

Swiss National Cohort

Linkage Volkszählung 2000 zur Registerzählung 2010/11

Zusammenfassender Bericht zu den verwendeten Daten und Methoden

Konzept

Die Personendaten der Volkszählung 2000 (VZ2000) wurden mit probabilistischen Methoden zur Registererhebung 2010/11 (R11) verlinkt.

Für die Verlinkung stand kein eindeutiger Identifikator, wie die AHV-Nummer, und auch keine Namen und Vornamen zur Verfügung. Das Linkage musste mit den vorhandenen Merkmalen der VZ2000 durchgeführt werden. Die hauptsächlichen Linkvariablen (Details siehe unten) waren:

- Geburtsdatum
- Geschlecht
- Wohnort und wo vorhanden, Koordinaten des Gebäudes oder EGID
- Nationalität
- Zivilstand
- Zudem wurde, wo vorhanden, zusätzliche Angaben von Partnern und Kindern (Geburtsdatum und Geschlecht) verwendet

Da sich über 10 Jahre relativ viele Merkmale verändern können, wie Wohnort und Zivilstand, zum Teil auch Nationalität, war die Verlinkung zur Registererhebung 2010 enorm schwierig. Die Registererhebung 2010 war neu und ersetzte die bisherigen Volkszählungen. Im Jahr 2010 gab es noch einige Datenprobleme bei den Lieferungen der Gemeinden. Zum Beispiel waren im 2010 die Haushaltsdaten gemäss Auskunft des BFS noch ungenügend. Wir haben deshalb in Absprache mit dem BFS entschieden, die Volkszählung 2000 mit der Registererhebung 2011 (R11) zu verlinken, und nicht mit 2010.

Die Daten aus 2010 (R10) wurden aber in das Linkage integriert. Details, siehe unten.

Damit ein möglichst gutes Ergebnis des Linkage erzielt werden konnte, wurden die Daten der VZ2000 und der R11 mit vorhandenen anderen Datenbeständen ergänzt. Zum Beispiel kann das Heirats- und Scheidungsregister Hinweise auf Veränderungen des Zivilstandes geben: Wer zum Beispiel im Jahr 2006 geheiratet hat, und vorher ledig war, konnte unter Umständen als ledige Person zur VZ2000 gelinkt werden. So wurde der Zivilstand in R11 für das Linkage mit VZ2000 von verheiratet auf ledig geändert.

Referenz Dateien

Für das Linkage wurden die offiziellen Daten der Volkszählung 2000 benutzt: Personendaten (VZ00_p004), N= 7'452'975, inklusive verschiedene wirtschaftliche Wohnorte. Für das Linkage wurde dann nur der zivilrechtliche und ein wirtschaftsrechtlicher Wohnort benutzt, falls diese unterschiedlich waren.

Bei der Registererhebung wurde vor allem mit den Personendaten aus dem Jahr 2011 (r11_pe_orig) mit N=8'077'045 gelinkt. Da aber zum Teil die Daten Missings hatten und auch die Datenqualität nicht bei allen Variablen gut war (gemäss Auskunft BFS), beschlossen wir, zu den R11 Daten die Daten von R10 und ein Jahr danach von R12, beim Pre-Processing mitzuverarbeiten. In R10 wurden zudem mögliche Personen erfasst, welche im 2011 entweder ausgewandert oder gestorben waren, aber im 2010 für das Linkage mit der VZ2000 von Bedeutung waren.

Bei fehlenden Informationen in R11 wurde entweder R10, und in seltenen Fällen die Information aus R12 in das Linkagefile übernommen. Die Registererhebungsdaten wurden untereinander mit Hilfe der Pseudo-AHV-Nummer des BFS verlinkt. R12 wurde primär für Konsistenzprüfungen verwendet, falls die Werte in R10 und R11 nicht identisch waren.

Vorbereitende Linkages

Für das Hauptlinkage VZ2000 zur Registererhebung 2011 wurde die VZ2000 mit Informationen aus anderen Quellen ergänzt.

- a) Die Daten des Heirats- und Scheidungsregisters wurde mit R11 verlinkt: Falls die Personen gefunden wurden, gab dies Hinweise auf den Zivilstatus, allenfalls auch andere Infos zum Zeitpunkt der Volkszählung 2000.
- b) PETRA-Daten wurden probabilistisch mit der VZ2000 verlinkt und falls möglich, die Linkagevariablen angepasst und ergänzt.
- c) Die Sterbedaten wurden direkt ins Linkage VZ2000 zu R11 integriert. So konnten Personen aus der VZ2000 einerseits in der Todesursachenstatistik oder andererseits in R11 gefunden werden.

