

BIJLAGE A. Reconstructie van socio-demografische levenslopen: nominale gegevenskoppeling

"Nominal record linkage" of nominale gegevenskoppeling is het volgens bepaalde regels samenbrengen van uiteenlopende gegevens over naamdragende eenheden in een samenhangend geheel (vgl. Wrigley 1973: 1). Die naamdragende eenheden zouden ook families of bedrijven kunnen zijn, maar hier gaat het om personen en het samenhangend geheel waarvan sprake is de geregistreerde levensloop.

Het heeft maar zin om aan nominale gegevenskoppeling te beginnen als aan twee voorwaarden voldaan is: het moet mogelijk zijn en nuttig. Het is mogelijk als er voldoende onderscheid kan worden gemaakt tussen correcte en verkeerde koppelingen. Het is nuttig als de koppeling het onderzoek informatie oplevert die er anders niet zou zijn. Het probleem is dat beide voorwaarden elkaar tegenwerken: hoe meer informatie observatie-eenheden bevatten, hoe beter het onderscheid tussen juiste en foute koppelingen te maken valt maar de informatiewinst die een nieuwe (juiste) koppeling in die situatie oplevert, neemt over het algemeen af naarmate de oorspronkelijke observaties rijker aan informatie waren (Wrigley 1973: 5-6). Veronderstel, om dit duidelijk te maken, dat er twee huwelijken geregistreerd zijn voor Jan Peters: één in 1849 en één in 1852. Als niet meer informatie voorhanden is dan deze naam en de huwelijksdata, valt nauwelijks uit te maken of de bruidegom in beide gevallen dezelfde persoon is. Wanneer voor beide observaties echter ook de geboortedatum en -plaats voorhanden is, de namen van de ouders en de namen van eventuele vorige vrouwen, dan wordt een correcte koppeling waarschijnlijker. Naarmate meer informatie voorhanden is, gaat die waarschijnlijkheid steeds dichterbij zekerheid grenzen. In het grensgeval zou de koppeling echter nutteloos worden omdat elke observatie-eenheid op zich al alle interessante informatie bevat. Veronderstel dat de tweede huwelijksakte alle informatie van de eerste correct overneemt, dan is een juiste koppeling met aan zekerheid grenzende waarschijnlijkheid mogelijk maar de informatiewinst is nihil. In het ideale geval bevat elke observatie-eenheid dus één uniek identificatie-item en verder informatie die niet bij andere eenheden aanwezig is: het onderscheid tussen juiste en foute koppelingen is dan gemakkelijk te maken en de koppelingen lijden niet onder een dalende marginale informatiewinst. In de praktijk identificeert

zelden één item een individu met voldoende zekerheid, zodat overlapping van informatie in meerdere informatievelden wenselijk is (Wrigley 1973: 5-6).

Met andere woorden moet bij koppeling een evenwicht worden gezocht tussen zekerheid omtrent de juistheid of accuraatheid van koppelingen en de winst aan informatie of het nut dat koppelingen opleveren. Bouchard (1992) verfijnde en operationaliseerde Wrigleys probleemstelling. De informatie in de gebruikte bronnen kan op drie manieren gebruikt worden: ten eerste als identificatie-item in het koppelingsproces, ten tweede als controle-informatie bij de validering van de koppelingen en ten derde als variabele bij de analyse. Het gevaar bestaat dat eenzelfde informatie op een ongeoorloofde manier in de drie fasen van het onderzoek gebruikt wordt. Om de accuraatheid van de koppelingen te maximaliseren, is het verleidelijk om zoveel mogelijk informatie als identificatie-items in de koppelingsvergelijkingen te betrekken. Het risico bestaat dan dat er geen variabelen meer overblijven voor een fatsoenlijke analyse. Als identificatie-items later ook als te analyseren variabelen gaan fungeren, loert het risico op vertekening immers altijd om de hoek¹ (Bouchard 1992: 69).

Om dit te verduidelijken een voorbeeld: veronderstel dat we informatie omtrent gemiddeld geboorte-interval gebruiken om kinderen al dan niet aan een bepaalde vrouw toe te schrijven (waarbij we *niet* koppelen bij een ongewoon lang of kort interval). De geboorte-intervals in de gekoppelde dataset zijn dan vertekend in de richting van een gemiddeld voortplantings-tempo. Bouchard (1992: 69) geeft een vergelijkbaar voorbeeld met betrekking tot nuptialiteit. De keuze en constructie van identificatie-items mag zich met andere woorden niet blind staren op het optimaliseren van de koppelingen op zich maar moet ook oog hebben voor de gevolgen voor de analysemogelijkheden.

Welke extra informatie of analysemogelijkheden levert nominale gegevenskoppeling op en op welke manier gebeurt die best?

A.1 Waarom koppelen?

Zoals in gezinsreconstructiestudies zijn de belangrijkste koppelingen voor onderhavige studie die tussen ouders en kinderen enerzijds en die tussen de ouders onderling anderzijds. Daartoe moeten in de eerste plaats observaties uit de bevolkingsregisters met elkaar worden verbonden. Binnen de bevolkingsregisters moet gezocht worden naar alle partners en alle kinderen

¹ 'Vertekening' is de Nederlandse vertaling van de Engelse term 'bias'.

van de drie referentiegeneraties. Dergelijke koppelingen van gegevens uit de bevolkingsregisters heten in wat volgt *interne koppelingen*. Maar omdat die registers systematisch de doodgeboren kinderen weglieten en omdat levend geboren maar jong-gestorven kinderen vaak ook niet werden opgenomen, werd bijkomend gezocht naar kinderen in de sterfte- en geboorteakten. Koppelingen tussen bevolkingsregistergegevens en de informatie uit die akten heten *externe koppelingen*.

De genoemde koppelingen tussen huwelijkspartners en tussen ouders en kinderen zijn uiteraard van essentieel belang om het reproductiepatroon van de steekproefgeneraties te reconstrueren. In principe kan men nadat deze koppelingen gemaakt zijn de verzamelde fragmenten uit de levensloop van de drie generaties op zich analyseren: elke inschrijving in het bevolkingsregister vertegenwoordigt een levensloopfragment. En elke inschrijving bevat informatie over het moment van inschrijving en een einddatum voor het fragment. Verder vermeldt het de leeftijd, het beroep, de burgerlijke staat en eventuele veranderingen daarin in de loop van het fragment. Met behulp van event history analysis is het mogelijk deze afgeknotte fragmenten apart te analyseren, ook al horen verschillende fragmenten mogelijk bij eenzelfde persoon.

Er zijn echter voldoende redenen om verschillende inschrijvingen van eenzelfde persoon in de bevolkingsregisters met elkaar te koppelen. Koppeling levert extra informatie op voor een meer diepgaande analyse, beperkt het aantal getrunceerde observaties en maakt een controle van de accuraatheid van de ruwe gegevens mogelijk. Om dit te verduidelijken is het zinvol de interne koppelingen verder op te delen naar intraregisterkoppelingen enerzijds en interregisterkoppelingen anderzijds.

Een *intraregisterkoppeling* is de verbinding van verschillende, samenhangende observaties binnen één bevolkingsregister, bijvoorbeeld het register van 1846. Een *interregisterkoppeling* is de verbinding van verschillende, samenhangende observaties uit verschillende, opeenvolgende registers, bijvoorbeeld de koppeling van gegevens uit de registers van 1846 met gegevens uit die van 1856. Concreet gaat het in beide gevallen om de verbinding van observaties die horen bij één en dezelfde persoon, met de bedoeling de geregistreerde levensloop te reconstrueren. Externe koppeling beoogt van verschillende bevolkingsregisters als het ware één doorlopend register te maken.

Er zijn twee belangrijke redenen waarom observaties voor een persoon op verschillende plaatsen in een bevolkingsregister verspreid kunnen staan. Ten eerste kan iemand herhaaldelijk in het register genoteerd zijn omwille van migratie. Iemand die Leuven verliet maar terugkeerde vóór het sluiten van het register, dus vóór de volgende volkstelling, werd opnieuw ingeschreven. De folio waar de gegevens werden genoteerd is afhankelijk van het huis waar de migrant gaat wonen. Herhaaldelijke emigratie en immigratie zorgde dus telkens voor een nieuwe inschrijving in het register, al dan niet op verschillende folio's.

Een tweede reden voor herhaalde inschrijving is verandering van burgerlijke staat. Na een huwelijk werden de echtgenoten plus de eventueel al aanwezige kinderen in de meeste registers overgeschreven naar de folio van het huis waar zij gingen wonen. Soms was dat in het huis waar één van de huwelijkspartners al woonde. Dit was vaak het geval bij hertrouw na verweduwing. In de registers van 1846 en 1856 werden huwenden echter haast nooit overgeschreven. Wel werden in die registers buitenechtelijk geboren kinderen die bij een huwelijk geëgitimeerd werden, overgeschreven naar het folio van de vader. Erkende en legitieme kinderen werden sowieso al op die folio ingeschreven.

Verhuis binnen de gemeente gaf meestal geen aanleiding tot het overschrijven van de gegevens voor de betreffende mensen, al gebeurde dat soms wel. Waarom is niet geheel duidelijk, maar meestal lijkt het een gevolg te zijn geweest van de behoefte om orde te scheppen op een met de jaren rommelig geworden folio.

Inter- en intraregisterkoppelingen beperken het aantal getrunceerde observaties. Dat geldt zowel voor linker- als voor rechtertruncatie. Een geval van linkertruncatie is bijvoorbeeld een gehuwde waarvan de huwelijksdatum onbekend is terwijl het wenselijk is de huwelijksduur te kennen. De truncatie verdwijnt als deze observatie gekoppeld kan worden aan een observatie waarin de huwelijksdatum wel vermeld is. Een voorbeeld van rechtertruncatie zou een koppel zijn dat een volgend kind krijgt na het afsluiten van een bevolkingsregister. Als dat koppel in het volgende bevolkingsregister geïdentificeerd kan worden, vervalt de rechtertruncatie.

Met niet-getrunceerde gevallen is een meer diepgaande analyse van het reproductieproces mogelijk. Zo is het enkel mogelijk de huwelijksduur in een regressievergelijking op te nemen als de huwelijksdatum bekend is. Het totale aantal geboren kinderen, levende en overledene, kan enkel in de analyse worden opgenomen als door inter- en intraregisterkoppelingen de volledige reproductieve levensloop gekend is.

Ten slotte maken dergelijke koppelingen het mogelijk de consistentie en de accuraatheid van de ruwe gegevens te controleren. Naarmate identificatiegegevens voor correct gekoppelde individuen conflicteren (bijvoorbeeld één keer geboortedatum 1/12/1830 en een volgende keer 1/11/1830), zijn de gegevens waarschijnlijk minder betrouwbaar. De registratie van geboortedata was wellicht minder betrouwbaar naarmate de registers vaker conflicterende data opgaven voor personen die herhaaldelijk opnieuw geregistreerd werden.

Interregisterkoppeling laat bovendien toe de accuraatheid van de registratie in te schatten, met name de registratie van migratie. Als iemand bij het afsluiten van een register geregistreerd staat als inwoner van Leuven, moet die bij het openen van het volgende register naar aanleiding van de volkstelling inderdaad in Leuven blijken te wonen, zo niet is er wellicht sprake van een ongeregistreerde emigratie (of overledene, maar de kans dat een overlijden door de mazen van het ambtelijke net glipte was veel kleiner, zie Gutmann & van de Walle 1978). En andersom: iemand die naar aanleiding van de volkstelling als inwoner van Leuven geregistreerd wordt, moet in principe al in het voorgaand bevolkingsregister geregistreerd staan. Als dat niet het geval is, zag de bevolkingsadministratie wellicht een immigratie (of geboorte) over het hoofd.

De externe koppeling tussen geboorteakten en bevolkingsregisters maakte ten slotte ook een externe validering mogelijk van de informatie die cruciaal is voor dit project: het aantal kinderen van de referentiegeneraties en de timing van de voortplanting.

A.2 Identificatie-items

Een identificatie-item is een stukje informatie over een bepaalde persoon die in beide eventueel te koppelen observatie-eenheden aanwezig is (Winchester 1973a: 19-21). De accuraatheid van een koppelingsprocedure hangt voor een groot stuk af van de kwaliteit en de beschikbaarheid van de identificatievariabelen. Deze paragraaf bespreekt deze problematiek. Vervolgens wordt de gevolgde koppelingsprocedure beschreven.

In het ideale geval voldoen identificatie-items volgens Winchester (1973a: 38) aan drie voorwaarden: ze zijn aanwezig voor alle observatie-eenheden, uniek voor elk individu en bij hun herhaling of kopiëring worden weinig fouten gemaakt. Bouchard (1992: 69) voegt daar aan toe dat ze best geen rechtstreeks te analyseren variabelen zijn.

A.2.1. Beschikbaarheid van identificatie-items

De meeste literatuur omtrent nominale gegevenskoppeling voor historische studies betreft gezinsreconstructiestudies op basis van parochieregisters (doop-, huwelijks- en begrafenisakten) of op basis van akten van de burgerlijke stand. Het aantal beschikbare identificatie-items is in dergelijke bronnen zeer beperkt en verschilt zeer sterk van land tot land, soms van streek tot streek en van periode tot periode. In die context heeft de naam nog de meeste kenmerken van de ideale identificatievariabele (zie Winchester 1973a; Wrigley & Schofield 1973; Herlihy 1973; Blayo 1973; Skolnick e.a. 1977; Schwartz e.a. 1984; Wesley e.a. 1987). In vele omstandigheden zijn alleen de namen van betrokken personen beschikbaar. In het beste geval is dat een reeks namen, bijvoorbeeld de namen zowel van de boreling als van de ouders in een geboorteakte of de namen van alle ouders van een huwelijkspaar in een huwelijksakte. In het slechtste geval vermeldt de akte slechts één naam (zie Wrigley 1973: 5-12; Wrigley & Schofield 1973: 64-67). Het geslacht moet meestal uit de voornaam worden afgeleid. Verder is de leeftijd van de betrokkene(n) op het moment van de gebeurtenis in kwestie nog het enige informatie-item dat verder vrij ruim beschikbaar is. Soms bevatten akten ook informatie over woonplaats, beroep en/of getuigen. Skolnick e.a. (1977) geven een overzicht voor een elftal projecten in zeven landen.

Het aantal en de beschikbaarheid van potentiële identificatie-items in de Belgische bevolkingsregisters is uitzonderlijk groot. Elke observatie, elke regel in het bevolkingsregister bevat naast familie- en voornaam, in principe ook de leeftijd of de geboortedatum, de geboorteplaats en het beroep. Daarenboven worden meestal ook familieverbanden aangeduid, vooral deze tussen ouders en kinderen. Over dit laatste volgt elders meer uitleg. Tabel A.1 geeft aan in welke mate de overige identificatie-items daadwerkelijk beschikbaar zijn voor gebruik bij de gegevenskoppeling. (In het vervolg staat de afkorting BR voor 'Bevolkingsregister'). Er is zo goed als altijd een familienaam en minstens één voornaam beschikbaar. De volledige geboortedatum is echter lang niet altijd beschikbaar.

