

Swiss National Cohort

Appariement entre le recensement de la population 2000 et le recensement 2010/2011 fondé sur les registres

Rapport de synthèse sur les données et les méthodes utilisées

Concept

Les données sur les personnes du recensement de la population 2000 (RFP2000) ont été appariées à celles du relevé structurel 2010/2011 fondé sur les registres (R11) à l'aide de méthodes probabilistes.

Aucun identifiant unique, tel que le numéro AVS, ni aucun nom ou prénom n'était disponible pour l'appariement. Il a donc dû être effectué à l'aide des caractéristiques du RFP2000 à disposition. Les principales variables d'appariement (voir détails ci-dessous) étaient les suivantes:

- date de naissance;
- sexe;
- domicile et, le cas échéant, coordonnées du bâtiment ou EGID;
- nationalité;
- état civil;
- d'autres indications sur les partenaires et les enfants (date de naissance et sexe) ont par ailleurs été utilisées lorsqu'elles étaient disponibles.

De nombreuses caractéristiques pouvant changer en dix ans, telles que le domicile et l'état civil voire la nationalité, il était extrêmement difficile d'apparier les données avec celles du relevé fondé sur les registres. Ce dernier a été réalisé pour la première fois en 2010 pour remplacer le recensement fédéral de la population. Certaines données livrées par les communes ne répondaient alors pas encore aux exigences: c'était le cas par exemple des données sur les ménages, selon les indications de l'OFS. Nous avons donc décidé, en accord avec l'OFS, d'apparier les données du recensement de la population 2000 à celles du relevé structurel de 2011 (R11) plutôt que du relevé de 2010.

Les données de 2010 (R10) ont toutefois été intégrées dans l'appariement. Voir ci-dessous pour les détails.

Afin d'obtenir le meilleur appariement possible, nous avons exploité d'autres sources de données pour compléter les informations du RFP2000 et du R11. Le registre des mariages et des divorces, par exemple, fournit des indications sur les changements d'état civil. Il se peut donc que les données de quelqu'un qui se serait marié en 2006 et qui était célibataire jusque-là soient appariées à celles d'une personne célibataire dans le RFP2000. Il a donc

fallu modifier l'état civil de cette personne dans le R11 — de mariée à célibataire — pour l'appariement au RFP2000.

Fichiers de référence

Pour l'appariement, on a utilisé les données officielles du recensement de la population 2000: données sur les personnes (VZ00_p004), N= 7 452 975, y compris différents domiciles économiques. S'ils étaient différents, seul le domicile au sens du Code civil et un domicile de droit économique ont été utilisés pour l'appariement.

Pour ce qui est du relevé fondé sur les registres, nous avons utilisé pour l'appariement principalement les données de 2011 sur les personnes (r11_pe_orig) avec N=8 077 045. Compte tenu des données manquantes et de la mauvaise qualité de certaines données (selon les informations de l'OFS), nous avons décidé d'ajouter aux données du R11 celles du R10 et, un an plus tard, celles du R12, dans le cadre du prétraitement. Dans le R10, nous avons en outre saisi les personnes qui avaient émigré ou étaient décédées en 2011, mais qui en 2010 étaient importantes pour l'appariement au RFP2000.

Les informations qui faisaient défaut dans le R11 ont été reprises, pour l'appariement, du R10 ou, dans de rares cas, du R12. Les données du relevé fondé sur les registres ont été appariées à l'aide du numéro AVS pseudonymisé de l'OFS. Celles du R12 ont été utilisées principalement pour les contrôles de cohérence lorsque les valeurs du R10 et du R11 n'étaient pas identiques.

Appariements préparatoires

Les données du RFP2000 ont été complétées par des informations provenant d'autres sources pour l'appariement principal entre le RFP2000 et le relevé 2011 fondé sur les registres.