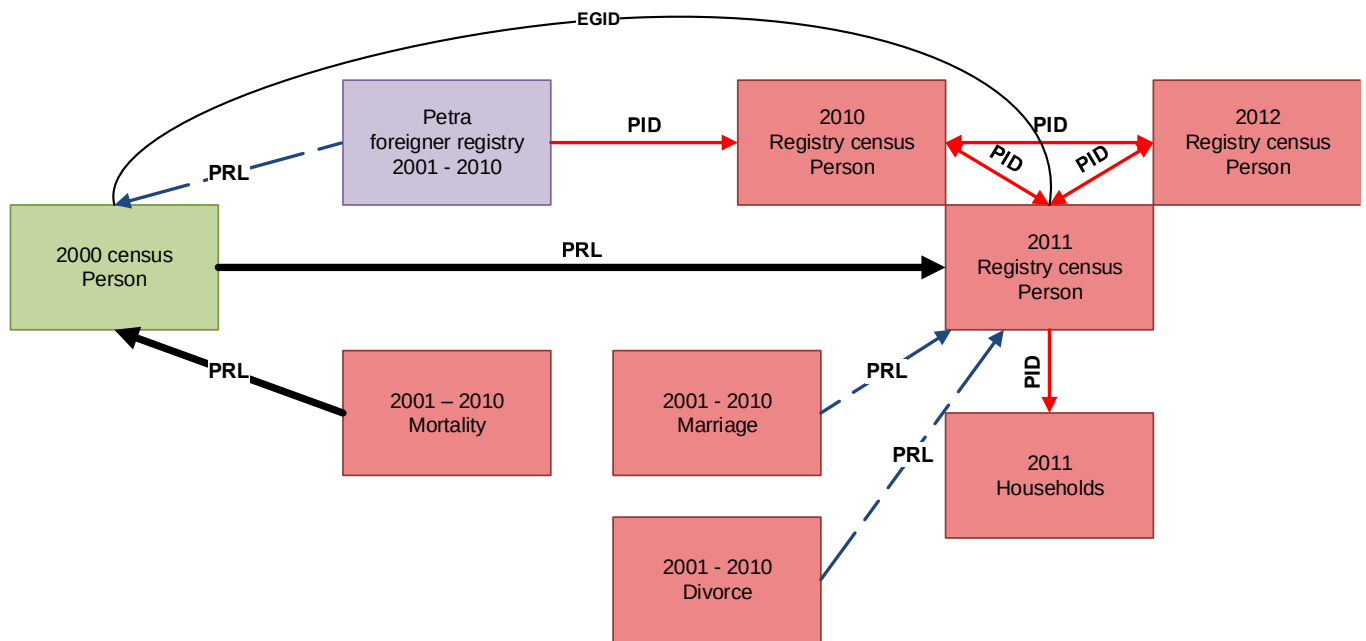


Figure 1: Linkage VZ2000 zu R11
PRL: Probabilistic Record Linkage
PID: Linkage über Pseudo-AHV

Linkagevariablen

Für das Linkage wurden folgende Variablen verwendet, je nach Datensatz:

- Geburtsdatum
- Geschlecht
- Zivilstand
- Wohnort, oder Koordinaten des Wohngebäudes oder EGID, wo vorhanden
- Wohnort vor 5 Jahren
- Zuzugsdatum Gemeinde
- Datum Emigration oder Immigration
- Typ des Wohnsitzes (z.B. zivilrechtlich)
- Nationalität
- Aufenthaltsbewilligung
- Geburtsort, Geburtsland
- Falls vorhanden: Geburtsdatum und Geschlecht der Haushaltsmitglieder, respektive Partner. Ehepartner wurden nicht explizit erfasst. Deshalb wurden Regeln erstellt, um die häufigsten Familienstrukturen (Ehepartner falls verheiratet, andere Partner, Alleinerziehende usw.) zu erfassen.
- Nationalität Partner
- Datum der Zivilstandsänderung (Heirat, Scheidung, Verwitwung)
- Religion (nur VZ2000 und Sterbedaten)

Widersprüche

Sobald mehr als eine Datenquelle für verschiedene Merkmale vorhanden ist, ergeben sich zwangsläufig Datenprobleme. Zum Beispiel: bei der Aufbereitung von R11 wurden R10 und R12 hinzugenommen. Bei ganz wenigen Personen gab es Diskrepanzen, zum Beispiel beim Geburtsdatum.