Bij de voorbereidingen van de volkstelling van 1846 waren de provinciale statistische commissies van Antwerpen, Brabant en Luik het er over eens dat het geen zin had om de mensen naar hun geboortedatum te vragen. Heel veel mensen, zelfs uit de hogere sociale klassen, kennen hun geboortedatum niet en het is belangrijk voor de kwaliteit van de volkstelling enkel vragen te stellen waar iedereen gemakkelijk op kan antwoorden, aldus het commissieverslag (Commission Centrale de Statistique 1845: 65). Daarom vroeg ook de Leuvense

volksteller op 15 oktober 1846 niet naar exacte geboortedata maar gewoon naar de leeftijd. Maar zelfs dan blijft het de vraag hoe accuraat daarop geantwoord werd. In de historische demografie werden vaak afrondingseffecten vastgesteld in de leeftijdsvariabele: mensen gaven blijkbaar frequenter een rond getal als leeftijd op, 5 of 40 jaar bijvoorbeeld in plaats van 6 of 39. Dergelijk afrondingseffect heeft zich bij de volkstelling van 1846 in Leuven waarschijnlijk niet voorgedaan. Dat mag worden afgeleid uit de frequentieverdeling van de opgegeven leeftijden: er is duidelijk geen sprake van systematisch hogere frequentie in de ronde leeftijdscategorieën. De gemiddelde frequentie in de door tien deelbare leeftijden ligt net iets lager dan in de andere categorie, respectievelijk 31,3 en 31,9. In de door vijf deelbare leeftijden ligt de gemiddelde frequentie net iets hoger dan in de niet door vijf deelbare leeftijden, respectievelijk 30,5 en 32,2. We nemen aan dat dit kleine verschil aan het toeval te wijten is. Het is immers erg onwaarschijnlijk dat er meer naar vijftallen dan naar tientallen werd afgerond.

Tabel A.1: Beschikbaarheid van identificatie-items in de Leuvense bevolkingsregisters (BR)

<i>Identificatie-item</i>	<i>BR1846</i>	<i>BR1856</i>	<i>BR1866</i>	<i>BR1880</i>	<i>BR1890</i>	<i>BR1900</i>
Familienaam	100,00%	100,00%	100,00%	99,99%	100,00%	100,00%
Voornaam	99,96%	99,95%	99,16%	99,98%	99,99%	99,98%
Geboortedatum						
enkel leeftijd	46,78%	11,17%	0,01%	0,00%	0,00%	0,00%
enkel geboortejaar	1,17%	5,03%	5,73%	7,78%	1,10%	0,04%
geboortejaar en -maand	0,18%	1,74%	0,73%	0,42%	0,27%	0,04%
volledige geboortedatum	51,64%	82,01%	93,36%	91,76%	98,62%	99,91%
Geboorteplaats	99,31%	99,80%	99,76%	99,80%	99,82%	99,78%
Beroep	97,80%	99,03%	93,31%	82,13%	91,77%	48,58%
waarvan 'zonder beroep'	44,27%	47,36%	51,96%	40,87%	47,82%	2,16%
N	6817	4430	17587	17417	12413	4547

Snel moet het de verantwoordelijken duidelijk zijn geworden dat leeftijd moeilijk te integreren valt in een dynamisch, voortdurend geactualiseerd systeem als de bevolkingsregisters. In de eerste plaats was het absurd voor kinderen geboren na de referentievolkstelling telkens leeftijd nul te noteren en daarom werd dat dus systematisch de geboortedatum (Oris 1990: 152). Verder waren er nog vele gevallen waar de datum van inschrijving in het bevolkingsregister ambigu was. Zo werden mensen die bij de volkstelling vergeten waren later in het register ingeschreven, bijvoorbeeld als hun verblijf in Leuven naar aanleiding van een huwelijk of een geboorte aan het licht gekomen was. Het is niet altijd duidelijk op welke datum de

vermelde leeftijd dan betrekking heeft zodat hun geboortjaar zelfs niet bij benadering te berekenen valt. Daarom registreerden sommige bevolkingsadministraties, waaronder de Leuvense, op eigen initiatief vaak wel de geboortedatum of -jaar. Wanneer er sprake was van een huwelijk was dat zelfs duidelijk systematisch het geval. Van 93% van de individuen die huwden in de looptijd van BR1846 is de geboortedatum gekend. Voor de niet-trouwers is dat slechts 46%.²

Dat toch nog bijna 52% van de observaties in het register van 1846 een geboortedatum vermelden, is dus vooral te danken aan het feit dat de Leuvense ambtenaren op eigen initiatief meestal de geboortedatum van borelingen en huwenden noteerden. Wellicht namen ze die over uit de geboorteakte. Daarenboven is nog van ruim 1% het geboortjaar bekend. Voor 47% van de observaties kunnen we een benaderend geboortjaar berekenen door de leeftijd af te trekken van het jaar waarin de persoon in het register werd ingeschreven. Slechts voor 0.1% van de persoonsgebonden observaties uit de registers van 1846 ontbreekt elke aanduiding van leeftijd of geboortjaar.

De registers van 1856 vermeldde al bij 82% van de inschrijvingen de volledige geboortedatum en slechts voor 0.05% ontbreekt elke leeftijds aanduiding. Vanaf de registers van 1866 was de registratie van de geboortedatum een wettelijke verplichting en de Leuvense praktijk voldeed hieraan sindsdien in 91,76 tot 99,91% van de gevallen. Het aantal observaties waarvoor geen enkele informatie omtrent geboortjaar voorhanden is, blijft altijd beperkt tot een handvol (zie Tabel A.1).

De geboorteplaats is het identificatiekenmerk dat doorheen de gehele reeks bevolkingsregisters het meest beschikbaar is, maar de uniciteit en daarmee het onderscheidingsvermogen van dit kenmerk is zeer beperkt. Beroep is geen tijdsinvariant kenmerk: mensen kunnen in hun leven van beroep veranderen. Bovendien is de gebruikte omschrijving niet altijd even nauwkeurig. Toch kan het kenmerk soms gebruikt worden om te beslissen in twijfelgevallen (zie Wrigley & Schofield 1973: 75). Sommige vormen van beroepsmobiliteit waren, zeker in de 19^{de} eeuw, zeer onwaarschijnlijk. Een dagloner die later als advocaat gaat werken, bijvoorbeeld, of een brouwersknecht die schooldirecteur wordt, zijn zeer onwaarschijnlijke scenario's en vormen bij twijfel een argument *contra* koppeling.

²De nulhypothese van onafhankelijkheid werd m.b.v. een chi-kwadraattoets verworpen (df=1, p<0.0001).

A.2.2. Unicité van identificatie-items

Naast de beschikbaarheid bepaalt ook de uniciteit de bruikbaarheid van identificatie-items voor nominale gegevenskoppeling. Hoe unieker een bepaalde waarde van een identificatie-item, hoe hoger zijn vermogen om in de gehele verzameling die observaties te onderscheiden die bij de persoon in kwestie horen. Het onderscheidingsvermogen van de geboorteplaats Leuven is bijvoorbeeld zeer laag omdat vele Leuvense mensen in Leuven geboren zijn. Sommige auteurs gebruiken de term informatiewaarde als synoniem voor het onderscheidingsvermogen.

De meest strategische koppelingsprocedure gaat uit van de identificatievariabelen die gemiddeld een grote informatiewaarde hebben. Chiaramella ontwikkelde een maat voor die gemiddelde waarde, de entropie H genoemd en gedefinieerd als volgt (zie Wesley e.a. 1987: 194-195):

$$H = - \sum_{k=1}^K p_k \log_2 (p_k) \quad (\text{A.1}) \quad (1)$$

waarin p_k staat voor de relatieve frequentie van waarde k van de variabele in kwestie. H is nul als de spreiding van de identificatievariabele nul is ($p_k=K=1$) maar heeft geen opperste limiet. De entropie is niet afhankelijk van steekproefomvang maar wel van het aantal waarden K dat de variabele aanneemt: hoe hoger K , hoe hoger *c.p.* de entropie H . Verder is H gevoelig aan de spreiding: hoe sterker de concentratie van observaties in een beperkt aantal categorieën, hoe lager het onderscheidingsvermogen van die variabele. Tabel A.2 geeft de entropie van een aantal potentiële identificatievariabelen naar bevolkingsregisterjaar. Deze waarden liggen voor overeenkomstige identificatie-items bijzonder dicht bij de waarden die Oris (1990: 153) vond voor het Brabantse stadje Huy en de waarden voor de Mormonen-dataset van Utah (Wesley e.a. 1987: 194).

Tabel A.2: Entropie (H) van de identificatievariabelen naar bevolkingsregister

Identificatievariabele	BR1846	BR1856	BR1866	BR1880	BR1890	BR1900
Familienaam	10,07	9,77	11,01	10,92	10,69	9,85
Gestandaardiseerde familienaam (*)	10,00	9,71	10,90	10,80	10,57	9,78
Soundex van gestand. Familienaam (*)	8,82	8,66	9,24	9,12	8,98	8,71
Voornaam	8,44	9,00	9,84	10,21	10,45	10,02
Geboortedatum	6,89	10,05	12,64	12,45	12,80	11,69
(Berekend) geboortejaar	5,96	5,64	6,05	5,96	5,91	5,74
Geboorteplaats	3,62	3,52	4,17	4,10	3,90	3,91
N	6817	4430	17587	17417	12413	4547

In het bevolkingsregister van 1846 heeft de familienaam het grootste onderscheidingsvermogen. Ook de gestandaardiseerde versies daarvan, waarover hieronder meer, hebben een grotere informatiewaarde dan de overige items. Die van de geboortedatum wordt in dat eerste register nog sterk uitgehold door het grote aantal observaties waarvoor de volledige datum niet beschikbaar is. Vanaf de registers van 1856 heeft de geboortedatum echter een groter onderscheidingsvermogen dan de familienaam; de kans dat twee observaties een zelfde familienaam hebben, is groter dan de kans dat ze eenzelfde geboortedatum hebben.

A.2.3. Variaties en verschillen

Wanneer een set unieke identificatie-items altijd beschikbaar is, valt in theorie gemakkelijk te beslissen of twee observaties geldig gekoppeld kunnen worden of niet: ja, als de identificatie-items dezelfde waarden hebben, nee, als de waarden verschillen. In de praktijk treden bij het herhalen van een identificatie-item meestal fouten en variaties op, dat wil zeggen afwijkende vormen van dezelfde 'ware' waarde. In het koppelingsproces komen sommige radicale fouten in bepaalde omstandigheden aan het licht maar wellicht vaker leiden fouten onherroepelijk tot gemiste koppelingen. Deze paragraaf behandelt enkel de manier waarop de koppelingsprocedure met variaties omgaat.

Het onderscheiden van variaties van werkelijk verschillende waarden is zo essentieel voor het koppelingsproces dat de meeste literatuur over nominale gegevenskoppeling in historisch onderzoek dáárover gaat. Vooral de variaties in schrijfwijzen van namen vormt een moeilijk-

heid.³ Verder zijn in het koppelingsproces procedures gebruikt om met variaties in geboorteplaats en -datum om te gaan.

A.2.3.1 Familienaam

De naamvariaties in een computerbestand weerspiegelen zowel de variaties in de oorspronkelijke bron als de variaties geïntroduceerd tijdens het invoeren van de gegevens in de computer.⁴ De ambtenaren schreven een in wezen zelfde naam soms lichtjes anders, maakten schrijffouten, vertaalden van het Nederlands naar het Frans of kortten af. De codeurs lazten een *n* in plaats van een *s*, typten een *L* in plaats van een *M* of vervormden een voor hen onbekende, moeilijk leesbare naam tot een daarop lijkende, vertrouwde variant. De lijst van variaties is praktisch eindeloos en is voor een groot stuk afhankelijk van de taal, de gebruikte bron, lokale gewoonten en administratieve nauwkeurigheid (Desama 1991; Winchester 1992: 152-156; Morton 1994). In de Leuvense bevolkingsregisters leken bijvoorbeeld de hoofdletters C en T sterk tot zeer sterk op elkaar, afhankelijk van de gemeentelijke klerk van dienst. Daardoor komt de familie Cornu in de bestanden vaak voor onder de naam Tornu en de familie Tosseyn onder de naam Cosseyn.

Bouchard en Pouyez (1980) ordenen familienaamvariaties in drie categorieën.

1. Spellingvariaties of orthografische variaties verschillen van schrijfwijze maar niet van uitspraak. De meerderheid van de familienaamvariaties zijn van dit type. *De Keyser* heette bijvoorbeeld later *Dekeyzer* en nog later *De Keizer*. De fonetische structuur van de naam blijft doorheen deze orthografische variaties constant. Daarom bestaat de gebruikelijke procedure om er mee om te gaan uit het omzetten van de familienaam in een fonetische standaardcode.
2. Bij fonetische varianten blijft de variatie niet beperkt tot de spelling: ook de uitspraak verandert. Soms is dat tamelijk drastisch, zoals bij vertaling van *Van Parijs* naar *Du Paris* of wanneer de codeur *VanArenberg* leest als *VanDenberg*. Soms is het uitspraakverschil kleiner, zoals bij *Berrewaarts* en *Berrenaarts* of bij *Janssens* en *Janssen*. Fonetische varianten zijn meestal moeilijker te herkennen dan orthografische. Eenvoudige standaard-

³ Zie bijvoorbeeld het dozijn bijdragen in de *Annales de démographie historique* van 1972, Blayo (1973), de proceedings van een *Conference on Methods of Automatic Family Reconstitution* (Skolnick e.a. 1977), Bouchard & Pouyez (1980), Schwartz e.a. (1984), Reher (1984), Nygaard (1992), Winchester (1992) of Bloothoof (1998).

⁴ De codeurs kregen de opdracht namen letterlijk in te typen zoals genoteerd in het register, zij het dat alles automatisch in hoofdletters werd omgezet.

disatie biedt vaak geen oplossing. Dit type varianten komt bij familienamen meestal op toevalsbasis tot stand maar is schering en inslag bij voornamen.

3. Familienamen die uit twee of meer delen bestaan, werden soms slechts gedeeltelijk genoteerd. Zo treft men de familie *De Dieudonné de Corbeek Overloo* soms aan als *Dedieu-donné*. Omdat die naam nooit afgekort wordt tot *de Corbeek Overloo* is die variatie nog relatief gemakkelijk op te sporen. Dat gaat echter moeilijk als de familie *Germus dit Wouters* nu eens als *Germus* geregistreerd is en dan weer als *Wouters* maar nooit met beide namen. De Franse uitdrukking *dit* komt in de Leuvense registers herhaaldelijk voor en de vrije vertaling luidt "in de volksmond bekend als". Meestal bestaat het eerste deel uit een Franse naam en het tweede deel uit een Nederlandse. Samengestelde familienamen met toponiemen zijn de regel voor adellijke families: *Dejonghe De Schietere de Lophem* of *De Bernard De Fauconvalle*, bijvoorbeeld.

