De staat van het Nederlands

Overzicht van de gegevensverwerking

Aantal deelnemers

De resultaten in dit document zijn gebaseerd op de definitieve data van 15 september 2016. Op dat moment hadden er 7406 mensen meegedaan, 3037 mensen in Nederland, 4233 mensen in het Vlaams Gewest en 136 mensen in het Brussels Hoofdstedelijk Gewest.

Opdeling

De data is gesplitst in drie groepen, waarbij elk van de groepen afzonderlijk verwerkt wordt. De drie groepen zijn: Nederland, Vlaams Gewest, Brussels Hoofdstedelijk Gewest. Het Brussels Hoofdstedelijk Gewest kan moeilijk samengenomen worden met het Vlaams Gewest. Op Wikipedia wordt van het Brussels Hoofdstedelijk Gewest gezegd:

"Het gewest heeft het Nederlands en het Frans als officiële talen... Toch is de voertaal op straat veelal het Frans, in overeenstemming met de indeling van de bevolking: 80 tot 90 procent hanteert het Frans als voertaal, 20 tot 10 procent Nederlands, afhankelijk van de bron en de gebruikte maatstaven."

Deelnemers die buiten Nederland, Vlaanderen Brussel wonen, worden uit de data weggefilterd. Dat betekent dat ook deelnemers uit het Waals gewest (inclusief de gemeenten van de Duitstalige Gemeenschap) niet worden verwerkt.

Te stratificeren variabelen

Stratificeren is het opsplitsen van de populatie in verschillende groepen of lagen (stratum = laag) op basis van criteria die voor het onderzoek relevant zijn. Bij stratificatie verdeelt men een onderzoekspopulatie in één of meerdere subcategorieën volgens bepaalde criteria, zoals leeftijd, geslacht, sociale status, etc. Zo ontstaat een gestratificeerde (gelede) steekproef.

Stratificatie is in ons geval noodzakelijk omdat de enquête op basis van vrijwilligheid ingevuld wordt, en dat kan een vertekening veroorzaken. Voor het stratificeren heb ik de variabelen in onderdeel 1 (Uw persoonlijke achtergrond) uit de enquête overwogen, namelijk:

Nederlandse of Vlaamse vragenlijst (A0)
Geslacht (B1)
Geboortejaar (B2)
Geboorteland (B4)
Postcode (B13)
Geloofsovertuiging (B5)
Onderwijs dat eventueel gevolgd wordt (B11a)
Hoogst behaalde onderwijsdiploma (B7NL/B7B)
Sector waarin men werkt (B6)

Wel of geen schoolgaande kinderen (B12)
Moedertaal (B10, B10a)

Het is belangrijk om voor zoveel mogelijk variabelen gelijktijdig te stratificeren. Stel we willen stratificeren voor een variabele A (met groepen a1 en a2) en B (met groepen b1 en b2). Als we beide variabelen combineren, krijgen we vier subgroepen, namelijk:

a1.b1
a1.b2
a2.b1
a2.b2

Als we weten dat in de populatie 50% van alle voorkomens behoort tot subgroep a1.b1, 25% tot subgroep a1.b2, 20% tot subgroep a2.b1 en 5% tot subgroep a2.b2, zullen deze proporties ook teruggevonden moeten worden in een steekproef die uit die populatie getrokken is.

Om te kunnen stratificeren heb ik gebruik gemaakt van gegevens van het CBS (voor Nederland) en AD Statistiek (Vlaanderen en Brussel). CBS en AD Statistiek geven voor een aantal van de variabelen in onderdeel 1 de aantallen voor de complete populaties (in Nederland in totaal 16.900.726 inwoners, in Vlaanderen in totaal 6.325.740 inwoners, in Brussel in totaal 1.136.778 inwoners).

Omdat de variabelen in combinatie met elkaar bekeken moeten worden, heb ik op de sites van het CBS en AD Statistiek steeds gezocht naar een tabel waarin een opdeling gemaakt werd volgens zoveel mogelijk van bovengenoemde variabelen tegelijkertijd. Het maximaal mogelijke voor beide landen was een gelijktijdige opdeling in geslacht, leeftijdsgroep, geboorteland en provincie. Uiteraard geldt voor Brussel geen opdeling in provincie.

Geslacht
In de enquête zijn er drie mogelijkheden: 'man', 'vrouw', 'ik omschrijf mezelf anders'. Zowel het CBS als AD Statistiek kennen slechts twee mogelijkheden: 'man' en 'vrouw'. Daardoor is het niet mogelijk om te straticiferen over 'ik omschrijf mezelf anders'. Ik heb deze groep daarom weggelaten. In de enquête-resultaten van 15 september zijn er voor Nederland 9 deelnemers die dit opgaven, voor Vlaanderen zijn dit er 5 en voor Brussel niemand.

Leeftijdsgroep
Geboren in 2015 heb ik geïnterpreteerd als: geboren op 1 januari 2015. Dus: iemand geboren in 2015 is 1 jaar oud op 1 januari 2016. Alle participanten geboren in 2002 of later zijn uit de data verwijderd. Deze groep is niet bruikbaar vanwege een te sterke ondervertegenwoordiging. Voor Nederland waren dit er 3 mensen, voor Vlaanderen 7 mensen en voor Brussel niemand. Als gevolg daarvan worden geen resultaten gegeven van leerlingen uit basisonderwijs en het middelbaar onderwijs (maar wel voor mbo, hbo, universiteit en cursusonderwijs).

Het CBS geeft een verdeling in vijf groepen, en AD Statistiek geeft een verdeling in 20 groepen. De verdeling volgens AD Statistiek is aangepast aan die van het CBS (ondersom zou niet mogelijk zijn). De groepsverdeling is: 0 tot 15 jaar (maar deze groep is weggelaten), 15 tot 30 jaar, 30 tot 45 jaar, 45 tot 65 jaar, 65 jaar of ouder.