Dazu mussten für jede Variable Regeln definiert werden, welcher Wert prioritär als korrekt eingestuft wurde. Das hängt von der Quelle ab. Wenn z.B. das Geburtsdatum in R10 und R12 identisch, aber in R11 unterschiedlich war, wurde R10 genommen.

In Fällen, wo zwei Werte vorlagen, die beide möglicherweise korrekt sind, wurden beide in die Datensets für das Linkage mitgenommen. Das heisst, es gab DOB1 (Geburtsdatum 1) und DOB2 (Geburtsdatum 2) und im Linkageprozess galten beide als potentiell korrekt mit dem gleichen Gewicht.

Auch beim Zivilstand oder bei der Nationalität gab es Widersprüche. Auch dort wurden Regeln erstellt, welcher Wert schlussendlich ins Linkagefile übernommen wurde.

Probabilistic Record Linkage

Das probabilistische Record Linkage PRL wurde mit der Software G-Link von Statistics Canada mit der Version 3.0 durchgeführt. G-Link ist eine Applikation, welche Daten auf SAS als Datenbank verarbeitet.

Beim PRL werden Records aus dem einen Datensatz (A) mit den Records aus dem zweiten Datensatz (B) verglichen: jede Linkagevariable erhält eine M (match) und eine U (unmatch) Wahrscheinlichkeit (refs). Die Wahrscheinlichkeiten wurden danach in ganzzahlige Gewichte (Logarithmen) umgerechnet. Für jeden Vergleich und für jede Variable wurde so ein Linkagegewicht (=Wahrscheinlichkeit) definiert und über alle Linkagevariablen addiert. Das ergibt ein Total des Linkagegewichtes. Hohe Werte haben eine hohe Wahrscheinlichkeit, dass diese beiden Records zu der gleichen Person gehören.

Das PRL erlaubt Personen zu finden, auch wenn die Angaben in einzelnen Variablen leicht differieren (z.B. Geburtsdatum mit Typo), oder sich Variablen über die Zeit ändern (Zivilstand, Nationalität).

Für das PRL wurden folgende Arbeitsschritte durchgeführt:

Pre-processing

Vorbereiten der Daten, wie Standardisierung der Variablennamen, Rekodierung bei unterschiedlichen Kategorien (z.B. Zivilstand VZ2000 mit 4 Kategorien und R11 mit 7 Kategorien) oder Harmonisierung der Gemeindenummern und Nationalitätscodes. Zudem wurden bei R11 Information von R10, respektive R12 geholt, falls in R11 fehlende Werte waren. Bei widersprüchlichen Angaben wurden Regeln erstellt, welche Werte in die Linkagefiles übernommen wurden.

Initial criteria (blocking)

Von den potentiell möglichen Vergleichen von zwei Datensets für das Linkage A X B wurde eine Vorwahl mit Hilfe von Kriterien getroffen. Es war nicht möglich und auch nicht sinnvoll, alle Datensätze der VZ2000 mit allen Datensätzen von R11 zu vergleichen. Das wären ungefähr 60.2×10^{12} mögliche Datenpaare, die verglichen werden müssten.

Für das Linkage VZ2000-R11 wurde ein Set von Blocks definiert, welche über logische ODER verknüpft wurden: a) Übereinstimmung im Geburtsdatum, b) Übereinstimmung bei Geschlecht, Nationalität und Wohnort bei Geburt, c) Übereinstimmung bei Geschlecht, Nationalität bei Geburt und Geburtsjahr, d) Übereinstimmung im Geburtsdatum des Partners (wo vorhanden).

Das führte zu 4.2×10^9 möglichen Paaren, welche danach einzeln verglichen wurden.