Elke koppelingsstrategie moet een methode voorzien om met dergelijke variaties om te springen. Vele variaties hebben een min of meer systematisch karakter. Sommige zijn typisch voor de gebruikte bron en de lokale omstandigheden. Andere variaties doen zich op louter toevallige wijze voor en zijn bijvoorbeeld een gevolg van nooit volmaakte codeurbetrouwbaarheid. Elk project heeft zo zijn eigen procedures ontwikkeld om met de principieel eindeloze reeks mogelijke variaties om te gaan, alle pogingen tot ontwikkeling van een algemeen systeem ten spijt. De meest complexe systemen ter automatische detectie van varianten en verschillen werden in het verleden ontwikkeld in het kader van het GENISYS-project aan de universiteit van Utah in Salt Lake City (Wesley e.a. 1987), door de SOREP-groep in Quebec (Bouchard 1992) en door de Cambridge Group (Schofield 1992). Deze teams investeerden zeer veel in de automatisatie van een zeer complex koppelingsproces omdat zij met zeer grote datasets werkten. Maar geen van de systemen is buiten de oorspronkelijke onderzoekscontext toegepast wegens te weinig algemeen en te weinig flexibel (Desama 1991: 119). Daarom geldt voor de meeste koppelingsprojecten nog steeds wat Wrigley en Schofield (1973: 100-101) jaren geleden al schreven: in geval van twijfel is menselijke beoordeling nodig om uit te maken of het om spellingsvarianten gaat dan wel om twee verschillende namen (vgl.

Schofield 1992: 76; Bouchard 1992: 68).⁵ Dat neemt niet weg dat de overgrote meerderheid automatisch herkend kan worden.

In de literatuur circuleren twee basisstrategieën om met naamvariaties om te gaan: standaardisatie en similariteitsmeting. Grotere koppelingsprojecten combineren beide strategieën. Standaardisatie zet elke geregistreerde naam om in een standaardvorm. Dat kan vooraf gebeuren of op het moment waarop twee observaties vergeleken worden, en de standaard kan via een stel regels in een algoritme voor elke observatie bepaald worden, dan wel door opzoeking in een soort elektronisch woordenboek. Similariteitsmeting probeert, altijd middels een algoritme, te becijferen hoe sterk of zwak twee voorliggende namen op elkaar lijken.

Virtueel alle projecten passen op de ruwe namen minstens een lichte vorm van standaardisatie toe. Nygaard (1992) onderscheidt naar vorm en toepassing vijf standaardisatietypes.

1. Fonetische codes proberen orthografische variaties te normaliseren door de uitspraak van een naam op een gestandaardiseerde manier voor te stellen. Zo ontwikkelende een team onderzoekers aan de universiteit van Quebec het *FONEM*-algoritme om spellingvariaties van Franse namen fonetisch te standaardiseren (Bouchard & Pouyez 1980). Het team van de universiteit van Utah ontwikkelde de *Utah Phonetic Traducer*, die het best aan het Engels is aangepast (Wesley e.a. 1987). De standaardisatievoorschriften zijn in deze algoritmes in een beperkt aantal regels gegoten die op elke geregistreerde naam worden toegepast. De meeste fonetische standaardisatiefuncties zijn echter *semi*-fonetisch: door de grote diversiteit aan orthografische variaties en uitzonderingen is het bijzonder moeilijk gebleken een sluitend systeem te ontwikkelen om alle geschreven namen in een precieze uitspraakcode om te zetten (Bloothoof 1994a: 29). Daarom, en omdat ze uitspraakvariaties niet standaardiseren, vormt fonetische codering in de meeste projecten slechts een eerste, oppervlakkig niveau van standaardisatie, gecombineerd met andere standaardisatiefuncties.
2. Etymologische codes pogen geschreven namen om te zetten in een basisvorm of stam waarvan de varianten afgeleiden zijn. Dergelijke codes kunnen in principe naast orthografische ook fonetische variaties herkennen. De standaardisatiefunctie wordt hier beschreven door een aantal afleidingsregels, eventueel gecombineerd met een elektronisch

⁵ In de toekomst mogen wellicht meer algemeen bruikbare procedures verwacht worden, wanneer een meer doorgedreven gebruik kan worden gemaakt van de onderzoeksresultaten rond kunstmatige intelligentie (Desama 1991; Bloothoof 1998: 46-48).

woordenboek. Deze vorm van standaardisatie is echter slechts in beperkte mate relevant voor familienamen. Onderzoekers die etymologische standaardisatie toepassen doen dat in regel enkel voor voornamen. Opnieuw is het virtueel onmogelijk een sluitend systeem van afleidingsregels te formuleren dat in meerdere contexten toepasbaar is (Bloothoof 1994b, 1998).

3. Gecomprimeerde vormen doen aan overstandaardisatie: ze kennen niet alleen dezelfde code toe aan naamvarianten maar ook aan vele verschillende namen. De naam wordt hier samengedrukt tot een skelet van een handvol karakters die enkel de contouren van de naam afbakenen. Het verlies aan informatie, inherent aan elke vorm van standaardisatie, is dan uiteraard zeer groot. Toch passen vele koppelingsprojecten compressie toe om het aantal te vergelijken observaties te beperken. Als het praktisch onmogelijk is alle potentiële koppelingen observaties in detail met elkaar te vergelijken, is het gebruikelijk om eerst sorteervakken te maken, in de Engelstalige literatuur 'sorting pockets' genoemd. Enkel observaties uit hetzelfde vak, dus met een zelfde gecomprimeerde familienaamvorm, worden dan meer in detail met elkaar vergeleken. Overstandaardisatie is dan wenselijk om er voor te zorgen dat alle denkbare varianten in hetzelfde sorteervak zitten (Bouchard & Pouyez 1980: 121; Nygaard 1992: 64; Winchester 1992: 154-155). De Russell Soundex Code en de door Louis Henry aan het Frans aangepaste variant zijn de bekendste compressiefuncties. Ze werden ontworpen als fonetische codes maar worden tegenwoordig niet meer als zodanig beschouwd omwille van hun hyperstandaardisatie (Bouchard & Pouyez 1980: 121; Nygaard 1992: 64)
4. Gebruikmakend van de expertise aanwezig in de naamkunde, is het volgens Nygaard (1992: 64) mogelijk numerieke codes aan namen toe te kennen op een manier die toelaat de verwantschap tussen twee namen te bepalen door eenvoudig optellen en aftrekken. Zo'n standaardisatiemethode is bij mijn weten in historische koppelingsprojecten nog niet toegepast.
5. Sommige onderzoekers zetten varianten om naar de verondersteld 'ware' of meest gebruikte vorm als een eerste vorm van standaardisatie. Tijdens de koppelingsprocedure van het SOREP in Quebec, bijvoorbeeld, wordt automatisch een steeds uitbreidend woordenboek aangelegd dat voor elke variant 1 de meest frequente variant 2 vermeld. Die modale naamvariant wordt dan in het vervolg van de procedure steeds gebruikt (Bouchard & Pouyez 1980: 123-124;).

Om fonetische varianten automatisch te herkennen zijn de meeste standaardisatie-algoritmes niet waterdicht genoeg. Daarom werken alle automatische koppelingsprocedures ook met similariteitsindices. Charbonneau c.s. (1972) deden pionierswerk op dit vlak, maar hun similariteitsmaat voldoet de meeste onderzoekers niet (bijvoorbeeld Bouchard & Pouyez 1980: 122). De meeste toegepaste indices zijn varianten van die van Guth (1976). Het Guth-algoritme meet de verwantschap tussen twee namen op basis van de relatieve posities van de letters. Het bekijkt letter na letter en is niet taalspecifiek omdat het geen rekening houdt met de fonetische aspecten van de letters. Het algoritme levert geen standaardvorm voor elke naam op, maar een getal dat moet aangeven in welke mate twee naamvormen op elkaar lijken. De onderzoeker bepaalt zelf de drempelwaarde om te beslissen of de observaties al dan niet verder met elkaar vergeleken moeten worden (Nygaard 1992: 64).

De te volgen strategie om naamvarianties te onderscheiden van naamverschillen, hangt af van de algemenere koppelingsstrategie. Volautomatische koppeling veronderstelt een mix van verschillende standaardisatiefuncties en similariteitsmeting. Als het aantal observaties hoog oploopt, is het bijvoorbeeld praktisch niet mogelijk om voor alle mogelijke paren een similariteitsindex te bepalen omdat dat enorm veel computertijd zou vergen. Het is dan raadzaam om eerst via hyperstandaardisatie sorteervakken te maken en enkel namen uit een zelfde vak in detail met elkaar te vergelijken (Winchester 1992: 154-155).

Om redenen uiteengezet in een volgende paragraaf is in dit project gekozen voor semi-automatische of *computer-assisted* koppeling: observaties met identieke identificatie-items of met eenvoudig te herkennen varianten worden automatisch gekoppeld, maar over de moeilijke gevallen beslist de onderzoeker.

De gevolgde procedures gebruiken een courante compressiefunctie in combinatie met semi-fonetische standaardisatie en similariteitsmeting. Deze worden hiërarchisch toegepast (vgl. Wesley e.a. 1987). Op het eerste niveau haalt een semi-fonetisch standaardisatie-algoritme een aantal systematisch en vaak voorkomende orthografische varianten uit de familienamen. Het gaat om een lichte vorm van standaardisatie met maximaal behoud van onderscheidingsvermogen. Op een tweede niveau comprimeert de Russell Soundex Code deze semi-fonetisch gestandaardiseerde namen. De gecomprimeerde waarden definiëren brede sorteervakken,

waarbinnen de semi-fonetisch gestandaardiseerde namen als één van de te vergelijken items fungeren. Die vergelijking gebeurt onder meer door het bepalen van similariteitsindices.

De semi-fonetische standaardisatieregels werden afgeleid uit een vooronderzoek op basis van de familienamen in de bevolkingsregisters van 1846. Dat leverde 23 eenvoudige regels op.⁶

Tabel A.3 geeft naast de regels ook telkens een voorbeeld. Namen die volgens deze regels werden gestandaardiseerd, worden in het vervolg aangeduid als **S(naam)**.

Tabel A.3: Primaire standaardisatieregels voor familienamen

STANDAARDISATIЕРЕГЕЛ	VOORBEELD
1. Verwijder alle spaties	DE BONT → DEBONT
2. Verwijder letterherhaling op het einde van een naam	VANDEKAA → VANDEKA; DEBONTT → DEBONT
3. Vervang QU door K	RASQUIN → RASKIN
4. Vervang CK door K	VANDENHOECK → VANDENHOEK
5. Vervang KK door K	HIKKENS → HIKENS
6. Vervang KX door KS	MERCKX → MERKX → MERKS
7. Vervang BB door B	ABBELOOS → ABELOOS
8. Vervang DT door T	DEBONDT → DEBONT
9. Vervang TH door T	MATHIJS → MATIJS
10. Vervang SCH door S	SCHREVENS → SREVENS DEBUSSCHERE → DEBUSSERE
11. Vervang CA door KA	DEFOUCAULT → DEFOUKAULT
12. Vervang CO door KO	DECORTE → DEKORTE
13. Vervang CU door KU	LOCUS → LOKUS
14. Vervang CE door SE	CERKEL → SERKEL
15. Vervang CI door SI	MERCIE → MERSIE
16. Vervang CR door KR	CRAB → KRAB
17. Vervang CHR door KR	CHRISTIAANS → KRISTIAANS
18. Vervang IJ door Y	SAVERIJS → SAVERYS
19. Vervang OY door OI	GODEFROY → GODEFROI
20. Vervang AE door AA	KRISTIAENS → KRISTIAANS
21. Vervang GH door G	BOOGHMANS → BOOGMANS
22. Vervang OEY door OEI	VANROEY → VANROEI
23. Vervang EY door Y	EYSKENS → YSKENS

De Russell Soundex Code is de bekendste compressiecode. De functie levert een code met vier posities op die het basisskelet van de naam voorstelt. De eerste positie is gewoon de eerste letter van het karakterveld dat gestandaardiseerd wordt. De overige letters van dat veld

⁶ In Luik formuleerde men 21 regels op dit minimaal niveau van standaardisatie (Oris 1990: 151 of Desama 1991) en in Utah 75 regels (Wesley e.a. 1987: 193). De Mormonen-databank in Salt Lake City, Utah, bevat dan ook niet alleen veel meer observaties maar bovendien namen uit een veelvoud aan talen.

worden gereduceerd tot drie getallen voor de medeklinkers met uitzondering van de letters W en H. De volledige functiebeschrijving luidt als volgt:

1. behoud de eerste letter;
2. laat van de rest van de naam de volgende letters vallen: A, E, H, I, O, U, W, Y;
3. ken de volgende getallen toe aan de overblijvende letters tot je drie getallen hebt:
 - B, F, P, V : 1
 - C, G, J, K, Q, S, X, Z : 2
 - D, T : 3
 - L : 4
 - M, N : 5
 - R : 6
4. als twee of meer opeenvolgende letters *in de oorspronkelijke naam* (vooraleer enige codering is doorgevoerd) dezelfde code hebben, laat ze dan allemaal weg behalve de eerste;
5. voeg indien nodig nullen aan de code toe indien je minder dan drie cijfers bekomt (Skolnick e.a. 1977: 16-17; Stephenson 1980).

De Russell Soundex van de semi-fonetisch gestandaardiseerde namen worden in het vervolg aangeduid als **RS(naam)**. Deze codering werd in historische nominale gegevenskoppeling zelden zonder meer toegepast. Zowel in experimenten als in de praktijk bleek het vermogen van de Soundex om variaties van eenzelfde naam te onderscheiden van verschillende namen immers te klein, vooral omdat verschillende namen te vaak een zelfde code krijgen. Voor het maken van sorteervakken is zo'n overstandaardisatie echter juist wenselijk (Wrigley & Schofield 1973: 98-99; Bouchard & Pouyez 1980: 121; Nygaard 1992: 64). Doms en Dams kunnen bijvoorbeeld verschillende familienamen zijn, maar even goed is het mogelijk dat de codeur ten onrechte een A voor een O heeft gelezen en ingevoerd. De Soundex van beide namen is D520. Problematischer is dat de soundex soms ook varianten een verschillende code geeft. Dat is systematisch het geval als de varianten met een verschillende letter beginnen. Tijdens de invoer van de Leuvense gegevens is gebleken dat de sierlijke eerste hoofdletters in de Leuvense registers regelmatig verkeerd gelezen werden. *Tamps* is zo een variant van *Camps*, ontstaan door een leesfout. De codes verschillen echter: T512 respectievelijk C512. Idem voor *Nilis* en *Milis*. Dergelijke eerste-letterfouten werden regelmatig gemaakt bij het lezen en invoeren van familienamen. Er zijn nog andere gevallen waar de Soundex een

verschillende code toekent aan potentiële varianten. Enkele voorbeelden uit de Leuvense praktijk: *Berrenaarts* (B656) en *Berrewaarts* (B663); *Bruglants* (B624) en *Bruylants* (B645); *Vandenberg* (V535) en *Vanarenberg* (V565). Wanneer de Soundex gebruikt wordt als sleutel voor het maken van sorteervakken, komen deze potentiële varianten niet in aanmerking voor verdere vergelijking. Daarom is in tweede instantie beslist om ook sorteervakken te maken op basis van intervallen rond het geboortejaar.