De metingen van het CBS zijn per 1 januari 2015, en die van ADS Statistiek per 1 januari 2016. Bij de verwerking van de geboortejaren in de enquête heb ik hier rekening mee gehouden. De Nederlandse geboortejaren worden als volgt omgerekend:

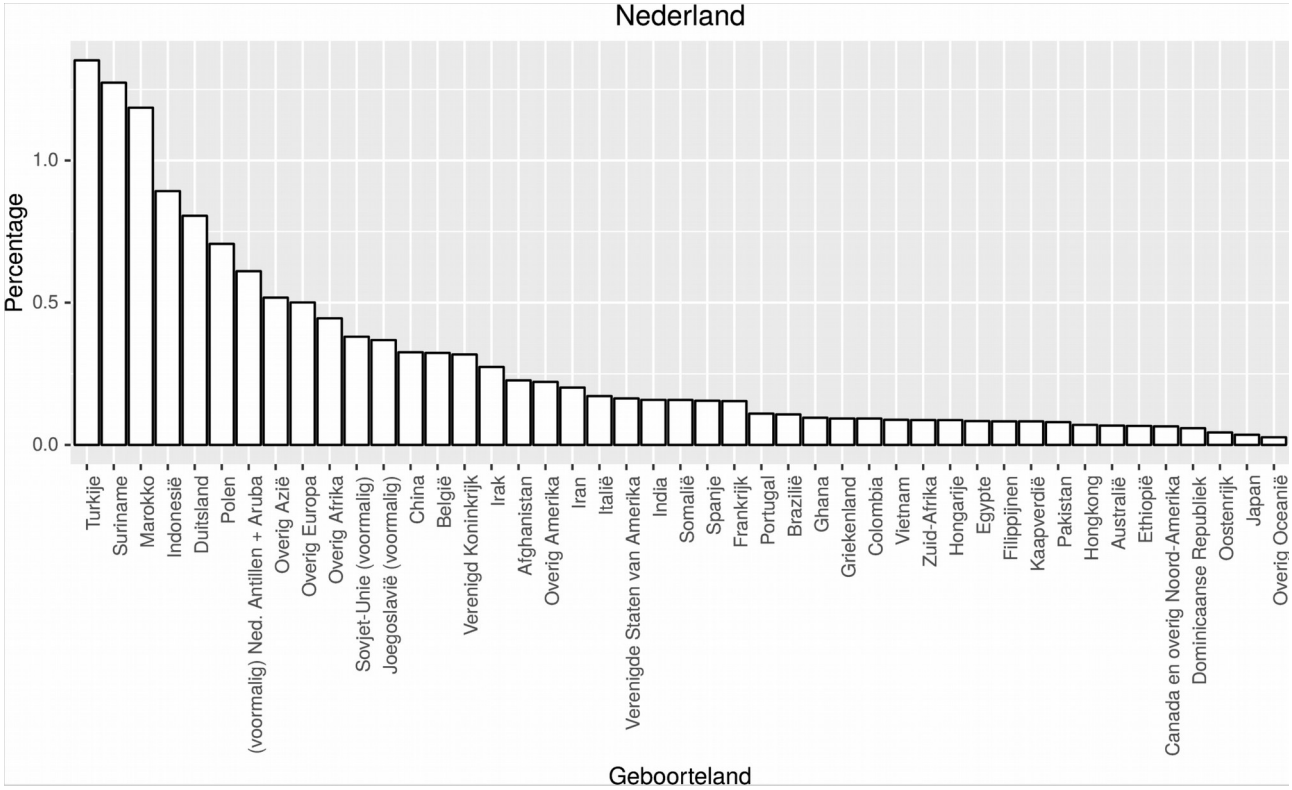
```
2015 - geboortejaar,
```

en voor België:

```
2016 - geboortejaar
```

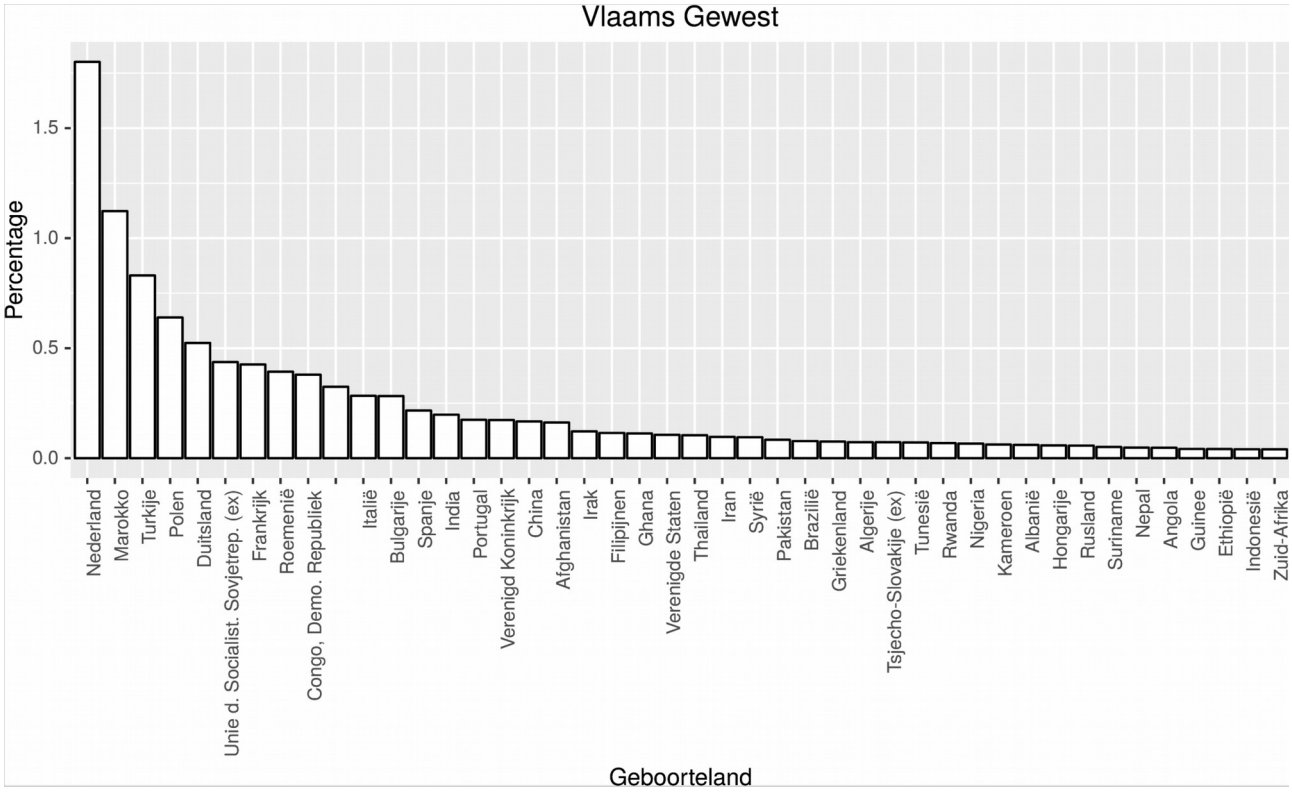
waarna de deelnemers geklasseerd worden volgens bovengenoemde leeftijdsgroepen.