- a) Les données du registre des mariages et des divorces ont été appariées avec celles du R11: lorsque les données correspondaient, nous disposions ainsi d'indications sur l'état civil des personnes, et éventuellement d'autres informations valables au moment du recensement de la population 2000.
- b) Pour les données PETRA (statistique de la population étrangère), nous avons procédé à un appariement probabiliste à celles du RFP2000 et nous avons adapté et complété les variables d'appariement, lorsque cela était possible.
- c) Les données sur la mortalité du RFP2000 et du R11 ont été intégrées directement dans l'appariement, afin de pouvoir trouver les concordances entre les données sur les personnes tirées du RFP2000 et celles de la statistique des causes de décès ou du R11.

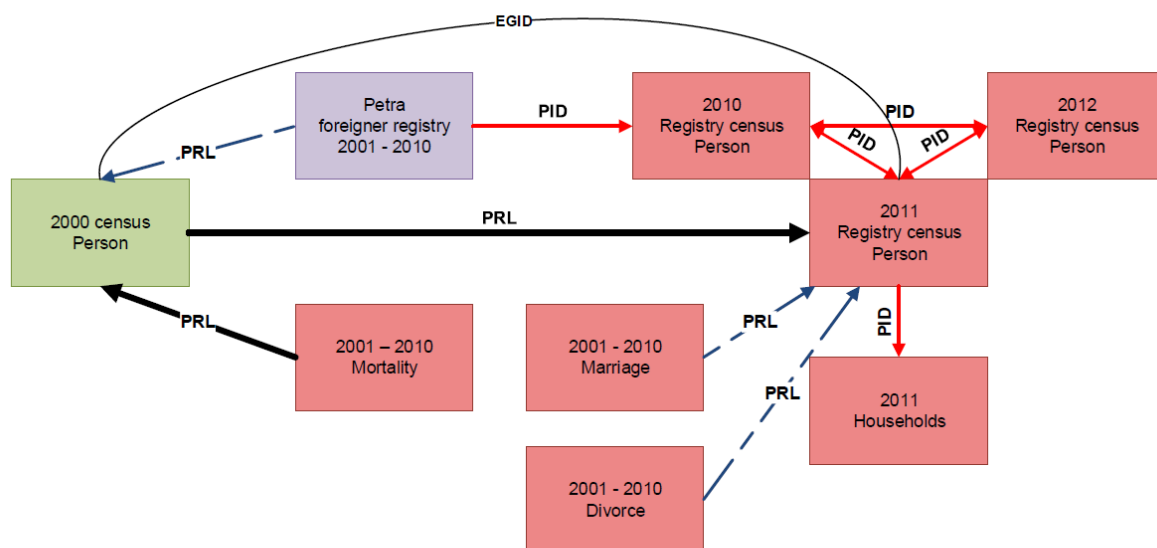


Figure 1: Linkage VZ2000 zu R11
PRL: Probabilistic Record Linkage
PID: Linkage über Pseudo-AHV

Variables d'appariement

Pour l'appariement, on a utilisé les variables suivantes selon le jeu de données:

- date de naissance;
- sexe;
- état civil;
- domicile, ou coordonnées du bâtiment d'habitation ou EGID, si disponible;
- domicile il y a cinq ans;
- date de l'arrivée dans la commune;
- date de l'émigration ou de l'immigration;
- type de domicile (par exemple, domicile au sens du Code civil);
- nationalité;
- autorisation de séjour;
- lieu de naissance, pays de naissance;
- Si disponibles: date de naissance et sexe des membres du ménage ou des partenaires. Les conjoints n'ont pas été recensés explicitement. Des règles ont par conséquent été établies pour couvrir les structures familiales les plus courantes (conjoints si mariés, autres partenaires, familles monoparentales, etc.);
- nationalité du ou de la partenaire;
- date du changement d'état civil (mariage, divorce, veuvage);
- religion (uniquement RFP2000 et données sur le décès).