Standaardisatie leidt altijd tot een zekere mate van informatieverlies. Het aantal waarden of categorieën wordt immers gereduceerd. De bedoeling is echter zoveel mogelijk categorieën samen te voegen die inderdaad dezelfde familienaam representeren, terwijl zo weinig mogelijk werkelijk verschillende namen onder dezelfde noemer mogen worden gebracht. Een standaardisatie is dus efficiënter naarmate het verlies aan onderscheidingsvermogen kleiner is dan de reductie in het aantal categorieën (vgl. Wesley e.a. 1987: 193-196). Een mogelijke efficiëntiemaat E voor een standaardisatie S is dan de verhouding tussen de proportionele reductie in aantal categorieën en het proportioneel verlies aan onderscheidingsvermogen:

$$E_S = 1 - \frac{\frac{H - H_S}{K - K_S}}{K} \quad (\text{A.2})$$

waarin H de entropie is voor het identificatie-item vóór standaardisatie en H_S de entropie ná standaardisatie; K het aantal waarden van het item vóór en K_S het aantal waarden na standaardisatie. Het theoretische maximum voor E_S is 1, wanneer de standaardisatie geen verlies aan informatie zou betekenen. In de praktijk bereikt geen enkele standaardisatiefunctie dat theoretisch maximum. E_S wordt negatief als het proportioneel verlies aan onderscheidingsvermogen H groter is dan de reductie in het aantal waarden K : het is dan beter de standaardisatie niet door te voeren. Tabel A.4 geeft de efficiëntiematen voor beide toegepaste niveaus van standaardisatie.

Tabel A.4: Efficiëntie van de primaire en secundaire (soundex) standaardisatie van familienaam

Aantal waarden (K)	BR1846	BR1856	BR1866	BR1880	BR1890	BR1900
Familienaam	2643	1455	4312	4101	3235	1525
Semi-fonetische gestand. familienaam	2512	1403	4060	3863	3032	1465
Soundex (gestand. familienaam)	1088	773	1469	1407	1231	801
$E_{\text{SEMIFONETISCHE CODE}}$	0,8271	0,8281	0,8258	0,8125	0,8116	0,8099
E_{SOUNDEX}	0,7601	0,7601	0,7609	0,7549	0,7474	0,7594

Hoe hoog zou E zijn als we 26 sorteervakken zouden maken voor de letters van het alfabet? Uit een experiment blijkt dat het comprimeren van de namen tot de eerste letter een E van 0,40 zou halen, toegepast op het register van 1846. De semi-fonetische standaardisatie bereikt een dubbel zo hoog efficiëntieniveau: E bedraagt 0,80 tot 0,83 naargelang van het register. Het verschil met het ideale 1 ligt vooral aan de beperkte reductie in aantal categorieën en nauwelijks aan het verlies van informatie (zie Tabel A.2). Ter vergelijking: het fonetische UTP-algoritme van het team van de universiteit van Utah behaalt een E tussen 0,82 tot 0,92 naargelang van het niveau in de standaardisatiehiërarchie (berekend op basis van Wesley e.a. 1987: 194).

De soundex presteert minder goed (E is ongeveer 0.76 voor elk register) omdat het verlies aan informatie zeer groot is. De E van de Soundex ligt hoger voor de gegevens van Wesley e.a. (1987: 194) ($E=0,82$). Wellicht heeft dit te maken met het feit dat de Soundex het best aan de Engelse taal is aangepast (Winchester 1992: 154).

Voorals omdat het verlies aan onderscheidingsvermogen zeer beperkt is, mogen we concluderen dat de semi-fonetische code voldoet als identificatie-item. Wanneer de namen van twee observaties dezelfde code krijgen, gaan we er in het vervolg van uit dat deze namen dezelfde zijn. Het omgekeerde is duidelijk *niet* waar: varianten kunnen volgens de semi-fonetische standaardisatie nog steeds als verschillend worden beschouwd. Bij niet-identiteit van de gestandaardiseerde familienaam wordt daarom een similariteitsindex bepaald.

In navolging van Bouchard en Pouyez (1980) werd een index gebruikt die de proportionele benadering van Charbonneau c.s. (1972) combineert met de restrictieve benadering van het Guth-algoritme. Het eerste team definieerde een similariteitsindex als volgt:

$$i = \frac{I}{I + D} \quad (\text{A.3})$$

waarbij I het aantal identieke letters in de twee namen is en D het aantal verschillende letters in de twee namen. Het Guth-algoritme telt eveneens het aantal overeenkomstige letters maar

houdt rekening met hun relatieve positie. Dit is ook het geval in de gebruikte similariteitsindex g . Die heeft de zelfde vorm als i maar I en D worden bepaald als volgt:

I is het aantal letters van de kortste naam dat *op dezelfde relatieve positie* voorkomt in de langste naam;

D is het aantal letters van de langste naam dat *niet op dezelfde relatieve positie* voorkomt in de kortste naam.

Een reeks letters heeft per definitie dezelfde relatieve positie in twee namen als en slechts als aan de volgende twee voorwaarden voldaan is:

1. voor elk paar uit de reeks geldt dat als de ene letter voor/na de andere komt in de eerste naam, die letter ook voor/na de andere komt in de tweede naam;
2. als de afstand tussen om het even welke twee letters uit de reeks d bedraagt in de ene naam, dan mag de afstand tussen die twee letters in de andere naam maximaal $d+3$ bedragen.

De sequentie ER heeft omwille van voorwaarde 1 niet dezelfde relatieve positie in de namen $VERHULST$ en $HULSBERGEN$: in de eerste naam komen de twee letters voor de sequentie $HULS$, in de tweede naam er na. De letter R in de namen $BORMANS$ en $BOUCHER$ heeft volgens voorwaarde 2 niet dezelfde relatieve positie: in de eerste naam bedraagt de afstand tussen de O en de R één positie en in de tweede naam vijf posities. Het afstandsverschil is vier en dus te groot volgens voorwaarde 2.

I kan niet eenvoudig bepaald worden door de namen van voor naar achter, letter na letter, te doorlopen. Dat zou in een aantal gevallen tot een onderschatting van I leiden. Volgend voorbeeld toont dit aan.

PETERS

PATERS

Het is duidelijk dat de vijf onderlijnde letters dezelfde relatieve positie hebben. Het letter na letter overlopen van deze naam leidt echter tot de verkeerde conclusie dat I maar vier bedraagt. Een letter-na-letter-algoritme telt immers als volgt:

1. Eerste gemeenschappelijke letter is een P ;
2. de volgende gemeenschappelijke letter is de E , in de eerste naam onmiddellijk volgend op de P , in de tweede naam op de T ;
3. de T situeert zich *niet* op dezelfde relatieve positie want in de eerste naam komt hij ná en in de tweede naam vóór de E uit stap 2.

4. *R* en de *S* hebben wél dezelfde relatieve positie in beide namen.

Om dergelijk verkeerd telwerk te voorkomen werd een algoritme ontwikkeld dat zo lang mogelijke gemeenschappelijke sequenties van letters zoekt en hun relatieve positie in beide namen vergelijkt.

Enkele voorbeelden verduidelijken de berekening van *g*.

Naam 1: JOKMANS

Naam 2: JOCHMANS

$$g = 6 / (6 + 2) = 0.75$$

Naam 1: VANDENBERG

Naam 2: VANARENBERG

$$g = 9 / (9 + 2) = 0.82$$

Naam 1: HERMANS

Naam 2: WAGEMANS

$$g = 5 / (5 + 3) = 0.62$$

Naam 1: KOPPERS

Naam 2: KUYPERS

$$g = 5 / (5 + 2) = 0.71$$

Naam 1: DJEAN

Naam 2: DEJAAN

$$g = 5 / (5 + 1) = 0.83$$

Vaak is het onderscheid tussen variaties en verschillen ondubbelzinnig te maken (Peeters is duidelijk geen variant van Jansens) maar vaak is er ook onzekerheid. Zo valt op basis van de geïsoleerde naam niet te zeggen of Jan Peters van een andere familie is dan Jan Pieters of dat de ambtenaar gewoon een *i* te veel of te weinig geschreven heeft. Om in zulke gevallen een beslissing te nemen, moeten bijkomende gegevens uit de observatie-eenheid in de vergelijking worden betrokken. Met name in gezinsreconstructiestudies zijn dat vaak weer andere namen omdat andere identificatie-items ontbreken. Parochieregisters vermelden inderdaad

meestal niet alleen de naam van de boreling, huwende of overledene maar ook van de ouders c.q. (voormalige) echtgenoten. Iets gelijkaardigs geldt voor de huishoudfiches in volkstellingen of bevolkingsregisters (Wrigley 1973: 8-9). Als de moeder van Jan Peters de naam Maria Matthijs draagt maar die van Jan Pieters heet Jacoba Lotti, dan gaat het om twee verschillende Jannen. In vele gevallen is dergelijke informatie echter niet voorhanden en moet een beroep worden gedaan op andere dan naaminformatie.

A.2.3.2 Voornaamvariaties

Familienamen hadden in de 19^{de} eeuw al een betrekkelijk grote stabiliteit en toch zijn variaties al moeilijk automatisch te onderscheiden van verschillen. Bij voornamen liggen de zaken nog een stuk ingewikkelder omdat er nog meer variatiemechanismen in het geding zijn, die meestal te maken lijken te hebben met het feit dat de voornaam ook de roepnaam voor dagelijks gebruik is. De problematiek van orthografische variaties is dezelfde als die bij familienamen. Er komen bij voornamen echter veel vaker fonetische variaties voor en die kunnen heel drastisch zijn. Een aantal systematische processen is in het geding. Bloothoofd (1994a; 1998) noemt volgende fonetische variaties, typisch voor voornamen (de voorbeelden komen uit de Leuvense dataset):

- verklein- en vleivormen komen in het Frans veel minder voor dan in het Nederlands (*Julie* → *Juliette*);
- verkorte vormen en deelvormen, vaak gebruikt als roepnamen (*Elisabeth* → *Elise*; *Alexandre* → *Alex*);
- samenvoegingen van (delen van) voornamen (*Marie + Rosa* → *Marierose*);
- initialen, die verkeerd geïnterpreteerd kunnen worden (*Jean Baptiste* → *JⁿB^{le}*; *Ferdinand* → *F^{and}*);
- vertalingen, waaronder (pseudo-)Latijnse vormen (*Joseph* → *Josephus*; *Jean* → *Johannus*; *Avelin* → *Avelinus*) maar soms ook vernederlandsingen. De Leuvense ambtenaren registreerden de voornamen virtueel altijd in het Frans. Toch zijn in het bestand ook enkele Nederlandse naamvormen te vinden, die waarschijnlijk een gevolg zijn van leesfouten (*Julie* → *Julia*; *Marie* → *Maria*);
- pre- en suffixvariaties (*Louis* → *Alouis*; *Manuel* → *Emmanuel*; *Marie* → *Mariette*);

Vooral vervelend bij voornamen is dat lang niet altijd alle voornamen in gebruik blijven die bij de geboorte worden toegekend. De bevolkingsregisters vermelden *François Xavier Leo-*

pold bijvoorbeeld enkel als *Xavier*, of nu eens als *François*, dan weer als *Xavier*. Vaak maar lang niet altijd wordt de eerste voornaam gebruikt. Bij tweelingen is dat trouwens meestal de laatste naam omdat de rest gemeenschappelijk is. Soms wordt een bijnaam zo gewoon dat hij ook bij de volkstelling wordt opgegeven en dan is er niet noodzakelijk nog een verband met de naam uit de geboorteakte (vgl. Bouchard & Pouyez 1980: 120).

Voornaamvariaties zijn net als bij familienamen vaak typisch voor de gegeven onderzoekscontext en dat geldt ook voor de manier waarop onderzoekers er mee omspringen. De te onderscheiden methodes voor standaardisatie en similariteitsmeting komen in grote lijnen overeen met die voor familienamen (Bouchard & Pouyez 1980; Nygaard 1992; Bloothoof 1994a, 1994b, 1998). Voor dit project loont het echter niet de moeite zwaar te investeren in min of meer sluitende systemen. In tegenstelling tot de situatie in vele andere koppelingsprojecten zijn er immers nog andere, minder aan variatie onderhevige identificatie-items voorhanden, die in combinatie met de naamgegevens gebruikt kunnen worden (vgl. Oris 1990).

De voornaamgegevens worden op drie manieren bij de koppelingsbeslissingen betrokken. In eerste instantie worden de letterlijke voornamen in hun geheel met elkaar vergeleken. Als die niet overeenkomen worden de soundexen van de samenstellende voornamen afzonderlijk bepaald (voor *Marie Barbe Hélène* bijvoorbeeld de soundexen van *Marie*, van *Barbe* en van *Hélène*), waarbij minstens één soundex gemeenschappelijk moet zijn voor de twee observaties.

Ten slotte worden alle samenstellende delen van beide voornamen met elkaar vergeleken met behulp van de similariteitsindex *g*, bepaald op dezelfde manier als bij de familienamen.

A.2.3.3 Geboorteplaatsen

Geboorteplaatsen zijn eigennamen en dus stellen zich in principe dezelfde variatieproblemen als bij familienamen. Het aantal verschillende geboorteplaatsen is echter veel kleiner dan het aantal familienamen. Dit aantal is zo beperkt dat een omzetting naar een numerieke standaardcode binnen een kort tijdsbestek mogelijk is. Bovendien bestaat een volledig elektronisch repertorium van alle Belgische gemeenten die sinds 1831 bestaan hebben (Foulon e.a. 1981; NIS 1983).

Bij de Belgische onafhankelijkheid bedroeg het aantal gemeenten 2.739. In 1839 stond het koninkrijk 124 Limburgse en 119 Luxemburgse gemeenten af aan Nederland respectievelijk

aan het Groothertogdom Luxemburg. Tot in 1928 werden vervolgens 161 nieuwe gemeenten opgericht, terwijl er slechts dertien werden afgeschaft. Een team verbonden aan het *Département de démographie* van de *Université Catholique de Louvain-la-Neuve* richtte een gegevensbank op met alle gemeenten die vanaf 1831 tot 1981 bestaan hebben. De spelling is in de eerste plaats de officiële schrijfwijze, vastgesteld door het Belgische Staatsblad. Voor de geschrapte gemeenten gaat het over de laatst van kracht zijnde spelling. Bovendien bevat het repertorium naast de officiële benamingen ook spellingsvarianten en vertalingen. Wanneer verschillende gemeenten dezelfde naam droegen of dragen, bevat de plaatsnaam het administratief arrondissement (Foulon e.a. 1981).

In de jaren 1960 heeft het Nationaal Instituut voor de Statistiek aan iedere toen bestaande Belgische gemeente een codenummer van vijf cijfers toegekend. Nadien kregen ook de vóór 1960 geschrapte gemeenten een code. Het eerste cijfer duidt de provincie aan, het tweede het administratieve arrondissement van die provincie en de laatste drie cijfers de gemeente in kwestie (Foulon e.a. 1981; NIS 1983). Deze numerieke NIS-codes zijn gebruikt om de geboorteplaatsen te standaardiseren. Dat impliceert dat ook de gemeenten die tussen 1831 en 1910 werden afgestaan aan een buurland een NIS-code krijgen.