Geboorteland
Voor Nederland ziet de verdeling van inwoners die niet in Nederland geboren zijn volgens het CBS er uit zoals gegeven in Figuur 1. Op basis van deze grafiek is voor 'geboorteland' een verdeling in zes groepen gemaakt: Nederland, Turkije, Suriname, Marokko, Indonesië, overig.



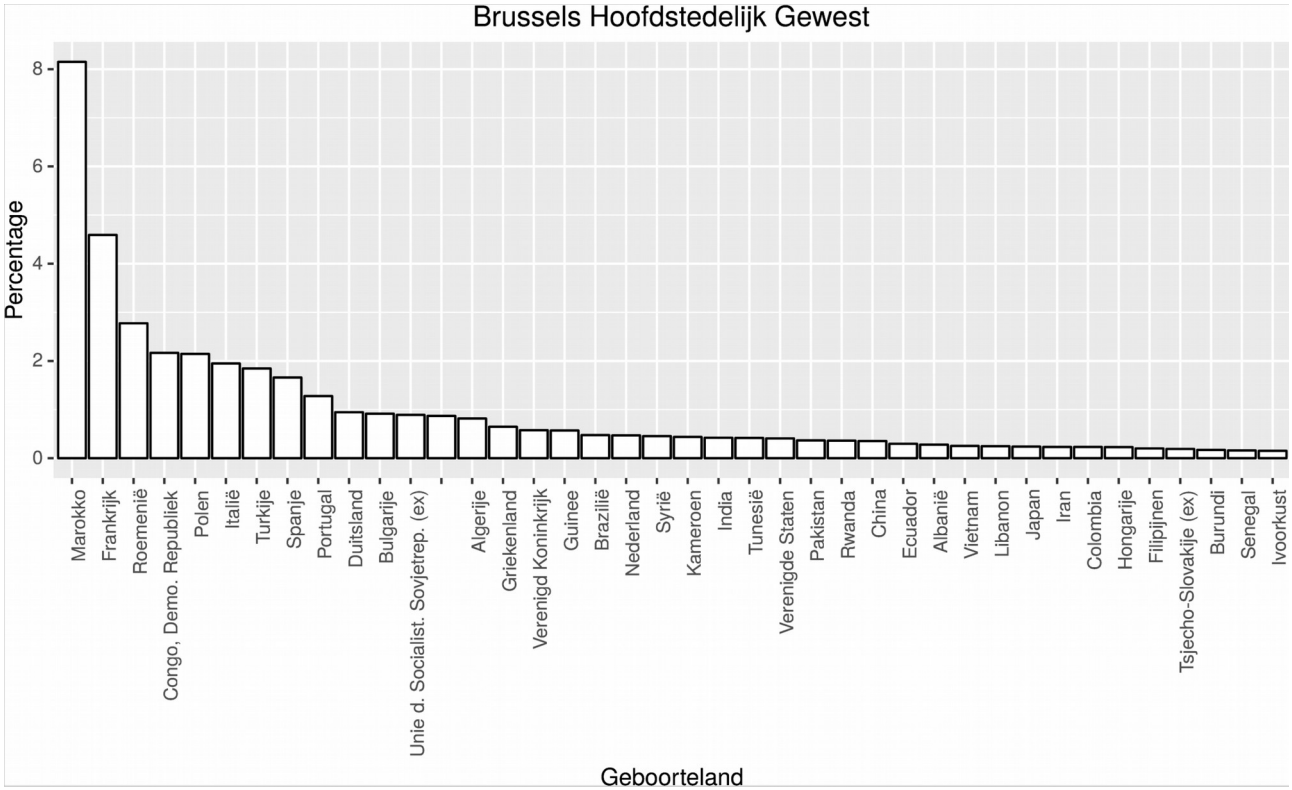
Figuur 1. Verdeling inwoners in Nederland naar geboorteland (exclusief Nederland) volgens de CBS-gegevens.

Voor Vlaanderen ziet de verdeling van inwoners die niet in België geboren zijn volgens AD Statistiek er uit zoals gegeven in Figuur 2. Op basis van deze grafiek is voor 'geboorteland' een verdeling in zes groepen gemaakt: België, Marokko, Nederland, Turkije, Polen, overig.



Figuur 2. Verdeling inwoners in Vlaanderen naar geboorteland (exclusief België) volgens de AD Statistiek-gegevens.

