.1 Echapper à l'indépendance conditionnelle ?

Morarity et Scheuren (2001) décrivent une procédure d'appariement qui préserve une valeur *spécifiée* de la corrélation entre Y et Z ou, plus généralement, de la matrice de covariance des variables spécifiques Y et Z. Ils traitent les fichiers A et B de façon symétrique, c'est-à-dire qu'ils reconstituent Z dans A et Y dans B. Cette valeur spécifiée de la covariance entre Y et Z permet de calculer les paramètres de la régression de Y sur X et Z et d'obtenir une estimation dans le fichier B. Ils rajoutent alors à $\widehat{Y}$ une erreur aléatoire e_Y de façon à ce que la matrice de variance-covariance de $\widehat{Y} + e_Y$ coïncide avec la matrice de variance-covariance de Y dans A. La loi de e_Y est la loi conditionnelle de Y sachant X et Z. Z est traité de manière symétrique. Ayant complété les deux fichiers avec l'information manquante, les auteurs procèdent à l'appariement statistique selon différentes méthodes du type plus proche voisin. Les meilleurs résultats (au sens de la conservation de la matrice de variance-covariance de Y et Z), sont obtenus en appariant les fichiers à l'aide des valeurs de $\widehat{Y} + e_Y$ et de $\widehat{Z} + e_Z$. L'appariement par les valeurs de X aboutit à l'indépendance conditionnelle des variables Y et Z pour les individus reconstitués.

Les auteurs ne commentent pas ce point, mais il est clair que leur méthode repose fortement sur l'hypothèse de normalité : en général, la matrice de covariance conditionnelle n'est pas constante, mais dépend de la valeur de X. D'autre part, si cette hypothèse de normalité est cruciale, on peut se demander quelle est l'utilité d'un appariement, une fois que les fichiers sont complétés. Moriarity et Scheuren ne donnent d'autre part aucune recommandation sur le choix en pratique de la valeur spécifiée de la covariance entre Y et Z. (On pourrait éventuellement l'obtenir grâce à une 3ème source de données C.)

E.2 Utilisation d'informations supplémentaires

En 1993, Singh et al. proposent plusieurs méthodes utilisant de l'information supplémentaire sur la distribution de Y et Z - ou X, Y et Z - provenant d'une source C. Les 8 méthodes évaluées dans l'article sont les suivantes : 4 variantes de régression et 4 variantes d'imputation par plus proches voisins (hot-deck).

Le but est de compléter A, et le point de vue est celui de la subsidiarité : les paramètres estimables dans A seulement sont évalués en n'utilisant que le fichier A ; sinon B est utilisé et en dernier lieu C, qui est supposé être une source de moins bonne qualité. On suppose que C ne

contient pas nécessairement de l'information au même niveau de détail que A et B. L'information contenue dans C est utilisée de deux manières différentes, mais toujours au niveau macro :

1. *par l'introduction des paramètres manquants de la loi conjointe de X, Y et Z*. Par une méthode analogue à celle de Moriarity et Scheuren, on calcule une valeur provisoire par régression[4] qui sera appariée avec une valeur authentique Z_{match} du fichier B par une méthode du plus proche voisin.
2. *par contraintes de catégorie*. Soit X^*, Y^* et Z^* les variables X, Y et Z catégorisées de telle façon que l'information existe dans les 3 fichiers[5]. On commence par un redressement des tables de contingence obtenues dans B et dans C, pour respecter les marges de la table $X^* \times Y^*$ dans A. Cela détermine des contraintes d'effectif pour la recherche de la valeur éligible Z_{match}. On doit alors tenir compte des distances pour choisir les valeurs Z_{match} : on apparie en priorité les paires les plus proches.

Les différentes variantes sont évaluées par simulation d'une loi normale multidimensionnelle. A notre avis, ces méthodes sont très ingénieuses, mais assez délicates à utiliser.

E.3 L'appariement statistique vu comme un problème de données manquantes

Rässler (2002) développe des alternatives aux méthodes d'appariement sans contrainte (plus proche voisin) et avec contrainte (appariement conservant les distributions marginales) en considérant le problème comme un problème d'imputation. Elle propose une procédure bayésienne non itérative (NIBAS). Les variables communes X sont quelconques, mais les variables spécifiques sont supposées normalement distribuées. L'aspect bayésien provient de la loi a priori posée sur la covariance conditionnelle de Y et Z sachant X. Les résultats de plusieurs tirages selon cette loi sont combinés pour fournir une évaluation de la variabilité résultante. Le principe est le même que la procédure d'imputation multiple de Rubin. Rässler compare sa méthode avec d'autres méthodes d'imputation multiples.