Aan 97 tot 98% van de individuele observaties kon een NIS-code voor de geboorteplaats worden toegekend. Deze heten in het vervolg de binnenlandse gemeenten, alhoewel enkele van de gemeenten dus ondertussen buiten Belgisch grondgebied liggen. Anderhalf tot 2% is geboren in een gemeente waarvoor geen NIS-code beschikbaar is (zie Tabel A.5). Deze gemeenten heten in het vervolg het buitenland en krijgen een landcode. Daarbij worden de huidige grenzen gehanteerd. Beieren, Westfalen en Pruisen krijgen bijvoorbeeld een *D* voor Duitsland.

Om de codes toe te kennen werd een gebruiksvriendelijk programma ontwikkeld dat gebruik maakt van een steeds uitbreidende lijst met varianten van plaatsnamen. Bij de aanvang bevat de lijst enkel de schrijfwijzen van het NIS-bestand (NIS 1983). Voor elke variant die niet in het bestand zit, kan met de hulp van het programma op een gemakkelijke wijze de juiste gemeente in de lijst worden opgezocht. Als die gevonden is, voegt het programma de variant automatisch met de bijhorende code aan de lijst toe.

Tabel A.5: Gestandaardiseerde geboorteplaatsen naar binnen- of buitenland en bevolkingsregister

	<i>BR1846</i>	<i>BR1856</i>	<i>BR1866</i>	<i>BR1880</i>	<i>BR1890</i>	<i>BR1900</i>
Gemeenten met NIS-code	97,50%	97,67%	98,32%	98,08%	98,40%	97,71%
Buitenlandse gemeente	1,99%	2,12%	1,44%	1,72%	1,42%	2,07%
Geboorteplaats onbekend	0,52%	0,20%	0,24%	0,20%	0,18%	0,22%
N	6791	4430	17587	17417	12413	4547

A.2.3.4 Geboortedata

Het heeft voor deze studie in theorie geen zin om over afwijkingen van de ware geboortedatum te spreken in termen van variatie. Een vermelde geboortedatum verwijst in het bevolkingsregister in principe naar de enige echte geboortedatum en anders gaat het om fouten. In gezinsreconstructiestudies die werken met parochiale doopboeken is dat anders: daar wijkt de datum van het doopsel met een onbekend aantal dagen af van de geboortedatum.

Fouten in de gecodeerde geboortedata ontstaan op twee manieren: door fouten in de bron zelf of door lees- en tyfouten van de codeurs. We gaan er van uit dat het bij de 19^{de}-eeuwse stadsambtenaren enkel om toevalsfouten gaat. Hetzelfde geldt voor de meeste lees- en tyfouten. Toch zit er ook een systematische component in de foutenmarge. De registers verwijzen courant naar de laatste vier maanden van het jaar op een deels Latijnse manier: *7^{bre}* staat voor *septembre* (zeven is in het Latijn *septem*), *8^{bre}* voor *octobre*, *9^{bre}* voor *novembre* en *X^{bre}* voor *decembre*. Hoewel de instructie, training en controle van de codeurs hier herhaaldelijk de nodige aandacht aan besteedde, werden data als *12 9^{bre} 1871* zoals verwacht regelmatig ten onrechte als 12/09/1871 ingevoerd. In het koppelingsproces is deze fout dan ook als een mogelijke variant van 12/11/1871 beschouwd.

A.3 Koppelingsstrategie en -resultaten

In de eerste plaats is een strategie nodig om persoonsgegevens te koppelen met het oog op de reconstructie van individuele levenslopen. Het gaat hierbij om interne koppelingen van observaties uit eenzelfde bevolkingsregister enerzijds en van observaties uit opeenvolgende bevolkingsregisters anderzijds. In de tweede plaats, cruciaal voor het onderzoek, zijn procedures nodig om de levenslopen van huwelijkspartners enerzijds en van ouders en kinderen anderzijds met elkaar te verbinden. Deze gezinsreconstructie omvat zowel interne als externe koppelingen.

A.3.1 Koppeling van individuele persoonsgegevens

Winchester (1973b; 1992) onderscheidt vier algemene koppelingssituaties naargelang er al dan niet fouten en variaties optreden in de identificatie-items en naargelang er al dan niet sprake is van multipliciteit. Van multipliciteit is sprake als verschillende personen dezelfde identificatie-items hebben. (1) Als de set identificatie-items variatieloos is en uniek voor elk individu, is koppeling eenvoudig: alle observaties met identieke items worden gekoppeld. (2) Wanneer de items weliswaar variatieloos zijn maar niet uniek, moet indien mogelijk bijkomende informatie worden gezocht. Anders moeten probabilistische methoden de beslissingen treffen. (3) De moeilijkste situatie is die waar er fouten en variaties zitten in de identificatie-items en er multipliciteit verwacht wordt. (4) Dit onderzoek zit echter in de vierde situatie: er zijn fouten en variaties in de identificatie-items maar we mogen aannemen dat er geen sprake is van multipliciteit. De koppelingsprocedure omvat dan minstens twee stappen: sorteren en in detail vergelijken.

A.3.1.1 Sorteren

Het is praktisch onmogelijk alle potentieel te koppelen observaties in detail met elkaar te vergelijken. Daarom moet een sleutel gekozen worden om sorteervakken te construeren waarbinnen een beperkter aantal gedetailleerde vergelijkingen mogelijk is. Gezien zijn universele beschikbaarheid is de soundex van de familienaam (SR(Familienaam)) een goede kandidaat. De bruikbaarheid van die gecomprimeerde vorm wordt echter uitgehold door het feit dat sommige variaties een verschillende code krijgen, waardoor ze ten onrechte niet in het zelfde sorteervak zitten. Bovendien veranderen sommige mensen van familienaam. Met name is dat het geval voor buitenechtelijk geboren, niet-erkende kinderen die later bij een huwelijk van hun moeder gelegitimeerd worden. Bij die gelegenheid veranderen zij van moeders familienaam naar die van de bruidegom en wettelijke vader. Dit leidt tot gemiste koppelingen.

Om dat te vermijden definieert in een tweede ronde het geboortjaar de sorteervakken. Met betrekking tot de registers van 1846 en 1856 kan een exact geboortjaar echter niet als sorteersleutel fungeren omdat vele geboortjaren op basis van de opgegeven leeftijd zijn berekend. Ook als de opgegeven leeftijd correct is, kan het berekende jaar daardoor één jaar hoger of lager dan het werkelijke liggen. Vanaf 1866 vermeldde de bevolkingsregisters wél systematisch het geboortjaar, zodat het exacte geboortjaar als sorteersleutel kan fungeren.

Het sorteren op basis van de soundex van de familinaam en op basis van geboortjaar is bij uitstek een taak voor de computer. In deze fase worden dus automatisch een groot aantal *negatieve* koppelingsbeslissingen genomen: observaties die nooit in hetzelfde vak zitten, worden niet gekoppeld. In de volgende fase worden ook *positieve* koppelingsbeslissingen genomen.

A.3.1.2 Vergelijken

Binnen de sorteervakken gaat de procedure na of de identificatie-items van de twee voorliggende observaties overeenstemmen. Om het even welke koppelingsprocedure steunt daarbij op het beginsel van identiteit van niet-onderscheidbare individuen ('Identity of Indiscernibles principle'). Dit beginsel houdt in dat als twee observaties identieke identificatie-eigenschappen hebben, we er van uitgaan dat ze bij één en hetzelfde individu horen (Winchester 1973a: 23-26).

Geen enkel identificatie-item is op zich altijd uniek voor een gegeven individu. In de bevolkingsregisters is echter altijd een combinatie van identificatie-items beschikbaar waarvan we aannemen dat hij wél uniek is. De gevolgde koppelingsprocedure gaat er meer bepaald van uit dat er in Leuven geen twee personen hebben geleefd met dezelfde familienaam, voornaam (waaruit ook het geslacht werd afgeleid), geboortedatum én geboorteplaats. Met andere woorden: een observatie waarvoor al deze identificatie-items voorhanden zijn, beschouwen we als volledig geïdentificeerd. Als één of enkele van deze identificatie-items niet volledig beschikbaar zijn, kan gebruikgemaakt worden van hulpitems: informatie over de huwelijks-carrière, migratie, beroep of familierelaties.

Omwille van de fouten en variaties moet voorzien worden in een aantal excuusprocedures: men kan niet altijd vasthouden aan de eis dat de identificatie-eigenschappen exact dezelfde moeten zijn. Sommige afwijkingen zien we door de vingers omdat we veronderstellen dat het om variaties of fouten gaat. De basis waarop beslissingen verantwoord worden, blijft echter het principe van de identiteit van niet-onderscheidbaren. Om uit te maken welke afwijkingen excuseerbaar zijn en welke niet, moet gesteund worden op eerder algemene achtergrondkennis enerzijds en kennis omtrent de specifieke bronnen waaruit de observatie-eenheden komen anderzijds (Winchester 1973a: 23-26). Een voorbeeld van de eerste soort kennis: 21/7/1850 is excuseerbaar ten opzichte van 12/7/1850 omdat twee getallen gewoon van plaats veranderd

zijn. Een voorbeeld van kennis omtrent de bron: 12/08/1850 is excuseerbaar ten opzichte van 12/10/1850 omdat de maand oktober in het register vaak als 8^{bre} genoteerd werd.

De nood aan excuusprocedures roept twee strategisch belangrijke vragen op. Ten eerste: in hoeverre kan een algoritme de koppelingsbeslissingen automatisch nemen? En ten tweede: streeft de koppelingsprocedure accuraatheid dan wel volledigheid na? Bij automatische koppeling moeten alle excuusprocedures exact in de algoritmes worden opgenomen. Bevatten de programma's te weinig excuses, dan worden reële koppelingen gemist en is de gekoppelde populatie dus onvolledig. Gaan de excuses te ver of zijn ze te algemeen, dan worden systematisch foute koppelingen gemaakt en is dus de accuraatheid suboptimaal.

1. De beslissing om koppelingen automatisch, semi-automatisch dan wel volledig met de hand te maken moet uiteraard rekening houden met het aantal observaties, het beschikbare budget, informatica-assistentie en computerkracht. Verder hangt de keuze samen met de beschikbaarheid van identificatie-items en de mate waarin er fouten en variaties in optreden. In de virtuele situatie waar unieke en foutloze identificatie-items altijd beschikbaar zijn, is automatisering gemakkelijk volledig door te voeren. Dat gaat moeilijker als nu eens deze (combinatie van) items een persoon moeten identificeren en dan weer een andere reeks items, waarbij er bovendien fouten en variaties optreden. In dat geval moet de koppelingsvergelijking de overeenstemming en tegenstrijdigheid in de identificatie-items wege om tot een koppelscore te komen (Winchester 1992: 157-158). Positieve gewichten weerspiegelen dan de mate van zekerheid die er is omtrent de overeenstemming in de items. Negatieve gewichten weerspiegelen de zekerheid omtrent de tegenstrijdigheid.

Automatische koppelingsprocedures omvatten altijd een systeem om koppelscores toe te kennen ('matchscoring'). Dit systeem kent aan een paar observaties een waarde toe die moet aangeven hoe waarschijnlijk het is dat het om hetzelfde individu gaat. De drempelwaarde waarboven wel en waaronder niet gekoppeld wordt, is door de onderzoeker te bepalen. De gewichten die men aan de indentificatievariabelen toekent, zijn hierbij uiteraard van cruciaal belang (Wrigley 1973: 10). In de praktijk verschillen de gehanteerde systemen voor weging en matchscoring van project tot project.

Schofield (1992) houdt een expliciet pleidooi voor een volautomatische koppelingsprocedure. Hij geeft een theoretisch-methodologisch en een praktisch argument. Het eerste heeft te maken met transparantie: als de beslissingen om al dan niet te koppelen op objectieve princi-

pes gebaseerd zijn, dan moeten deze principes ook in de vorm van een algoritme en dus van een computerprogramma uit te drukken zijn. De praktische reden heeft te maken met het aantal te maken koppelingsvergelijkingen en dus met het aantal te nemen beslissingen. De werkuren die daarvoor nodig zijn overschrijden al snel de capaciteit van bestaande onderzoeksteams. Schofield waarschuwt echter dat ook automatisering zeer grote budgettaire implicaties heeft. De Cambridge Group heeft zeventien jaar gewerkt aan de ontwikkeling en implementatie van een automatische koppelingsprocedure en zelfs dan blijft een eindbeoordeling door de onderzoeker noodzakelijk:

"As has been mentioned, before the records can be linked, the somewhat wayward spelling on the part of parish clerks in the past needs to be standardized. Indeed, spelling variations were so arbitrary that the historian must check the results carefully and take final responsibility for deciding which are variant spellings and which are separate names" (Schofield 1992: 78).

Principieel is het transparantieargument van Schofield terecht. Maar gegeven de budgettaire beperkingen en de relatief kleine omvang van dit project is de voorkeur gegeven aan semi-automatische of *computer-assisted* koppeling: de gemakkelijke beslissingen worden via de computerprogramma's genomen maar waar twijfel mogelijk is, beslist de onderzoeker. Twee argumenten verantwoorden dit. Ten eerste vergt de ontwikkeling van een geldige en automatische procedure om variaties te onderscheiden van verschillen een bijzonder grote tijdsinvestering, die maar loont als het aantal observaties zeer hoog oploopt. In Cambridge (Schofield 1992), Quebec (Bouchard 1992) of Utah (Wesley e.a. 1987) gaat het over miljoenen observaties en loont een forse investering de moeite. In onderhavig eenmansproject gaat het om ongeveer 70.000 observaties van persoonsgegevens (huis- en gebeurtenisobservaties niet meegerekend, al worden die bij twijfelgevallen wel in het koppelingsproces betrokken). Over de grote meerderheid kan eenvoudig en automatisch beslist worden al dan niet te koppelen. Het volume met de hand te behandelen vergelijkingen blijft dus hanteerbaar. Ten tweede beschouwt de literatuur het handmatig onderscheiden van varianten en verschillen impliciet als geldiger. De geldigheid van automatische procedures wordt in de methodologische bijdragen immers geëvalueerd door de beslissingen van het computerprogramma te vergelijken met handmatige resultaten. Die laatste fungeren in de evaluatie als de 'ware waarden' (Schwartz e.a. 1984; Adman e.a. 1992; Atack e.a. 1992; Davies 1992; Vetter e.a. 1992; Harvey & Green 1994; Bloothoof 1994b, 1998). Adman e.a. (1992) betogen expliciet dat handgemaakte koppelingen geldiger zijn dan volautomatische en bepleiten een computer-

assisted of semi-automatisch koppelingsproces (zie ook Vetter e.a. 1992).⁷ De te verwachten prijs van die menselijke interventie is echter een onbetrouwbaarheidsmarge: bij het nemen van een groot aantal koppelingsbeslissingen maken mensen onherroepelijk fouten, al was het maar uit verstrooidheid. We nemen aan dat het hierbij enkel om toevalsfouten gaat.