Voor Brussel ziet de verdeling van inwoners die niet in België geboren zijn volgens AD Statistiek er uit zoals gegeven in Figuur 3. Op basis van deze grafiek is voor 'geboorteland' een verdeling in zes groepen gemaakt: België, Marokko, Frankrijk, Roemenië, Democratische Republiek Congo, overig.



Figuur 3. Verdeling inwoners in Brussel naar geboorteland (exclusief België) volgens de AD Statistiek-gegevens.

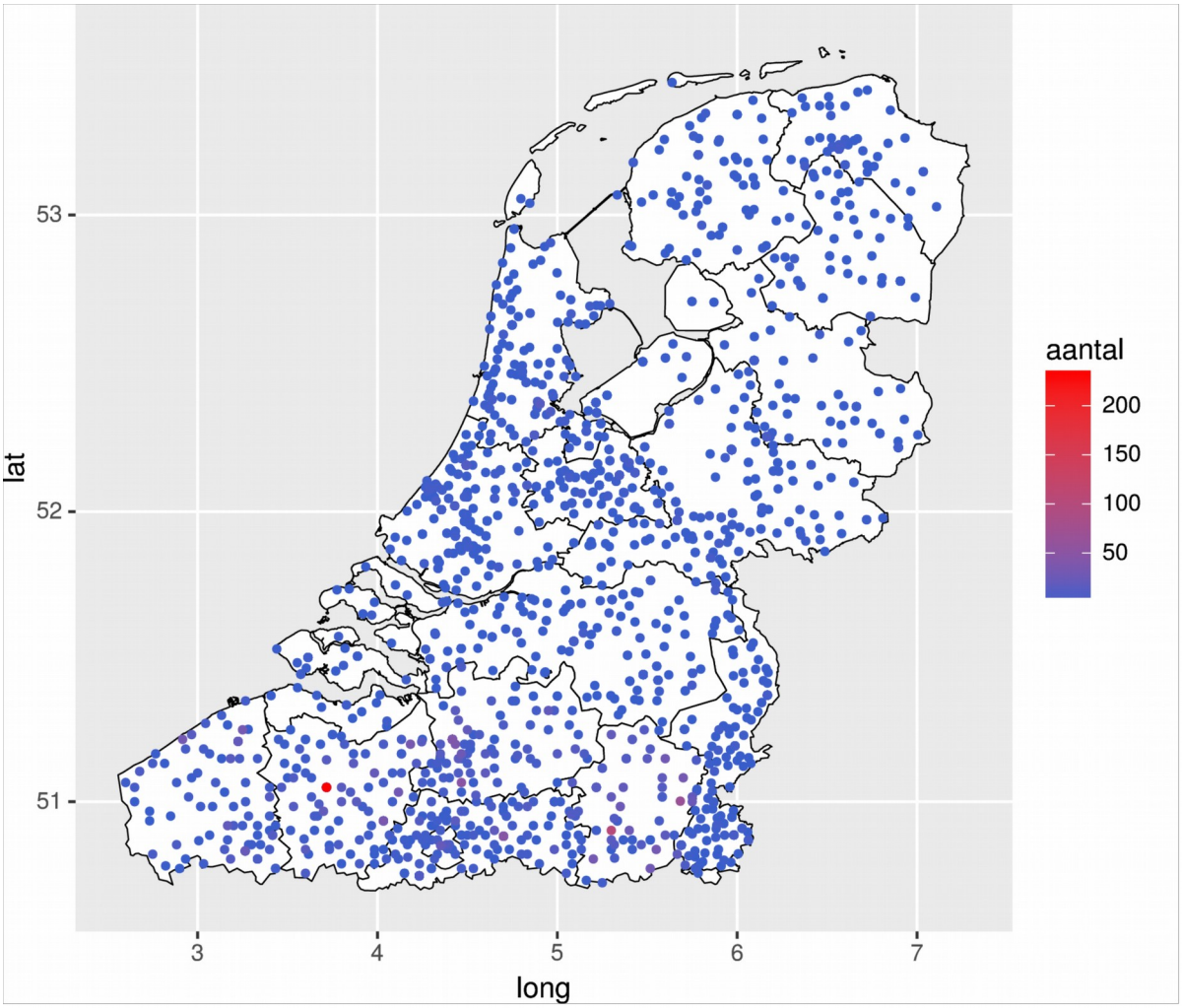
Provincie
Provincies worden bepaald aan de hand van postcodes. Deelnemers met ongeldige postcodes worden uit de data weggelaten. Onder 'geldig' wordt verstaan dat:

- een deelnemer voor de Nederlandse enquête een postcode opgeeft die verwijst naar een woonadres in Nederland;

- een deelnemer voor de Vlaamse enquête een postcode opgeeft die verwijst naar een woonadres in het Vlaams Gewest of het Brussels Hoofdstedelijk Gewest.

Dit betekent ook dat deelnemers met postcodes bestaande uit minder of meer dan vier cijfers niet verwerkt worden.

De geografische spreiding van de deelnemers aan de enquête wordt weergegeven in Figuur 4.



Figuur 4. Geografische spreiding van de deelnemers aan de enquête. Elk stip representeert een plaats. Hoe roder de plaats, hoe meer inwoners in die plaats hebben meegedaan.