Cette approche est techniquement assez difficile, mais des programmes existent. La méthode de Moriarity et Scheuren peut être vue, dans sa phase d'estimation, comme un cas particulier de NIBAS, où la covariance conditionnelle est fixée a priori. Cela mène, du point de vue bayésien, à des estimations a posteriori impropres, ce que NIBAS évite. La méthode de Rässler s'arrête à l'imputation et ne pratique pas, comme Moriarity et Scheuren, de fusion proprement dite entre A et B (ce qui nous semble dans la logique des hypothèses posées).

L'avantage de l'approche par imputation est qu'il devient possible - sous de nombreuses hypothèses - d'incorporer de l'information donnée par une 3ème source qui pourrait être une enquête contenant toutes les variables d'intérêt, mais de taille beaucoup plus petite.

[4]Comme les auteurs ne rajoutent pas d'erreur aléatoire (à la manière de Moriarity et Scheuren), cette valeur n'a pas la bonne distributuion.

[5]On doit alors prendre garde au piège du paradoxe de Simpson !

Références

[1] Afifi, A.A. & Azen, S.P. (1972). *Statistical Analysis. A Computer oriented Approach*. Academic Press, New-York, 366 p.

[2] Arts, Koos, Bakker, Bart F.M. & van Lith, Erik (2000), "Linking administrative registers and household surveys", *Netherlands Official Statistics*, **15 (2)** , 16-22

[3] Cabinet Office (2002). "Privacy and Data-sharing. The Way foward for Public Services." http ://www.cabinet-office.gov.uk/innovation/2002/privacy/report/downloads/piu-data.pdf

[4] Dayal,Neela, Gomulka, Johanna, Mitton, Lavinia, Sutherland, Holly & Taylor, Rebecca (2000), "Enhancing Family Resources Survey income data with expenditure data from the Family Expenditure Survey : data comparison", *Department of Applied Economics, Univ. of Cambridge, Note MU/RN/40*

[5] D'Orazio, Marcello, Di Zio, Marco & Scanu, Mauro (2001), "Statistical matching : a tool for integrating data in National Statistical Institutes" *Proceedings of the "ETK-NTTS 2001*, 18-22 June 2001, Hersonissos (Crete),vol. I, pp. 433-441.

[6] D'Orazio, Marcello, Di Zio, Marco & Scanu, Mauro (2003), "Statistical Matching and Official Statistics", *Italian Review of Official Statistics - ISTAT Research Paper*, à paraître.

[7] Gill, Leicester (2001),"Methods for Automatic Record Matching and Linkage and their Use in National Statistics", *National Statistics Methodological Series*, **25**, 153 p.

[8] Heady, Patrick & Evers, Michael (1997), "A Brief Review of Data Fusion", *Document ONS, Central Methods Unit and Quality Division*

[9] Kamakura, Wagner A. & Wedel, Michel (1997), "Statistical data fusion for cross-tabulation", *Journal of Marketing Research*, **34** , 485-498

[10] Krewski, D., Wang, Y., Bartlett, S., Zielinski, J.M. & Mallick, R. (2001), "Impact d'erreurs d'appariement sur l'inférence statistique lors d'études de mortalité par cohortes", *Recueil du Symposium 2001 de Statistique Canada*

[11] Kroese, Bert, Renssen, Robbert H. & Trijssenaar, Marjolijn (2000), "Weighting or imputation : constructing a consistent set of estimates based on data from different sources", *Netherlands Official Statistics*, **15 (2)** ,23-31

[12] Moriarity, Chris & Scheuren, Fritz (2001), "Statistical matching : A paradigm for assessing the uncertainty in the procedure", *Journal of Official Statistics*, **17 (3)** , 407-422

[13] Radner, Daniel (1981), " Report on Exact and Statistical Matching Techniques, chap. III", *Statistical Policy Working Paper* **5**, Federal Comittee on Statistical Methodology.

[14] Rässler, Suzanne (2002), " *Statistical Matching A Frequentist Theory, Practical Applications and Alternative Bayesian Approaches* ", Springer Series Lecture Notes in Statistics **168**, 2002 XVIII, 238 p

[15] Rinott, Yoseph & Tam, Michael, "Monotone regrouping, Regression and Simpson's paradox", *document internet*.