Incohérences

Des problèmes se posent inévitablement dès lors qu'il existe plus d'une source de données pour plusieurs caractéristiques. Par exemple: les données des relevés R10 et R12 ont été ajoutées lors de la préparation des données R11; celles de quelques personnes ne coïncidaient pas pour certaines variables (la date de naissance par exemple). Il a donc fallu

définir pour chaque variable des règles pour déterminer quelle valeur considérer prioritairement comme correcte. Pour ce faire, nous nous sommes basés sur la source, en recourant par exemple aux données du relevé R10 si la date de naissance était identique dans les relevés R10 et R12, mais différente dans R11.

Lorsqu'il y avait deux valeurs, toutes deux potentiellement correctes, elles ont été prises en compte toutes les deux dans le jeu de données pour l'appariement. En d'autres termes, deux valeurs étaient disponibles pour la date de naissance (DOB1, date de naissance 1, et DOB2, date de naissance 2), lesquelles étaient toutes deux considérées comme potentiellement correctes et avaient le même poids dans le processus d'appariement.

Il y a également eu des incohérences au niveau de l'état civil ou de la nationalité. Là aussi, des règles ont été établies pour savoir quelle valeur serait finalement reprise dans le fichier d'appariement.

Appariement probabiliste des données

L'appariement probabiliste des données (PRL pour *probabilistic record linkage*) a été réalisé à l'aide de la version 3.0 du logiciel G-Link de Statistique Canada. G-Link est une application qui traite les données dans SAS sous la forme d'une base de données.

Dans le PRL, les éléments d'un jeu de données (A) sont comparés à ceux d'un deuxième jeu de données (B): chaque variable d'appariement se voit attribuer une probabilité M (match) et une probabilité U (unmatch) (refs). Les probabilités ont ensuite été converties en poids entiers (logarithmes). Un poids d'appariement (= probabilité) a ainsi été défini pour chaque appariement et pour chaque variable d'appariement. Chacun de ces poids a été additionné. Il en résulte un poids total d'appariement. Si la valeur est élevée, la probabilité est forte que les deux données concernent la même personne.

Le PRL permet de coupler les données des personnes, même si les données de certaines variables diffèrent légèrement (p. ex. date de naissance refusée) ou si des variables changent au fil du temps (état civil, nationalité).

Les étapes de travail ci-après ont été mises en œuvre pour le PRL :

Prétraitement

Préparation des données comprenant la standardisation des noms de variables, le recodage en cas de catégories différentes (p. ex. état civil: quatre catégories dans le RFP2000 et sept catégories dans le R11) et l'harmonisation des numéros de commune et des codes de nationalité.

En outre, des informations ont été récupérées du R10 ou du R12 lorsqu'elles manquaient dans le R11. En cas de données contradictoires, des règles ont été établies pour déterminer quelles valeurs reprendre dans les fichiers d'appariement.

Critères initiaux (procédure de blocage)

Une présélection a été effectuée à l'aide d'une série de critères parmi les appariements potentiellement possibles de deux jeux de données en vue de l'appariement A X B . Il n'aurait pas été possible, ni judicieux, de comparer tous les jeux de données du RFP2000 et du R11. Cela serait revenu à comparer environ $60,2 \times 10^{12}$ paires de données possibles. Pour l'appariement RFP2000-R11, on a donc défini une procédure de blocage, en reliant les

variables par des OU logiques: a) la date de naissance coïncide, b) le sexe, la nationalité et le lieu de domicile à la naissance coïncident, c) le sexe, la nationalité et l'année de naissance coïncident, d) la date de naissance du ou de la partenaire coïncide (si disponible). Il en a résulté $4,2 \times 10^9$ paires possibles, qui ont ensuite été comparées individuellement.