Verdeling

Uitgaande van de vier variabelen *geslacht*, *leeftijdsgroep*, *geboorteland* en *provincie* krijgen we voor Nederland:

```
2 geslachten * 4 leeftijdsgroepen * 6 geboortelanden * 12 provincies = 576 subgroepen
```

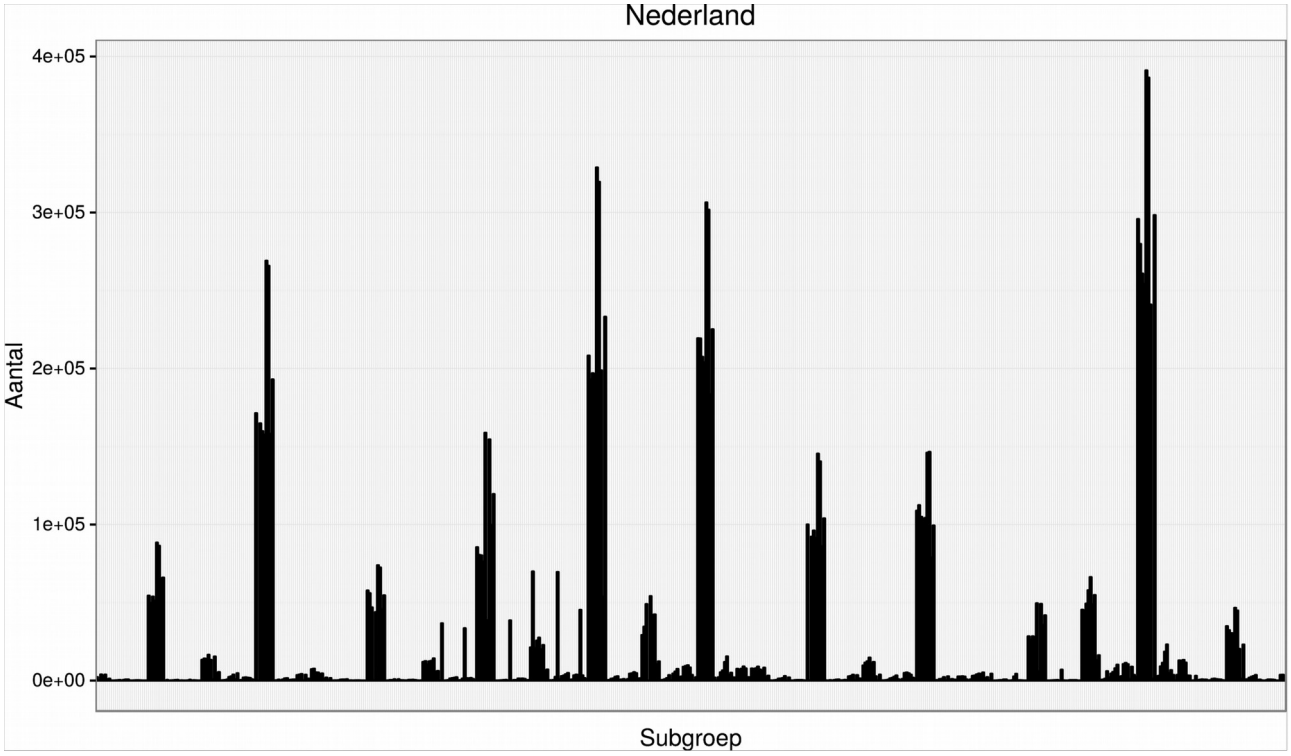
Voor Vlaanderen krijgen we:

```
2 geslachten * 4 leeftijdsgroepen * 6 geboortelanden * 5 provincies = 240 subgroepen
```

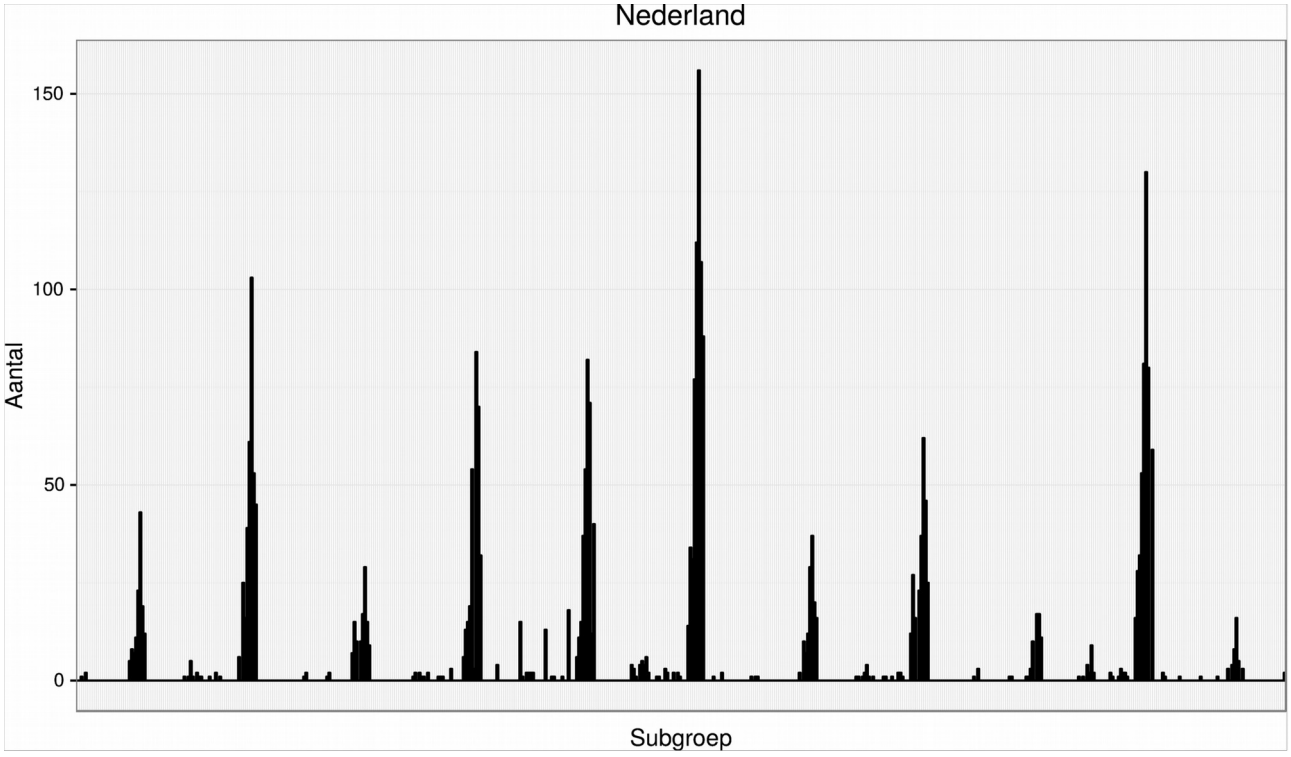
En voor Brussel:

```
2 geslachten * 4 leeftijdsgroepen * 6 geboortelanden * 1 provincie = 48 subgroepen
```