[16] Rodgers, Willard L. (1988), "Statistical matching", *Encyclopedia of statistical sciences (Vol. 1)*, **8** 663-664

[17] Rodgers, Willard L. (1984), "An evaluation of statistical matching", *Journal of Business and Economic Statistics*, **2** , 91-102

[18] Rodgers, Willard L. & DeVol, Edward B. (1981), "An Evaluation of Statistical Matching", *ASA Proceedings of the Section on Survey Research Methods*, 128-132

[19] Saporta, Gilbert (2002), "Data fusion and data grafting", *Computational Statistics and Data Analysis*, **38 (4)** , 465-473

[20] Sims, Christopher A. (1972), "Comments". *Annals of Economic and Socual Measurements*, **1**, 343-346

[21] Singh, A. C. , Mantel, H. J. , Kinack, M. D. , and Rowe, G. (1993), "Statistical matching : Use of auxiliary information as an alternative to the conditional independence assumption", *Survey Methodology*, **19** , 59-79

[22] Smith, Jillian, Blackwell, Louisa & Lynch, Kevin (2002), "The ONS Longitudinal Study, quality issues from 30 Years of data linkage", *Working Paper no. 11, Joint UNECE-EUROSTAT Work Session on Registers and Administrative Records for Social and Demographic Statistics, Geneva*

[23] Winkler, William E. (2001), "Record Linkage Software and Methods for Merging Administrative Lists", *Statistical Research Division, Methodology and Standards Directorate, U.S. Bureau of the Census*

Methodenberichte des Dienstes Statistische Methoden des BFS
Rapports de méthodes du Service de méthodes statistiques de l'OFS
Methodological reports of the Statistical Methods Unit of SFSO

Graf, M. (2004). Fusion de données. Etude de faisabilité. Numéro de commande: 338-0022

Potterat, J. (2003). Mietpreis-Strukturerhebung 2003. Entwicklung des Stichprobenplans und Ziehung der Stichprobe. Bestellnummer: 338-0021

Potterat, J. (2003). Landwirtschaftliche Betriebszählung 2003. Stichprobenplan der Zusatzerhebung. Bestellnummer: 338-0020

Renaud, A. (2003). Estimation de la couverture du recensement de la population de l'an 2000. Echantillon pour l'estimation de la sur-couverture (E-sample). Numéro de commande: 338-0019

Hulliger, B. (2003). Erhebung über Forschung und Entwicklung in der schweizerischen Privatwirtschaft 2000. Bereinigung der Stichprobe, Ersatz fehlender Werte und Schätzverfahren. Bestellnummer: 338-0018

Renfer, J.-P. (2003). Enquête 2000 sur la recherche et le développement dans l'économie privée en Suisse. Plan d'échantillonnage. Numéro de commande: 338-0017

Potterat, J. (2003). Kosten und Nutzen der Berufsbildung aus Sicht der Betriebe. Schätzverfahren. Bestellnummer: 338-0016

Graf, M., Matei, A. (2003). Stratégie de choix des modèles de désaisonnalisation. Application aux séries de l'emploi total. Numéro de commande : 338-0015

Potterat, J., Salamin, P.A. (2002). Betriebszählung 2001. Methoden für die Datenbereinigung. Bestellnummer: 338-0014

Renaud, A. (2002). Programme international pour le suivi des acquis des élèves (PISA). Plans d'échantillonnage pour PISA 2000 en Suisse. Numéro de commande: 338-0013

Renfer, J.-P. (2002). Enquête 2001 sur les coûts et l'utilité de la formation des apprentis du point de vue des établissements. Plan d'échantillonnage. Numéro de commande: 338-0012

Potterat, J., Salamin, P.A. (2002). Betriebszählung 2001. Stichprobenplan und Schätzverfahren für die provisorischen Ergebnisse. Bestellnummer: 338-0011

Graf, M. (2002). Enquête suisse sur la structure des salaires 2000. Plan d'échantillonnage, pondération et méthode d'estimation pour le secteur privé. Numéro de commande: 338-0010

Renaud, A., Eichenberger P. (2002). Estimation de la couverture du recensement de la population de l'an 2000. Procédure d'enquête et plan d'échantillonnage de l'enquête de couverture. Numéro de commande: 338-0009

Kilchmann, D., Hulliger, B. (2002). Stichprobenplan für die Obstbaumzählung 2001. Bestellnummer: 338-0008