Rules

Für jede Linkagevariable wurde ein Set von Regeln definiert:

- a) Agreement (A): vollständige Übereinstimmung
- b) Partial agreement (PA): teilweise Übereinstimmung, z.B. Geburtstag +-1, oder Regeln beim Zivilstand (ledig in VZ2000 und verheiratet in R11 war möglich, umgekehrt nicht etc.). Auch bei den räumlichen Merkmalen (Koordinaten, Gemeindenummern) wurden Regeln erstellt: bei den Koordinaten wurde z.B. +-100m besser gewertet, als +-1000m, oder bei Nichtübereinstimmung der Koordinaten, aber Übereinstimmung der Gemeinde Linkagegewichte für PA vergeben.
- c) User rules (U): Da gewisse Variablen nicht unabhängig waren, wurden spezielle Regeln entwickelt und detaillierte Linkagegewichte definiert. Als Beispiel dient hier das Geburtsdatum, respektive das Alter, und der Zivilstand: unter 25 sind fast alle ledig, über 65 steigt der Anteil der verwitweten Personen usw. Die Gewichte des Linkages wurden entsprechend der Häufigkeit definiert.
- d) Disagreement (D): falls keine Übereinstimmung und auch keine partielle Übereinstimmung gefunden wurde, wurde das Gewicht für eine Nichtübereinstimmung benutzt (z.B. völlig unterschiedliche Geburtsdaten, oder nichtübereinstimmendes Geschlecht, obwohl sich in ganz seltenen Fällen das Geschlecht tatsächlich geändert hat).

PRL

Durchführung des Linkage: Für jedes Paar von möglichen Records (siehe initial criteria) aus dem Datensatz VZ200 wurde jedes Record aus dem Datensatz R11 verglichen. Dazu wurde jede Linkagevariable überprüft und die entsprechenden Ergebnisse (A, PA, U oder D) mit den entsprechenden Linkagegewichten vergeben.

Das PRL verläuft iterativ, mit Tests nach jeder Linkagerunde. Falls gewisse Vergleiche von Variablen zu stark gewichtet wurden, wurden die Gewichte oder Regeln angepasst und das Linkage erneut laufen gelassen.

Klassifikation

Jedes Datenpaar, welches im PRL verglichen wurde, erhält ein Linkagegewicht. Die Gewichte wurden anhand von Regeln (cut-offs) eingeteilt in:

- definitive Datenpaare (mit sehr hohem Gewicht)
- mögliche Datenpaare
- verworfene Datenpaare

Manuell wurde die mittlere Gruppe im Detail angeschaut und die Gewichtung der einzelnen Linkagevariablen beurteilt. Falls Problem auftraten wurden die Gewichte und Regeln angepasst und das PRL wiederholt.

Da bei dieser grossen Menge von Daten die Einzelfälle nicht überprüft werden konnten, hängen die Grenzwerte für die Gruppen vor allem von der Erfahrung mit Linkageprojekten und der Kenntnisse der Daten ab.

Nach der manuellen Überprüfung wurde ein harter Grenzwert von 190 als finaler cut-off definiert: oberhalb wurden die Datenpaare als zur gleichen Person gehörend definiert, unterhalb als zu unterschiedlichen Personen gehörend.
Das führt zwangsläufig zu falsch positiven wie auch zu falsch negativen Verlinkungen.

Post-processing

Bei der Nachbearbeitung der Linkageresultate wurde überprüft, wie viele Personen aus der VZ2000 mit einem Record aus R11 verbunden werden konnte, wie häufig es vollständige oder partielle Übereinstimmungen gab.

Diskrepanzen, wenn Werte bei den Datenpaaren nicht übereinstimmen, können Hinweise auf die Datenqualität und die Linkagequalität geben. Bei nicht nachvollziehbaren Resultaten wurden die Regeln erneut angepasst und das PRL erneut gestartet.

Resultat des Linkages VZ2000 – Registererhebung 2011

Das PRL hat zu folgenden Resultaten geführt, welches danach in die Erweiterung der Swiss National Cohort zur Version 2.0 geführt hat:

	%	<i>N</i>
Anteil der Sterbefälle, welche zu VZ2000 gelinkt wurden	97.0%	655'485
VZ2000 gelinkt zu:		7'280'246
Registererhebung 2011	80.8%	5'881'740
Sterberegister 2001-2010(11)	9.0%	655'485
Ausgewandert, Link zu PETRA	1.7%	125'493
Ausgewandert und wieder eingewandert bis 2011	0.3%	18'207
Nicht gelinkte Personen	8.2%	599'321

Adrian Spörri, 13. August 2021