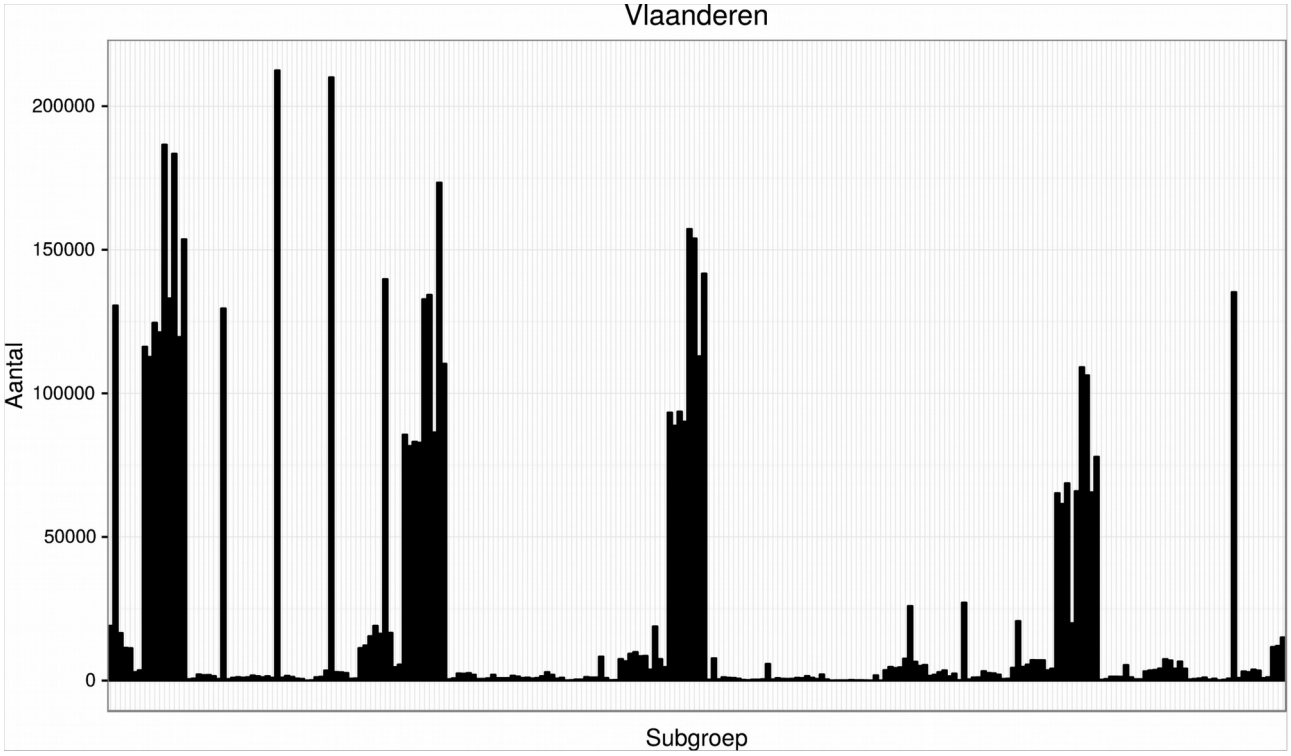
De verdeling van de aantallen volgens het CBS en AD Statistiek worden hieronder getoond in de Figuren 4a, 5a en 6a. De verdeling van de aantallen volgens de enquete worden getoond in de Figuren 4b, 5b and 6b.



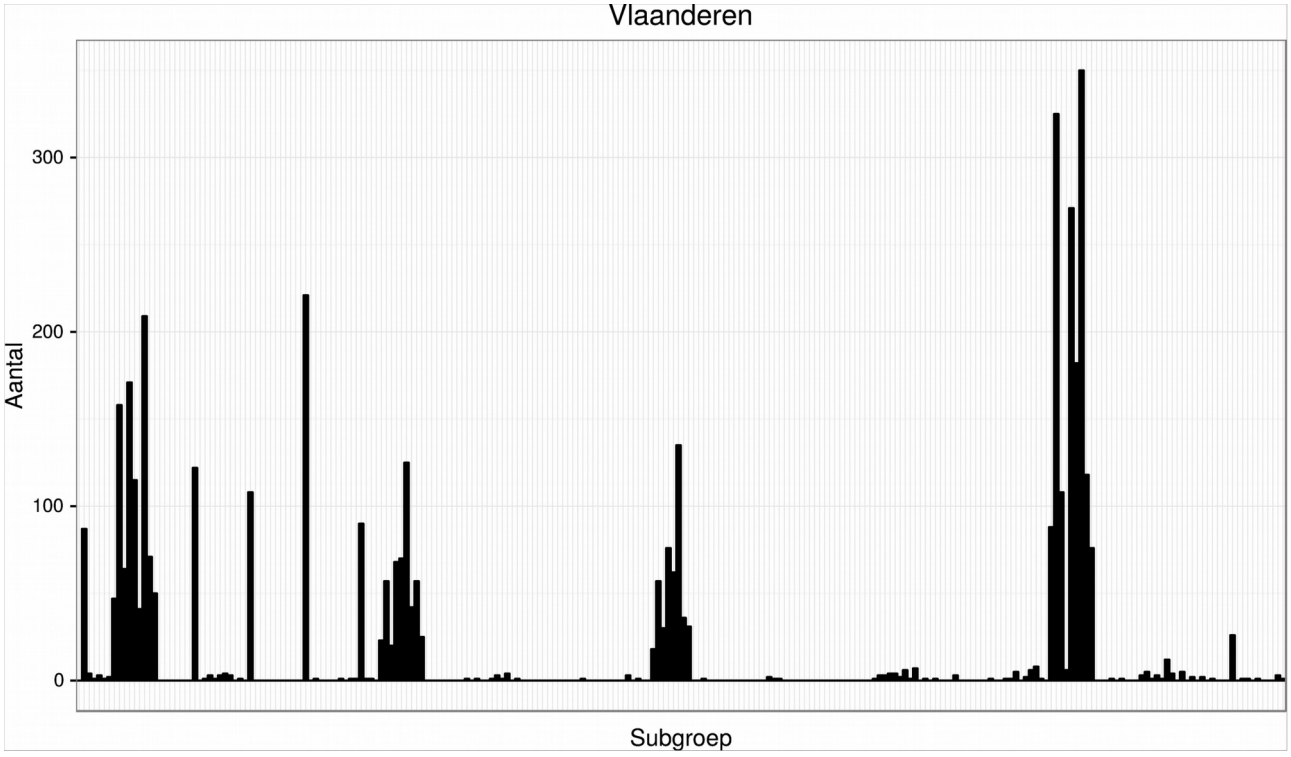
Figuur 4a. Verdeling van de 576 groepen voor Nederland volgens het CBS.