In de praktijk combineren de meeste projecten computeralgoritmes met menselijke interventie, vooral omwille van de naamvariaties (Harvey & Green 1994). Dat is ook in dit project zo. Meer bepaald werd de vergelijkingsfase opgesplitst in twee subfasen, de selectie- en de beslissingsfase, waarvan alleen de eerste volledig automatisch verloopt. De selectiefase hanteert zeer ruime excuusprocedures en streeft volledigheid na: alle observaties die mogelijk kandidaat zijn voor koppeling, worden automatisch als zodanig aangeduid. In de beslissingsfase worden enkel de observaties met identieke identificatie-items automatisch gekoppeld. Over de rest beslist de onderzoeker handmatig.

2. De tweede strategische kwestie inzake koppelingsvergelijkingen is of ze de accuraatheid van de koppelingen wil maximaliseren dan wel de volledigheid. De accuraatheid is de verhouding van het aantal *terechte* koppelingen tot het totale aantal *gemaakte* koppelingen. De volledigheid is de verhouding van het aantal *gemaakte* koppelingen tot het aantal *te maken* koppelingen. Als niet alle observaties volledig geïdentificeerd zijn en er fouten optreden in de identificatie-items, zijn volledigheid en accuraatheid contradictoir (Bouchard 1992: 69-70). Preciezer: een procedure die de accuraatheid maximaliseert, bereikt niet de maximale volledigheid. Een procedure die volledigheid maximaliseert gaat ten koste van maximale accuraatheid. De mate waarin beide maxima met elkaar te verzoenen zijn, hangt af van de kwaliteit van de data.

De optimale strategie hangt af van het doel van de studie. Voor genetische studies is niet de volledigheid maar wel de accuraatheid prioritair. Bij betwifelbare variaties of onzekere identificatie wordt dan ook niet gekoppeld, ook al leidt dit eventueel tot een niet-representatieve steekproef van gekoppelde observaties. In demografische studies is die representativiteit meestal belangrijker en daarom streven ze meestal eerder volledigheid na. De submaximale accuraatheid is geen probleem als de onterechte koppelingen via toevalsmechanismen totstandkomen (Bouchard 1992: 70).

⁷Bloothoofthoof schrijft bijvoorbeeld: "Nominal record linkage is a technology and should be judged by its ability to come to the same results as humans do" (1998: 54).

Dit koppelingsproject legt de prioriteit bij volledigheid. Die wordt in de eerste plaats al bevorderd door het hanteren van twee verschillende gecomprimeerde sleutels om sorteervakken te definiëren: de soundex van de (semi-fonetisch gestandaardiseerde) familienaam en een interval van geboortejaren. Dit gaat niet ten koste van de accuraatheid omdat op dit niveau nog geen positieve koppelingsbeslissingen worden genomen. Er is wel een kost in termen van accuraatheid bij de volgende beslissingsregels:

1. het ontbreken van één of enkele identificatie-items vormt geen beletsel voor koppeling; als de items die voorhanden zijn overeenkomen, wordt er (tenminste voorlopig) gekoppeld;
2. als er twijfel mogelijk is of twee vormen varianten zijn dan wel verschillen, wordt (voorlopig) aangenomen dat het om varianten gaat;
3. koppelingen (verder voorgesteld als "=") zijn transitief: als $A = B$ en $B = C$, dan geldt (voorlopig) automatisch $A = C$ (zie Bouchard 1992: 68);
4. om de volledigheid te bevorderen geldt een negatieve variant van regel 3 niet (in tegenstelling tot Bouchard 1992: 68); de procedure neemt met name *niet* automatisch aan dat als $A = B$ en $A \neq C$ dan $B \neq C$. B wordt nog expliciet met C vergeleken. Als het verdict luidt dat $B = C$ dan geldt volgens regel 3 automatisch dat ook $A = C$, terwijl eerder het omgekeerde was beslist. De oplossing van deze tegenstrijdigheid bestaat in het laten vallen van de zwakste koppeling in de constellatie.

A.3.1.3 Selecteren van koppelingskandidaten

De automatische selectie van koppelingskandidaten verloopt in twee ronden voor twee sorteersleutels. De eerste ronde vergelijkt de procedure alle koppels van observaties met dezelfde SR(Familienaam) én met hetzelfde geslacht. De tweede ronde vergelijkt alle koppels met (bij benadering) hetzelfde geboortjaar maar met een verschillende SR(Familienaam) of met een verschillend geslacht. Omdat de registers van 1846 en 1856 vaak enkel de leeftijd vermeldden, kan het berekende geboortjaar in principe met één jaar afwijken van het juiste geboortjaar. Daarom worden met betrekking tot deze registers alle observaties vergeleken waarvan het (berekend) geboortjaar maximaal met één jaar afwijkt. Vanaf 1866 is de sorteersleutel een exact geboortjaar. Het aantal extra koppelingen dat via deze tweede ronde ontdekt zal worden, is naar verwachting laag: het gaat enkel om familienaamvariaties met een verschillende soundex of om mensen wiens familienaam in de loop van hun leven veranderde

(bijvoorbeeld na legitimatie). Toch ligt het aantal te vergelijken observaties in de tweede ronde het hoogst want er zijn veel minder verschillende geboortejaren (ongeveer 100) dan verschillende soundexen (800 tot 1400, naargelang register).

Binnen elk sorteervak gaat het computerprogramma na of de identificatievariabelen van de twee voorliggende observaties voldoende overeenkomen om te besluiten dat het mogelijk om dezelfde persoon gaat. Dat gebeurt via een weegstelsel dat kwalitatief genoemd kan worden. Om de doorzichtigheid van de beslissingen te verhogen houdt het programma immers bij voor welke identificatie-items er overeenstemming of gelijkheid is en voor welke niet, in plaats van de mate van gelijkheid voor de gehele reeks van identificatie-items in één getal uit te drukken. Het resultaat en de gehanteerde logica zijn echter dezelfde: kandidaat voor koppeling zijn alle paren van observaties waarvan de combinatie van overeenstemmende identificatie-items een voldoende hoog onderscheidingsvermogen heeft om een persoon in Leuven te identificeren. De combinatie van enkel geboorteplaats en -jaar heeft dat niet. Als echter ook gelijkheid is op het vlak van familienaam en voornaam, gaat het mogelijk om dezelfde persoon. Deze procedure garandeert dat alle potentiële kandidaten voor koppeling geselecteerd worden. Een fout in één identificatie-item verhindert immers niet dat de observaties toch nog als koppelingkandidaat geselecteerd worden.

Figuur A.1 geeft de precieze criteria die de computerprogramma's hanteerden voor de selectie van koppelingkandidaten. Ze zijn zowel voor de intra- als voor de interregisterkoppelingen gehanteerd. Alle observatieparen die voldoen aan de voorwaarden voor één van de 23 getallen in de laatste kolom van de figuur, zijn kandidaat voor koppeling. Enkele voorbeelden van voldoende voorwaarden zijn (met verwijzing naar de figuur):

- de combinatie van een identieke familienaam met een identieke voornaam (lijn 1);
- gelijkheid op het vlak van familienaam en voornaam en een identieke geboortedatum (lijn 10);
- een verschillende familienaam maar gelijkheid *qua* voornaam en een identieke geboortedatum en -plaats (lijn 15);
- gelijkheid *qua* familienaam, een identieke voornaam, geboorteplaats en een zelfde geboortjaar en -maand maar een verschillende geboortedag (lijn 17).

In de tweede ronde, waar het geboortjaar de sorteersleutel is, hoeven de observaties met een identieke familienaam en hetzelfde geslacht niet meer met elkaar vergeleken te worden omdat dat in de eerste ronde al gebeurd is.

De gelijkenis tussen twee familienamen wordt bepaald met behulp van de eerder beschreven similariteitsmaat g . Minstens twee op drie letters moet in beide namen op dezelfde relatieve positie voorkomen ($g > 0.66$). Dezelfde maat wordt ook toegepast op de voornamen, waarbij voor minstens één van de samenstellende voornamen moet gelden dat g minstens twee derde bedraagt. Als voornaam 1 bijvoorbeeld *Anne Marie* is en *Mariette* voornaam 2, dan is er sprake van gelijkenis want g is voor de samenstellende voornamen *Marie* en *Mariette* groter dan 0.66.

Figuur A.1: Automatische selectie-procedure van koppelingskandidaten

Sorteersleutel	S(Familienaam)	Voornamen	Geboortedatum	(Berekend) geboortejaar	Geboorteplaats		
SR(Familienaam) + geslacht	Identiek	Identiek				1	
		Gelijkenis ²	Identiek				2
			Dag óf maand óf jaar verschillend			Identiek	3
			Onbekend	Maximaal verschil van 1 cijfer of 2 jaren	Identiek	4	
		Verschillend	Identiek				5
			Onbekend	BR1846&'56: Maximaal verschil 1 jaar Vanaf BR1866: Identiek	Identiek	6	
	Gelijkenis ¹		Identiek				7
		Dag óf maand óf jaar verschillend			Identiek	8	
		Onbekend	Maximaal verschil 1 cijfer of 2 jaren	Identiek	9		
	Gelijkenis ²	Identiek				10	
		Dag óf maand óf jaar verschillend			Identiek	11	
		Onbekend	Maximaal verschil 1 cijfer of 2 jaren	Identiek	12		
	Verschillend	Identiek				13	
		Onbekend	BR1846&'56: Maximaal verschil 1 jaar Vanaf BR1866: Identiek	Identiek	14		
		Verschillend	Gelijkenis ²	Identiek	Identiek	15	
Geboortejaar*	Gelijkenis ¹	Identiek				16	
		Dag óf maand óf jaar verschillend			Identiek	17	
		Onbekend			Identiek	18	
	Gelijkenis ²	Identiek				19	
		Dag óf maand óf jaar verschillend			Identiek	20	
		Onbekend			Identiek	21	
	Verschillend	Identiek				22	
	Verschillend	Gelijkenis ²	Identiek			Identiek	23

* Sorteersleutel BR1846 en BR1856: interval van drie geboortejaren $[j-1, j+1]$; vanaf BR1866: exact geboortejaar. Binnen deze sorteervakken worden de observaties met dezelfde SR(Familienaam) buiten beschouwing gelaten.

¹ Er is gelijkenis tussen twee familienamen N_1 en N_2 als $g(N_1, N_2) > 0.66$.

² Er is gelijkenis tussen twee voornamen als voor minstens één van de samenstellende voornamen geldt dat $g(N_1, N_2) > 0.66$.

Deze selectieprocedure is permissief om de volledigheid te optimaliseren. Niettemin reduceert ze het aantal in de beslissingsfase te vergelijken paren tot een tien- tot een honderdduizendste van het totale aantal combinaties (van minstens 9,8 miljoen tot een paar honderd of een paar duizend, afhankelijk van het register).⁸ De selectiekracht van de procedure neemt uiteraard toe naarmate de observaties vollediger geïdentificeerd zijn. Vooral het ontbreken van een volledige geboortedatum verzwakt in de registers van 1846 de selectiekracht: het aantal geselecteerde koppelingkandidaten is daar bijna de helft van het aantal observaties. In de overige registers is dat een tiende (BR1900) tot ruim een kwart (BR1866, zie Tabel A.6). Niettemin blijkt de efficiëntie van de selectieprocedure in elk register uit de reductie van het totale aantal combinaties (minstens 9,8 miljoen) tot een paar honderd à een paar duizend geselecteerde combinaties.

Tabel A.6: Aantal geselecteerde kandidaten voor intraregisterkoppeling naar periode en selectieronde

	<i>BR1846</i>	<i>BR1856</i>	<i>BR1866</i>	<i>BR1880</i>	<i>BR1890</i>	<i>BR1900</i>
▪ N	6817	4430	17587	17417	12413	4547
▪ Aantal combinaties van 2 observaties	23,2 mln.	9,8 mln.	154,6 mln.	151,6 mln.	77,0 mln.	10,3 mln.
▪ Geselecteerd:	2699	894	4872	3799	1822	465
1.Soundex	1618	638	3857	3082	1431	409
2.Geboortejaar	1081	256	1015	717	391	56

De interregisterkoppelingen gebeuren volgens dezelfde selectieprocedure als de intraregisterkoppelingen maar beperken zich tot observaties uit de drie steekproefgeneraties omdat anders het aantal te vergelijken observaties onhaalbaar hoog zou oplopen. In een eerste stap worden alle referentiepersonen (uit G1830, G1850 en G1864) uit één register vergeleken met alle personen van dezelfde generatie in de twee andere registers. In een tweede stap worden alle personen van elke referentiegeneratie waarvoor ten onrechte nog geen koppeling gevonden is in een volgend of vorig register, vergeleken met alle individueën uit de respectievelijke sorteervakken van het vorige of volgende register. Er is 'ten onrechte' nog geen koppeling voor alle referentiepersonen die bij het afsluiten of opstarten van een bevolkingsregister geregistreerd zijn als wonend in Leuven, terwijl ze nog niet teruggevonden zijn in het volgende, respectie-

⁸Het aantal in principe te vergelijken paren is het aantal combinaties uit N in een groep van 2 elementen.

velijk het vorige register.⁹ Tabel A.7 geeft het aantal voor interregisterkoppeling geselecteerde paren van observaties naar de generatie en de combinatie van bevolkingsregisters waar het om gaat en naar selectieronde.