Graf, M. (2002). Passage du concept établissement au concept entreprise. Numéro de commande: 338-0007

Salamin, P.A. (2001). La technique de la double enquête pour la statistique du transport routier de marchandise. Numéro de commande: 338-0006

Peters, R., Renfer, J.-P. et Hulliger, B. (2001). Statistique de la valeur ajoutée 1997-1998. Procédure d'extrapolation des données. Numéro de commande: 338-0005

Potterat, J., Hulliger, B. (2001). Schätzung der Sägereiproduktion mit der Sägerei-Erhebung PAUL. Bestellnummer: 338-0004

Graf, M. (2001). Désaisonnalisation. Aspects méthodologiques et application à la statistique de l'emploi. Numéro de commande: 338-0003

Hüsler, J., Müller, S. (2001). Schlussbericht Betriebszählung 1995 (BZ 95), Mehrfach imputierte Umsatzzahlen. Bestellnummer: 338-0002

Renaud, A. (2001). Statistique suisse des bénéficiaires de l'aide sociale. Plan d'échantillonnage des communes. Numéro de commande: 338-0001

Hulliger, B., Eichenberger, P. (2000). Stichprobenregister für Haushalterhebungen: Umstellung auf Telefonnummern ohne Namen und Adressen, Abläufe für Erstellung und Stichprobenziehung. Bestellnummer: 338-0000

de Rossi, F.-X. (1998). Méthodes statistiques pour le compte routier suisse.

Hulliger, B., Kassab, M. (1998). Evaluation of Estimation Methods for the Survey on Environment Protection Expenditures of Swiss Communes.

Salamin, P.A. (1998). Etablissement d'une clef de passage pondérée entre l'ancienne (NGAE 85) et la nouvelle nomenclature (NOGA 95) générale des activités économiques.

Peters, R. (1998). Extrapolation des données de l'enquête de structure sur les loyers.

Bender, A., Hulliger, B. (1997). Enquête suisse sur la population active: rapport de pondération pour 1996.

Salamin, P.A. (1997). Evaluation de la Statistique de l'emploi.

Peters, R. (1997). Etablissement du plan d'échantillonnage pour l'enquête 1996 sur la recherche et le développement dans l'économie privée en Suisse.

Peters, R. (1997). Enquête 1996 sur la structure des salaires en Suisse: établissement du plan d'échantillonnage.

Peters, R. (1996). Pondération des données de l'enquête sur la famille en Suisse.

Comment, T., Hulliger, B., Ries, A. (1996). Gewichtungsverfahren für die Schweizerische Arbeitskräfteerhebung (1991-1995).

Hulliger, B. (1996). Haushalterhebung Familie 1994: Stichprobenplan, Stichprobenziehung und Reservestichproben.

Peters, R., Hulliger, B. (1996). Schätzverfahren für die Lohnstruktur-Erhebung 1994 / Procédure d'estimation pour l'enquête de 1994 sur la structure des salaires.

Peters, R. (1996). Schéma de pondération des indices PAUL.

Hulliger, B., Peters, R. (1996). Enquête sur le comportement de la population suisse en matière de transport en 1994: plan d'échantillonnage et pondération.

Hulliger, B. (1996). Gütertransportstatistik 1993: Schätzverfahren mit Kompensation der Antwortausfälle.

Salamin, P.A. (1995). Estimation des flux pour le module II des comptes globaux du marché de travail.

Peters, R. (1995). Enquête de structure sur les loyers: établissement d'un plan d'échantillonnage stratifié.

Hulliger, B. (1995). Konjunkturelle Mietpreiserhebung: Stichprobenplan und Schätzverfahren.

Schwendener, P. (1995). Verbrauchserhebung 1990 - Vertrauensintervalle.

Peters, R., Hulliger, B. (1994). La technique de pondération des données: application à l'enquête suisse sur la santé.

Hulliger, B., Peters, R. (1994). Enquête sur la structure des salaires en Suisse: stratégie d'échantillonnage pour le secteur privé.

Publikationsprogramm BFS

Das Bundesamt für Statistik (BFS) hat – als zentrale Statistikstelle des Bundes – die Aufgabe, statistische Informationen breiten Benutzerkreisen zur Verfügung zu stellen.