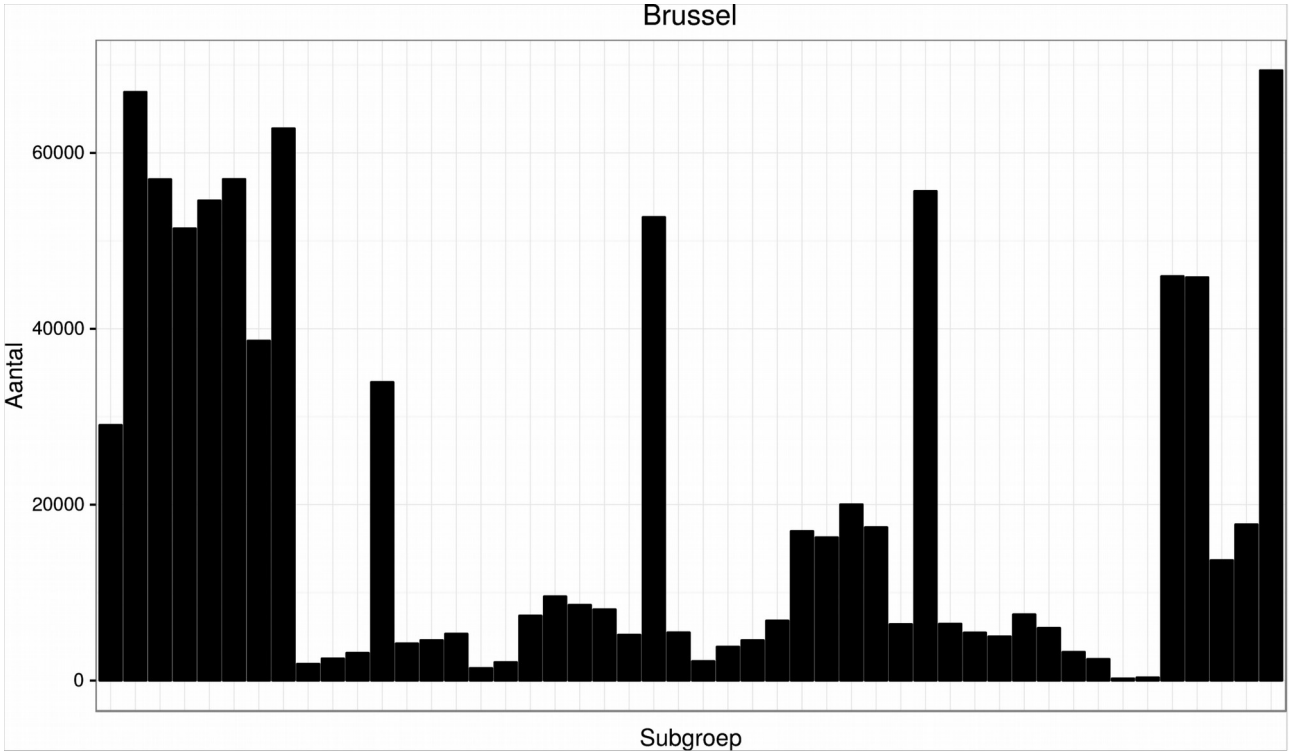
Figuur 4b. Verdeling van de 576 groepen voor Nederland volgens de enquête.