Tabel A.7: Aantal geselecteerde kandidaten voor interregisterkoppeling naar periode en selectieronde

GENERATIE	AANTAL POTENTIELE KOPPELINGEN					
	Sorteervak (zie tekst):		SNDX	GBJR	SNDX	GBJR
GENERATIE 1830	BR1846-1856		BR1856-1866		BR1846-1866	
Stap 1: vgl. G1830ers uit beide BR met elkaar	416	(*)/	371	/	286	/
Stap 2: nog geen link (zie tekst)						
- 2a. G1830ers uit 1 ^{ste} met alle uit 2 ^{de} BR	294	159	385	126	380	14
- 2b. G1830ers uit 2 ^{de} met alle uit 1 ^{ste} BR	156	202	73	156	163	10
GENERATIE 1850	BR1866-1880		BR1880-1890		BR1890-1900	
Stap 1: vgl. G1850ers uit beide BR met elkaar	728	/	596	/	575	/
Stap 2: nog geen link (zie tekst)						
- 2a. G1850ers uit 1 ^{ste} met alle uit 2 ^{de} BR	60	267	15	118	21	168
- 2b. G1850ers uit 2 ^{de} met alle uit 1 ^{ste} BR	21	15	15	69	11	95
GENERATIE 1864	BR1880-1890		BR1890-1900		BR1880-1900	
Stap 1: vgl. G1864ers uit beide BR met elkaar	802	/	589	/	535	/
Stap 2: nog geen link (zie tekst)						
- 2a. G1864ers uit 1 ^{ste} met alle uit 2 ^{de} BR	25	233	7	86	4	130
- 2b. G1864ers uit 2 ^{de} met alle uit 1 ^{ste} BR	19	44	12	17	3	22

(*) /: Niet van toepassing

A.3.1.4 Beslissen

Wanneer twee, automatisch geselecteerde koppelingkandidaten voldoende gelijke identificatie-items hebben, wordt de daadwerkelijke koppeling automatisch gemaakt. Concreet worden twee kandidaten als voldoende gelijk beschouwd als de semi-fonetisch gestandaardiseerde familienamen, de voornamen, de geboortedata en -plaatsen voor zover bekend identiek zijn; als er dus geen enkele tegenstrijdigheid bestaat tussen gekende indentificatie-items (en dat is op een handvol uitzonderingen na minstens familienaam, voornaam, geboorteplaats en leeftijd of geboortjaar). In alle andere gevallen moet de onderzoeker de knoop doorhakken, geassisteerd door een aantal computerprogramma's. De assistentie bestaat uit het op het scherm

⁹Met andere woorden: voor een gegeven persoon uit register 1 is 'ten onrechte' nog geen koppeling gevonden in een volgend register 2 als die persoon bij het afsluiten van register 1, naar aanleiding van een nieuwe volkstelling, nog geregistreerd staat, maar niet wordt teruggevonden in register 2. Vice versa: een persoon uit register 2 is 'ten onrechte' nog niet gekoppeld in register 1 als die persoon bij de aanvang van register 2, naar aanleiding van de volkstelling, teruggevonden wordt maar niet in het vorige register 1. Het gebrek aan koppeling kan in dergelijke gevallen in feite terecht zijn als het gaat om emigraties respectievelijk immigraties die niet in de register 1 geregistreerd werden. De paragraaf over de betrouwbaarheid van de registratie gaat hier verder op in.

accentueren van de precieze afwijkingen in de identificatie-items enerzijds, en uit het weergeven van bijkomende informatie anderzijds. Die bijkomende informatie omvat de identificatie van vader en moeder voor zover bekend en alle geregistreerde gebeurtenissen (vooral huwelijk, verhuis, migratie en overlijden maar ook erkenningen en legitimaties) die horen bij de twee observaties.

Mensen waarvoor 'ten onrechte' geen koppeling tot stand kon worden gebracht, mogen we niet uit de steekproef weglaten. Een steekproef die enkel rekening houdt met de succesvol gekoppelde individuen, zal vertekend zijn omdat dan over het algemeen het stabiele en relatief meer welvarende deel van de bevolking oververtegenwoordigd zal zijn. Het mobielere en armere deel van de bevolking liep immers een grotere kans door de mazen van het registratiesysteem te ontsnappen (Wrigley 1973: 12-13; Herlihy 1973).

A.3.1.5 Valideren

Elke voorlopig gemaakte koppeling kan ongedaan worden gemaakt als bij de validering blijkt dat ze fout is; als met name blijkt dat ze leidt tot incompatibiliteit, multipliciteit of demografisch onrealistische constellaties (Wrigley & Schofield 1973). Hoewel gedurende de gehele beslissingsfase controleprocedures worden ingebouwd, is de validering pas definitief nadat huwelijks- en ouderschapsrelaties zijn gelegd (zie infra) omdat foute koppelingen vooral in de context van een geheel van koppelingen aan het licht komen.

Wrigley en Schofield (1973: 75-78) brengen de problemen die opduiken bij de beoordeling van kandidaat-koppels onder in twee categorieën: incompatibiliteit en multipliciteit. Van incompatibiliteit is sprake als A gekoppeld kan worden met B en B met C, maar er is geen koppeling mogelijk tussen A en C. Bij multipliciteit kan A gekoppeld worden met B of met C, maar niet met beide.

Om dit te verduidelijken twee realistische voorbeelden, eerst van incompatibiliteit. Voor A zijn de ouders geïdentificeerd maar voor B niet. A en B komen overeen en worden gekoppeld volgens de voorziene beslissingsregels. C is volgens diezelfde regels dezelfde persoon als B, die volgens de vorige beslissing dezelfde persoon is als A. De ouders van C zijn echter, net als die van A, geïdentificeerd maar de respectievelijke ouderparen komen duidelijk niet overeen. A en C kunnen dus, zo blijkt bij de validering, niet dezelfde personen zijn, tenzij één van beide ouderparen niet correct geïdentificeerd is. Vanuit het standpunt van B is er sprake van

multipliciteit: B kan ofwel met A, ofwel met C worden gekoppeld maar niet met beide want dan ontstaat incompatibiliteit.

Om in zulke gevallen te kunnen beslissen, worden twee stappen gezet: eerst worden de identificatoren in de oorspronkelijke bron gecontroleerd. Als na eventuele correctie de incompatibiliteit en/of multipliciteit nog blijft bestaan, worden de zwakste koppelingen verwijderd tot het probleem is opgelost. De sterkte van een koppeling wordt bepaald door middel van *match-scoring*: het toekennen van een score voor de mate van overeenstemming van de identificatoren. De volgende paragraaf legt de concreet gevolgde scoreprocedure uit.

Sommige koppels zijn incompatibel omdat ze tot demografisch onrealistische constellaties leiden. Daarvan is sprake als kandidaat-koppels demografisch onmogelijke toestanden impliceren. Veronderstel volgende kandidaat-koppeling: 'A is kind van C' en 'B is kind van C' maar tussen de geboortedata van A en B liggen slechts drie maanden. Dan is ofwel een geboortedatum, ofwel één van beide koppelingen verkeerd. Al tijdens de beslissingsfase werden, ter validering van de koppelingen, een aantal demografische restricties ingevoerd:

1. niemand sterft later dan op honderdjarige leeftijd;
2. een moeder kan ten vroegste overlijden op de dag van de geboorte van haar kind;
3. bij de geboorte van een kind is de moeder nooit jonger dan vijftien en nooit ouder dan vijftig jaar;
4. een vader kan ten vroegste acht maanden voor de geboorte van zijn (veronderstelde) kind overlijden.
5. bij de geboorte van een kind is de vader nooit jonger dan vijftien en nooit ouder dan 75 jaar;
6. het geboorte-interval tussen twee kinderen van dezelfde moeder is nooit minder dan tien maanden;
7. bij een huwelijk zijn noch bruid noch bruidegom jonger dan zestien jaar.

De gevolgde procedure om demografisch onrealistische toestanden te vermijden – en meteen de koppelingen te valideren – is dezelfde als die bij multipliciteit en incompatibiliteit: eerst de identificatoren controleren en dan, indien nog nodig, de zwakste koppelingen laten vallen.

Aan het einde van het koppelingsproces moeten alle koppelingen een grote test doorstaan (vgl. Wrigley & Schofield 1973: 76). Elke observatie die aan een persoon is toegeschreven moet gekoppeld zijn aan elke observatie die aan dezelfde persoon is toegeschreven en geen

enkele observatie die met observaties omtrent de ene persoon is gekoppeld, mag ook nog gekoppeld zijn aan observaties die niet met de andere observaties van die persoon gekoppeld zijn. Met andere woorden: er mag geen incompatibiliteit meer zijn en geen multipliciteit.

Zelfs dan nog zou het naïef zijn te denken dat alle gemaakte koppelingen terecht zijn of dat er je alle eigenlijke koppelingen ook hebt gevonden:

"there will always be doubtful cases in historical record linkage. Furthermore, there will be cases where there is no apparent doubt [...] but where the link is yet false, chance coincidence in names and other characteristics having produced a spurious link" (Wrigley 1973: 15).

A.3.1.6 Matchscoring

Om hier al een idee te krijgen van de betrouwbaarheid van de genomen beslissingen, krijgen alle koppelingen een score. 'Matchscoring' of het bepalen van een koppelscore is het toekennen van een waarde aan een paar van observatie-eenheden. Die waarde moet aangeven hoe waarschijnlijk het is dat het inderdaad om hetzelfde individu gaat (Wrigley 1973: 10). In volautomatische koppelingsprocedures wordt het beslissen doorgaans overgelaten aan een complex systeem van koppelscores. Hier is *tijdens* het beslissingsproces geen gebruik gemaakt van koppelscores, enkel *achteraf* om de genomen beslissingen te evalueren en om tot een beslissing te komen bij multipliciteit en incompatibiliteit.

In een eenvoudig systeem van koppelscore levert overeenstemming tussen twee waarden van elke identificatievariabele één positief punt op, kost elk conflict een negatief punt en blijft de koppelscore 0 in andere gevallen, bijvoorbeeld wanneer de informatie in minstens één van de identificatievelden niet voorhanden is. In een complexer systeem wordt aan elke identificatievariabele een verschillend gewicht toegekend: overeenkomst van naam zou bijvoorbeeld meer punten opleveren dan overeenkomst van geboorteplaats (Wrigley & Schofield 1973: 88-89).

In dit onderzoek werd aan elke besliste koppeling als volgt een score S toegekend. Voor elke identificatievariabele werd de mate van overeenkomst vermenigvuldigd met de entropie H van die variabele in het betreffende bevolkingsregister, en S is dan de som van al die producten. De overeenkomst van een bepaalde identificator weegt dus zwaarder door in S naarmate de uniciteit of informatiewaarde van die variabele groter is, en hoe groter de overeenkomst, hoe groter S . De mate van overeenkomst voor familie- en voornamen werd gemeten met de eerder beschreven similariteitsmaten (een proportioneel Guth-algoritme). De overeenkomst voor (de NIS-code van) geboorteplaats, geboortejaar, -maand en -dag is -1 ingeval van ver-

schil, 0 als een identificator ontbreekt, en 1 bij identiteit.¹⁰ De koppelscore is maximaal als er een perfecte overeenkomst is voor alle identificatoren.

A.3.1.7 Resultaten

De volgende tabellen geven een overzicht van de koppeling van individuele persoonsgegevens *na* validering. Dat betekent dat het gaat om de definitieve resultaten *nadat* ook de gegevens van huwelijkspartners, ouders en kinderen gekoppeld en gevalideerd zijn. Tabel A.8 bevat alle kandidaten die in aanmerking zijn genomen voor intraregisterkoppeling naar koppelscore en naar de uiteindelijk genomen beslissing.

Tabel A.8: Kandidaten voor intraregisterkoppeling naar bevolkingsregister en genomen beslissing (al dan niet gekoppeld): procentuele verdeling naar koppelscore S

Koppelscore S	BR1846		BR1856		BR1866		BR1880		BR1890		BR1900	
	Nee	Ja	Nee	Ja	Nee	Ja	Nee	Ja	Nee	Ja		
Gekoppeld?												
<0,00	0,0	0,0	0,5	4,6	0,01	0,0	0,0	0,0	0,0	0,0	0,0	0,0
0,00-4,99	4,2	0,0	13,4	0,0	8,1	0,0	10,4	0,0	11,4	0,0	17,0	0,0
5,00-9,99	13,1	0,0	26,2	0,0	11,7	0,0	7,4	0,0	7,2	0,0	19,0	0,0
10,00-14,99	42,2	0,0	25,7	0,8	26,1	0,1	25,1	0,1	20,2	0,0	0,7	0,0
15,00-19,99	34,4	2,2	19,4	3,8	26,4	0,4	16,3	0,2	14,3	0,0	11,6	0,0
20,00-24,99	5,3	17,7	8,0	3,1	11,9	2,5	17,1	2,9	18,2	0,8	12,9	0,0
25,00-29,99	0,7	20,7	4,2	12,2	9,6	4,6	15,2	3,0	20,4	3,1	22,4	2,8
30,00-34,99	0,0	59,5	2,1	13,0	4,0	9,6	5,3	7,3	6,0	5,5	11,6	0,6
35,00-39,99	0,0	0,0	0,5	62,6	2,3	82,8	3,2	86,6	2,2	90,6	4,8	96,5
<i>Totaal N = 100%</i>	2467	232	763	131	3009	1863	2467	1332	970	852	147	318
<i>Maximale S</i>		33,37		36,24		39,45		39,49		39,44		38,07
<i>Gemiddelde S</i>	13,71	29,75	12,59	31,09	16,13	37,59	17,54	37,99	17,88	38,43	18,54	37,74

Uit Tabel A.8 blijkt ten eerste dat er een zinvolle relatie bestaat tussen de koppelscore en het al dan niet aanvaarden van een kandidaat. Verworpen kandidaten situeren zich vooral in de onderste helft van de verdeling van de score, en aanvaarde koppelingen vooral in de bovenste helft. Niettemin blijkt ook dat, op basis van deze *scoring*, een automatische procedure niet tot juiste resultaten zou hebben geleid, want de score-verdelingen van verworpen en aanvaarde koppels overlappen elkaar. Toch blijkt ook, ten derde, dat aanvaarde koppelingen nooit een score S lager dan tien kregen. Kandidaten met een lagere S hadden dus, achteraf bekeken, ook al in de selectiefase geweerd kunnen worden. Enige uitzondering is BR1856, waar een

¹⁰ De exacte functiebeschrijving van S is als volgt, waarin *i* staat voor de eerder beschreven similariteitsmaat, H voor de entropie van de identificator, en o voor het al dan niet (-1, 0 of 1) gelijk zijn van twee kenmerken:

$$S = i_{SD(\text{familienaam})}H_{SD(\text{familienaam})} + i_{\text{voornaam}}H_{\text{voornaam}} + o_{NIS(\text{geboorteplaats})}H_{NIS(\text{geboorteplaats})} + o_{\text{geboortjaar}}H_{\text{geboortjaar}} + o_{\text{geboortemaand}}H_{\text{geboortemaand}} + o_{\text{geboortedag}}H_{\text{geboortedag}}$$

aantal aanvaarde koppelingen een negatieve S kregen. Het gaat om een familie die blijkbaar eerst door een ambtenaar met geheel verkeerde gegevens was ingeschreven en later opnieuw, met juiste gegevens, werd overgeschreven. De beslissing om mensen met afwijkende geboortedata, -plaatsen en namen toch te koppelen werd (manueel) genomen nadat hun gebeurtenissen en onderlinge kruisverwijzingen in het register in detail waren bekeken.

Ten slotte blijkt uit Tabel A.8 dat de selectieprocedure efficiënter werd naarmate het om recentere registers ging. Dit is een gevolg van de meer volledige identificatie van de recentere observaties. In BR1846 moesten 2.467 van de 2.699 (91%) kandidaat-koppels (manueel) verworpen worden; slechts 9% van de kandidaten werd daadwerkelijk gekoppeld in de beslissingsfase. Naarmate de bevolkingsregisters recenter worden, zijn observaties vollediger geïdentificeerd (zie Tabel A.1) en beschikt de selectieprocedure dus over meer identificatoren om in die fase al koppeling uit te sluiten. In BR1900 kon dan ook 68% van de kandidaten effectief gekoppeld worden.

Tabel A.9 geeft de verdelingen van koppelscores voor de interregisterkoppelingen. In grote lijnen hebben die hetzelfde verloop als de verdelingen bij de intraregisterkoppelingen. Een verschil is dat een beperkt aantal koppelingen bevestigd werd ondanks een score S lager dan tien. Excuses om toch te koppelen zijn onder meer expliciete naamsveranderingen (bijvoorbeeld ten gevolge van de wettiging van buitenechtelijke kinderen) en ambtelijke fouten.