Die Verbreitung der statistischen Information geschieht gegliedert nach Fachbereichen (vgl. Umschlagseite 2) und mit verschiedenen Mitteln

Programme des publications de l'OFS

En sa qualité de service central de statistique de la Confédération, l'Office fédéral de la statistique (OFS) a pour tâche de rendre les informations statistiques accessibles à un large public.

L'information statistique est diffusée par domaine (cf. verso de la première page de couverture); elle emprunte diverses voies:

Diffusionsmittel	*Kontakt* *N° à composer*	*Moyen de diffusion*
Individuelle Auskünfte	032 713 60 11 info@bfs.admin.ch	Service de renseignements individuels
Das BFS im Internet	www.statistik.admin.ch	L'OFS sur Internet
Medienmitteilungen zur raschen Information der Öffentlichkeit über die neusten Ergebnisse	www.news-stat.admin.ch	Communiqués de presse: information rapide concernant les résultats les plus récents
Publikationen zur vertieften Information (zum Teil auch als Diskette/CD-Rom)	032 713 60 60 order@bfs.admin.ch	Publications: information approfondie (certaines sont disponibles sur disquette/CD-Rom)
Online-Datenbank	032 713 60 86 www.statweb.admin.ch	Banque de données (accessible en ligne)

Nähere Angaben zu den verschiedenen Diffusionsmitteln liefert das laufend nachgeführte Publikationsverzeichnis im Internet unter der Adresse www.statistik.admin.ch>>News>>Neuerscheinungen.

La Liste des publications mise à jour régulièrement, donne davantage de détails sur les divers moyens de diffusion. Elle se trouve sur Internet à l'adresse www.statistique.admin.ch >>Actualités>>Nouvelles publications.

Methodenberichte des Dienstes Statistische Methoden
Rapports de méthodes du Service de méthodes statistiques
Methodological reports of the Statistical Methods Unit

Die Methodenberichte beschreiben die mathematischen und statistischen Methoden, die den Resultaten und Analysen der öffentlichen Statistik zu Grunde liegen. Sie enthalten ausserdem die Evaluation und Entwicklung von neuen Methoden im Hinblick auf eine zukünftige Anwendung. Diese Publikationen sollen einerseits die verwendeten Methoden dokumentieren, um Transparenz und Wissenschaftlichkeit sicher zu stellen, und sie sollen andererseits die Zusammenarbeit mit den Hochschulen und der Wissenschaft fördern.

Zur Illustration der beschriebenen mathematischen Konzepte, werden im Bericht numerische Resultate aufgeführt. Diese sind allerdings nicht als offizielle Resultate der betreffenden Erhebungen zu verstehen. Ebenfalls können die tatsächlich angewendeten Methoden leicht von den hier beschriebenen abweichen.

Die Methodenberichte sind auf der Internetseite des BFS in elektronischer Form verfügbar.

Les rapports de méthodes décrivent les méthodes mathématiques et statistiques à la base des résultats et des analyses de la statistique publique. Ils présentent également l'évaluation et le développement de nouvelles méthodes en vue d'une application future. Ces publications visent d'une part à documenter les méthodes utilisées ou envisagées dans un souci de transparence et de rigueur scientifique, et d'autre part à favoriser la collaboration avec le monde scientifique et universitaire.

Les résultats numériques présentés dans les rapports de méthodes illustrent les concepts mathématiques décrits, mais ne sont pas des résultats officiels des enquêtes concernées. De même, les méthodes réellement appliquées peuvent différer légèrement de celles décrites dans ces rapports.

Les rapports de méthodes sont disponibles sous forme électronique sur le site internet de l'OFS.

Il existe deux types de fusion de données: le couplage d'enregistrements et l'appariement statistique. On procède à un couplage lorsque la même unité statistique est recherchée dans deux sources de données différentes, et à un appariement lorsque la fusion des deux sources est basée sur la création d'unités synthétiques. Ces deux modes de fusion comportent quelques analogies techniques, mais sont des procédés de nature totalement différente, aboutissant à des fichiers de résultats qui n'ont pas les mêmes conditions d'application. Le rapport commence par présenter plus en détail ces deux modes de fusion, puis se termine par quelques recommandations. En Annexe, on trouve quelques applications tirées de la littérature internationale, ensuite les principales notions et difficultés liées à l'appariement statistique sont illustrées par des exemples.

N° de commande: 338-0022
Commandes: 032 713 60 60 Fax: 032 713 60 61 E-Mail: order@bfs.admin.ch
Prix: gratuit
ISBN 3-303-00275-4