Règles

Un ensemble de règles a été défini pour chaque variable d'appariement:

- a) Concordance (A, agreement): concordance totale
- b) Concordance partielle (PA, partial agreement): concordance partielle, p. ex. date d'anniversaire ± 1 , ou règles concernant l'état civil (la combinaison célibataire selon le RFP2000 et marié selon le R11 était possible, l'inverse ne l'était pas, etc.) Des règles ont également été établies pour les caractéristiques spatiales (coordonnées, numéros de commune): pour les coordonnées, ± 100 m était par exemple mieux coté que ± 1000 m, ou des poids d'appariement étaient attribués pour les concordances partielles en cas de non-concordance des coordonnées, si la commune coïncidait.
- c) Règles pour l'utilisateur (U, user rules): certaines variables n'étant pas indépendantes, il a fallu développer des règles spécifiques et des poids d'appariement détaillés. Citons pour exemple la date de naissance ou l'âge, et l'état civil: en dessous de 25 ans, presque tout le monde est célibataire, au-delà de 65 ans, la proportion de personnes veuves augmente, etc. Les poids d'appariement ont été définis en fonction de la fréquence.
- d) Non-concordance (D, disagreement): si aucune concordance ni aucune concordance partielle n'a été trouvée, le poids a été utilisé pour une non-concordance (par exemple, des dates de naissance complètement différentes ou un sexe non concordant, bien que dans de très rares cas le sexe ait effectivement changé).

PRL

Déroulement de la procédure d'appariement: pour chaque paire de données possible (voir les critères initiaux), on a comparé chaque donnée du jeu de données du RFP200 avec celle du jeu de données du R11. Chaque variable d'appariement a fait l'objet d'un contrôle et le poids correspondant a été attribué à chacun des résultats (A, PA, U ou D).

La PRL est un processus itératif qui prévoit des tests après chaque série d'appariements. Lorsqu'une pondération trop forte a été attribuée à un appariement, on a adapté le poids et les règles et relancé l'appariement.

Classification

Chaque paire de données comparée dans le PRL se voit attribuer un poids d'appariement.

Les poids sont attribués selon les critères suivants (cut-offs):

- paires de données définitives (poids très élevé),
- paires de données possibles,
- paires de données rejetées.

Le groupe moyen a été contrôlé manuellement en détail, en évaluant la pondération des différentes variables d'appariement. En cas de problème, les poids et les règles ont été adaptés et le PRL relancé.

Il n'a pas été possible de vérifier chaque cas individuellement, compte tenu de la grande quantité de données, de sorte que les valeurs limites pour les groupes ont avant tout été définies sur la base des enseignements tirés d'autres projets d'appariement et de la connaissance des données.

Une valeur limite de 190 a été définie (final cut-off) après le contrôle manuel: au-dessus de ce seuil, les paires de données ont été définies comme appartenant à la même personne, en dessous comme appartenant à des personnes différentes.

Il en résulte inévitablement des appariements faussement positifs et d'autres faussement négatifs.

Post-traitement

Lors du post-traitement des résultats de l'appariement, nous avons vérifié combien de données sur les personnes du RFP2000 ont pu être reliées à des données du R11 et combien de fois il y a eu des concordances totales ou partielles.

Les non-concordances, soit les cas où les valeurs des paires de données ne correspondent pas, peuvent donner des indications sur la qualité des données et celle des appariements. Lorsque les résultats n'étaient pas plausibles, on a adapté à nouveau les règles et relancé le PRL.

Résultat de l'appariement des données du RFP2000 et du relevé 2011 fondé sur les registres

Le PRL a abouti aux résultats suivants, qui ont ensuite conduit à étendre la Swiss National Cohort à la version 2.0:

	%	N
Part de décès appariés avec le RFP2000	97,0%	655 485
Appariement du RFP2000 avec:		7 280 246
Relevé fondé sur les registres 2011	80,8%	5 881 740
Registres des décès 2001-2010 (2011)	9,0%	655 485
Personnes ayant émigré, lien vers PETRA	1,7%	125 493
Personne ayant émigré puis immigré à nouveau jusqu'en 2011	0,3%	18 207
Personnes sans appariement	8,2%	599 321

Adrian Spörri, le 13 août 2021