A.3.2 Koppeling van huwelijkspartners

Voor de analyse van (huwelijks)vruchtbaarheid is het leggen van de juiste huwelijksbanden uiteraard cruciaal. De te volgen strategie verschilde van register tot register. Systematisch werd er enkel gezocht naar huwelijkspartners van mensen uit de drie steekproefgeneraties, indien nodig en mogelijk ook in de Leuvense huwelijksakten (indien bijvoorbeeld de datum van een in Leuven afgesloten huwelijk ontbrak).

Tabel A.9: Kandidaten voor interregisterkoppeling naar generatie en genomen beslissing (al dan niet gekoppeld): procentuele verdeling naar koppelscore S

GENERATIE 1830						
Koppelscore S <i>Gekoppeld?</i>	BR1846-1856		BR1856-1866		BR1846-BR1866	
	<i>Nee</i>	<i>Ja</i>	<i>Nee</i>	<i>Ja</i>	<i>Nee</i>	<i>Ja</i>
<0,00	0,3	0,0	14,1	0,0	0,4	0,0
0,00-4,99	6,0	0,2	19,2	0,0	15,5	0,0
5,00-9,99	19,0	0,7	29,8	1,4	21,0	0,8
10,00-14,99	45,5	2,9	20,5	2,7	39,6	3,5
15,00-19,99	21,5	13,7	8,1	6,0	19,5	9,6
20,00-24,99	7,1	19,7	7,3	7,7	3,3	23,5
25,00-29,99	0,2	16,8	0,0	16,4	0,0	11,1
30,00-34,99	0,0	21,4	1,0	26,3	0,0	9,9
35,00-39,99	0,5	24,7	0,0	39,3	0,0	41,5
<i>Totaal N =100%</i>	637	590	396	715	457	395
<i>Maximale S</i>		39,75		39,69		36,41
<i>Gemiddelde S</i>	12,36	27,00	11,13	32,94	11,09	28,89
GENERATIE 1850						
Koppelscore S <i>Gekoppeld?</i>	BR1866-1880		BR1880-1890		BR1866-1890	
	<i>Nee</i>	<i>Ja</i>	<i>Nee</i>	<i>Ja</i>	<i>Nee</i>	<i>Ja</i>
<0,00	2,4	0,0	2,0	0,0	1,6	0,0
0,00-4,99	38,0	0,5	26,9	0,0	41,0	0,0
5,00-9,99	12,7	0,2	17,8	0,2	13,2	0,2
10,00-14,99	25,5	0,0	34,3	1,7	27,1	0,7
15,00-19,99	9,0	1,7	7,4	3,3	8,6	0,7
20,00-24,99	8,8	5,9	8,1	8,5	6,0	4,3
25,00-29,99	1,8	10,5	2,0	10,7	1,6	12,3
30,00-34,99	1,8	18,0	1,3	11,4	0,7	12,6
35,00-39,99	0,0	63,2	0,0	64,1	0,2	69,2
<i>Totaal N =100%</i>	502	589	297	516	432	438
<i>Maximale S</i>		39,48		39,46		39,45
<i>Gemiddelde S</i>	9,47	35,26	10,45	34,64	9,03	36,01
GENERATIE 1864						
Koppelscore S <i>Gekoppeld?</i>	BR1880-1890		BR1890-1900		BR1880-1900	
	<i>Nee</i>	<i>Ja</i>	<i>Nee</i>	<i>Ja</i>	<i>Nee</i>	<i>Ja</i>
<0,00	1,9	0,0	0,0	0,0	0,7	0,0
0,00-4,99	37,8	0,2	26,4	0,2	38,4	0,0
5,00-9,99	15,2	0,3	23,1	0,2	21,2	0,0
10,00-14,99	25,7	1,6	31,7	0,6	30,1	1,0
15,00-19,99	9,4	1,6	13,5	1,4	6,8	1,5
20,00-24,99	4,6	4,2	3,8	2,6	1,4	4,5
25,00-29,99	1,7	8,5	0,0	7,8	0,3	8,7
30,00-34,99	1,7	9,2	1,4	13,9	0,7	11,9
35,00-39,99	2,1	74,5	0,0	73,4	0,3	72,4
<i>Totaal N =100%</i>	479	644	208	503	292	402
<i>Maximale S</i>		39,46		37,53		37,56
<i>Gemiddelde S</i>	9,91	36,07	10,21	34,91	8,47	34,64

Vanaf de bevolkingsregisters van 1866 schreven de Leuvense ambtenaren bij een huwelijk de partners plus de eventueel reeds aanwezige, inwonende kinderen over naar de folio van de gezinswoning. Als dat de (ouderlijke) woonplaats van één van de partners was, werd de andere er dus gewoon bijgeschreven met verwijzing naar het huwelijk. In principe staan zo alle gehuwde koppels op minstens één folio bij elkaar en zijn ze dus via kruiselingse verwijzingen overal in het register op te sporen. Aangezien die verwijzingen tijdens de dataverzameling mee gecodeerd zijn, kan dit grotendeels automatisch verlopen.

In de registers van 1846 en 1856 was dat echter nog niet het geval: huwelijk gaf meestal geen aanleiding tot het overschrijven van de partners of hun kinderen. Dat betekent dat als iemand van de 1830-generatie trouwde, de gegevens van de huwelijkspartner niet in de steekproef zaten tenzij die partner ook van 1830 was of in een huis gewoond had waar ook een 1830'er stond ingeschreven. De rest moest manueel worden opgezocht. Dit kostte bijzonder veel tijd maar het was voor dit onderzoek nodig en mogelijk doordat bij huwelijk het bevolkingsregister naast de naam ook het adres (in BR1846) of het folionummer (in BR1856) van de huwelijkspartner vermeldde. Dit kon uiteraard enkel als die partner in Leuven woonde of er onmiddellijk na het huwelijk kwam wonen.

Tabel A.10 geeft een overzicht van het aantal gekoppelde echtgenoten. Van 85 tot 95% van de huwelijken kon de huwelijkspartner in het bevolkingsregister van Leuven worden teruggevonden, behalve bij een beperkt aantal 'late' huwelijken van 1830'ers in BR1866. Wanneer de partner bij een huwelijk niet geïdentificeerd kon worden, woonde niet alleen die partner haast altijd buiten Leuven, maar emigreerde de referentiepersoon op papier vaak binnen één of twee maanden (in werkelijkheid vond die emigratie wellicht al vroeger plaats) weg uit Leuven. Van de gehuwden kon ruim negen op tien partners in het Leuvense bevolkingsregister worden geïdentificeerd. Personen die bij het inschrijven in het bevolkingsregister al gehuwd waren, stonden in principe op dezelfde bladzijde als de partner. Waar dat niet het geval was, was er meestal sprake van een feitelijke, eventueel seizoensgebonden scheiding.

Tabel A.10: Aantal geïdentificeerde huwelijksrelaties naar generatie en bevolkingsregister

GENERATIE	BEVOLKINGSREGISTER			
GENERATIE 1830	BR1846	BR1856	BR1866	TOTAAL
Aantal geregistreerde huwelijken	111	147	28	286
- waarvan partner geïdentificeerd	106 (95,5%)	141 (95,9%)	23 (82,1%)	270 (94,4%)
Aantal keer geregistreerd als 'gehuwd'	26	254	334	614
- waarvan partner geïdentificeerd	26 (100,0%)	236 (92,9%)	322 (96,4%)	584 (95,1%)
GENERATIE 1850	BR1866	BR1880	BR1890	TOTAAL
Aantal geregistreerde huwelijken	369	73	140	582
- waarvan partner geïdentificeerd	313 (84,8%)	64 (87,7%)	138 (98,6%)	515 (88,5%)
Aantal keer geregistreerd als 'gehuwd'	118	442	280	840
- waarvan partner geïdentificeerd	113 (95,8%)	428 (96,8%)	265 (94,6%)	806 (96,0%)
GENERATIE 1864	BR1880	BR1890	BR1900	TOTAAL
Aantal geregistreerde huwelijken	278	355	341	974
- waarvan partner geïdentificeerd	238 (85,6%)	341 (96,1%)	336 (98,5%)	915 (93,9%)
Aantal keer geregistreerd als 'gehuwd'	63	186	78	327
- waarvan partner geïdentificeerd	62 (98,4%)	178 (95,7%)	66 (84,6%)	306 (93,6%)

A.3.3 Koppeling ouders-kinderen

Absoluut essentieel voor dit onderzoek is dat de juiste kinderen aan de juiste ouders worden gekoppeld. Daarom was dit een bijzonder aandachtspunt bij de opleiding van de codeurs, betrokken bij de dataverzameling. Indien de vader in het huishouden aanwezig is, staat het kind in principe geïdentificeerd als kind van die vader. Alleen als enkel de moeder aanwezig is, wordt naar háár verwezen. In het eerste geval moet dus de relatie tussen moeder en kind afgeleid worden uit de huwelijksband met de vader. Aangezien er echter ook vaak kinderen uit vorige huwelijken aanwezig zijn, is die band niet altijd met zekerheid te leggen. In vele gevallen vermeldde het bevolkingsregister dat een bepaald kind bijvoorbeeld stamde "du premier mariage", maar lang niet systematisch. Daarom zijn bij de interpretatie en het toeschrijven (onder meer op basis van geboorteplaats) ongetwijfeld en onvermijdelijk fouten gemaakt. In de meeste gevallen is er echter nauwelijks twijfel mogelijk, vooral wanneer de huwelijksdatum én het geboortjaar van het kind bekend zijn.

Door het patronieme systeem van naamgeving scheidt het toeschrijven van het wettige vaderschap bij wettig geboren kinderen weinig problemen. Bij buitenechtelijk geboren kinderen werd de juridische procedure overgenomen: erkende kinderen en/of gelegitimeerde kinderen werden toegeschreven aan de juridische vader.

Ouderschapsbanden werden in twee richtingen gelegd: van referentiepersoon naar zijn of haar ouders, en van referentiepersoon naar zijn of haar kinderen. Tabel A.11 geeft een overzicht van het aantal individuele inschrijvingen van personen uit de drie referentiegeneraties naar gelang vader, moeder, beide of geen van beide ouders geïdentificeerd zijn. Zoals verwacht mag worden, zijn de ouders telkens vooral geïdentificeerd in het eerste van de drie opeenvol-

gende registers waarin de generaties gevolgd worden. Naarmate de cohorten ouder worden, vormen ze een eigen huishouden en sterven hun ouders. Wellicht vooral omwille van de hogere vrouwelijke levensverwachting, werden in elk register en voor elke generatie meer moeders dan vaders geïdentificeerd.

Tabel A.11: Identificatie van ouders naar generatie en bevolkingsregister

GENERATIE	BEVOLKINGSREGISTER					
GENERATIE 1830	BR1846		BR1856		BR1866	
Aantal individuele observaties	702	(100,0%)	806	(100,0%)	588	(100,0%)
- waarvan geïdentificeerd						
- moeder en vader	319	(45,4%)	87	(10,8%)	17	(2,9%)
- moeder wel, vader niet	107	(15,2%)	68	(8,4%)	24	(4,1%)
- moeder niet, vader wel	32	(4,6%)	30	(3,7%)	13	(2,2%)
- noch moeder noch vader	244	(34,8%)	621	(77,1%)	534	(90,8%)
GENERATIE 1850	BR1866		BR1880		BR1890	
Aantal individuele observaties	1655	(100,0%)	902	(100,0%)	662	(100,0%)
- waarvan geïdentificeerd						
- moeder en vader	597	(36,1%)	54	(6,0%)	11	(1,7%)
- moeder wel, vader niet	232	(14,0%)	64	(7,1%)	28	(4,2%)
- moeder niet, vader wel	100	(6,0%)	19	(2,1%)	6	(0,9%)
- noch moeder noch vader	726	(43,9%)	765	(84,8%)	617	(93,2%)
GENERATIE 1864	BR1880		BR1890		BR1900	
Aantal individuele observaties	1499	(100,0%)	1107	(100,0%)	670	(100,0%)
- waarvan geïdentificeerd						
- moeder en vader	523	(34,9%)	128	(11,6%)	13	(1,9%)
- moeder wel, vader niet	198	(13,2%)	129	(11,6%)	42	(6,3%)
- moeder niet, vader wel	80	(5,3%)	31	(2,8%)	3	(0,4%)
- noch moeder noch vader	698	(46,6%)	819	(74,0%)	612	(91,3%)

De identificatie van de ouders van de drie steekproefcohorten is vooral interessant omdat het iets kan leren over hun gezin van herkomst. De identificatie van de eigen kinderen van die cohorten is essentieel voor dit onderzoek omdat op die basis hun vruchtbaarheid gereconstrueerd wordt. Tabel A.12 geeft voor elke generatie het aantal gevonden kinderen naar bevolkingsregister. Zoals verwacht worden vooral kinderen gevonden in het tweede en derde register; gedurende de looptijd van het eerste register zijn de referentiepersonen tussen 16 en 26 of 30 jaar oud. De meeste kinderen werden pas na die leeftijd geboren.

Tabel A.12: Identificatie van ouders naar generatie en bevolkingsregister

GENERATIE	BEVOLKINGSREGISTER					
Generatie 1830	BR1846		BR1856		BR1866	
- kinderen / vrouwen	178 / 364	(49 / 100)	617 / 375	(165 / 100)	735 / 284	(259 / 100)
- kinderen / mannen	98 / 331	(30 / 100)	531 / 376	(141 / 100)	750 / 260	(288 / 100)
- totaal kinderen/G1830ers	276 / 695	(40 / 100)	1148 / 751	(153 / 100)	1485 / 544	(273 / 100)
Generatie 1850	BR1866		BR1880		BR1890	
- kinderen / vrouwen	672 / 611	(110 / 100)	1055 / 401	(263 / 100)	868 / 315	(276 / 100)
- kinderen / mannen	497 / 580	(86 / 100)	970 / 391	(248 / 100)	856 / 307	(279 / 100)
- totaal kinderen/G1850ers	1169 / 1191	(98 / 100)	2025 / 792	(256 / 100)	1724 / 622	(277 / 100)
Generatie 1864	BR1880		BR1890		BR1900	
- kinderen / vrouwen	418 / 599	(70 / 100)	906 / 456	(199 / 100)	796 / 325	(245 / 100)
- kinderen / mannen	220 / 575	(38 / 100)	829 / 444	(187 / 100)	802 / 298	(269 / 100)
- totaal kinderen/G1864ers	638 / 1174	(54 / 100)	1735 / 900	(193 / 100)	1598 / 623	(257 / 100)

Bron: TELKID10.prg

* * *

De koppeling van individuele persoonsgegevens binnen en tussen de bevolkingsregisters, en de koppelingen tussen echtgenoten en tussen ouders en kinderen, levert een steekproef van onderling verbonden, socio-demografische levensloopfragmenten op. Bijlage B schetst de belangrijkste kenmerken van de steekproeven voor de drie generaties en onderzoekt de betrouwbaarheid van de registratie. Daarbij gaan we er van uit dat de gemaakte koppelingen correct zijn en er geen nominale koppelingen ten onrechte niet zijn gelegd.