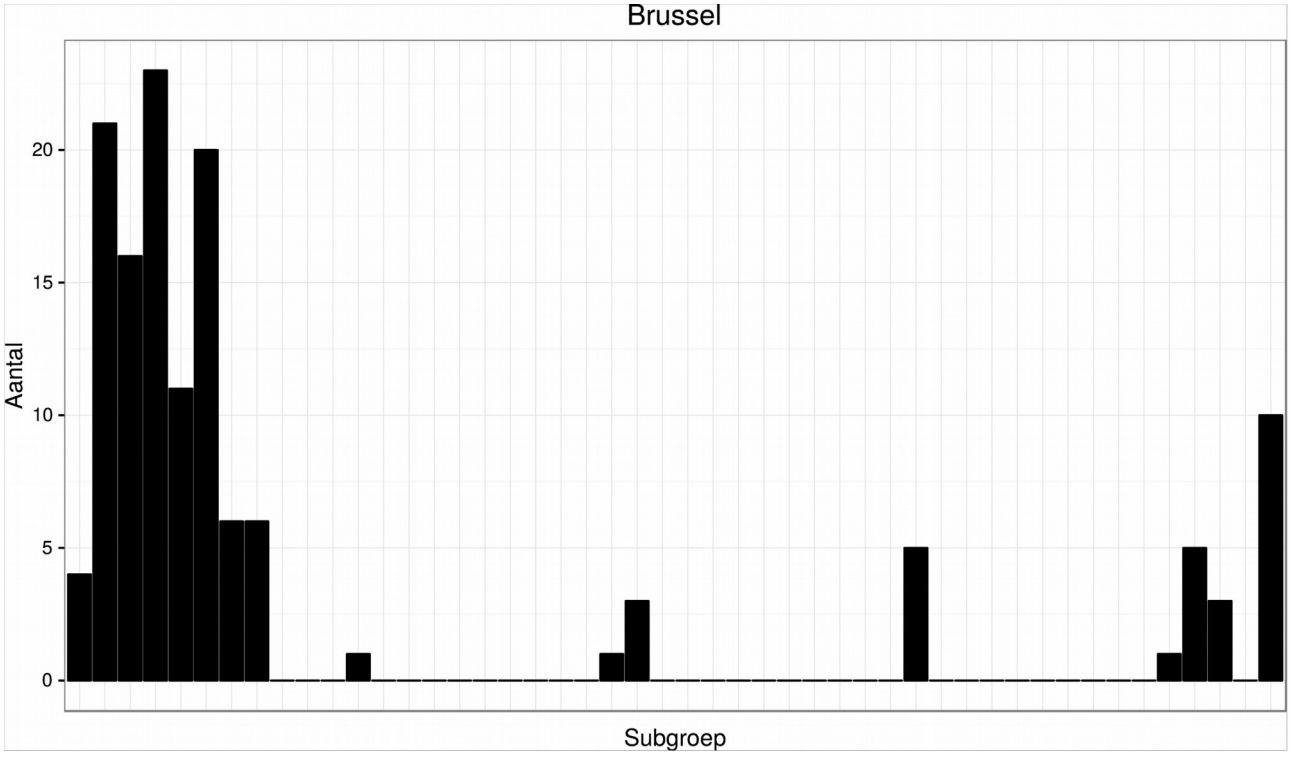
Figuur 5a. Verdeling van de 240 groepen voor Vlaanderen volgens AD Statistiek.



Figuur 5b. Verdeling van de 240 groepen voor Vlaanderen volgens de enquête.



Figuur 6a. Verdeling van de 48 groepen voor Brussel volgens AD Statistiek.



Figuur 6b. Verdeling van de 48 groepen voor Brussel volgens de enquête.

Stratificatie

De correlaties tussen de verdelingen volgens CBS/AD Statistiek en die volgens de enquête zijn:

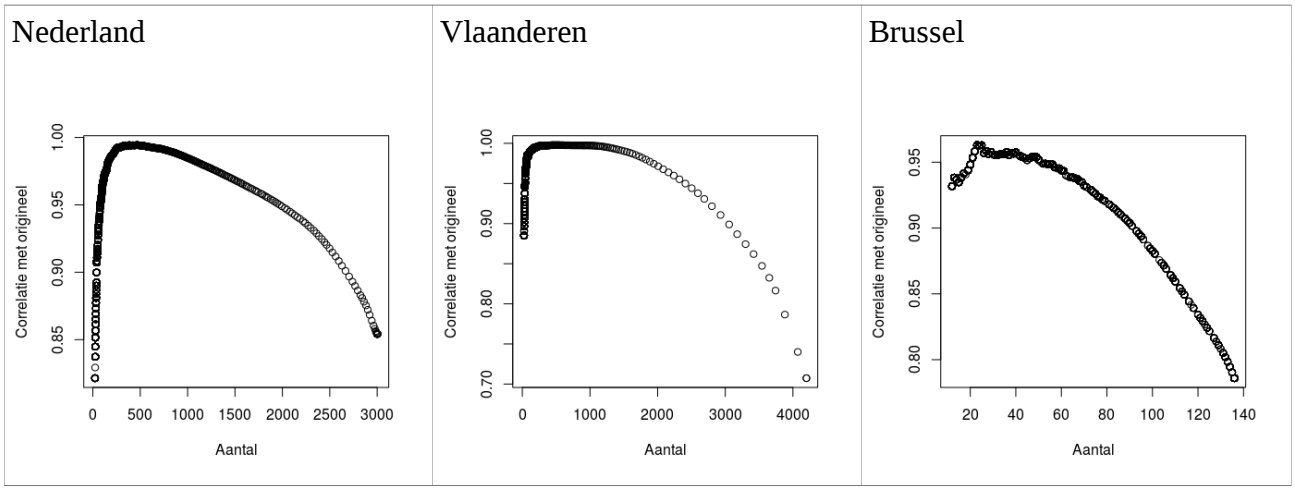
	correlatie *r*	aantal deelnemers
Nederland	0.854	3003
Vlaanderen	0.707	4200
Brussel	0.786	136

Ik heb een procedure ontwikkeld die de verdelingen volgens de enquête meer gelijkend op de verdelingen volgens CBS/AD Statistiek te maken. De procedure kan dat echter alleen maar doen door deelnemers in overtegenwoordigde groepen weg te laten. Deelnemers toevoegen in ondervertegenwoordige groepen kan de procedure uiteraard niet. Wanneer we de verdelingen volgens onze enquête zodanig willen aanpassen dat ze zo goed mogelijk correleren met de verdelingen volgens het CBS/AD Statistiek, krijgen we de volgende correlaties met de aantallen deelnemers:

	correlatie *r*	aantal deelnemers
Nederland	0.994	470
Vlaanderen	0.998	454
Brussel	0.963	23

We zien dat de correlaties bijna gelijk zijn aan 1, maar een correlatie van $r=1$ blijkt voor geen van de drie landen/gewesten mogelijk. Wat ook opvalt is dat het aantal deelnemers dat na deze strenge stratificatie nog overgebleven is, dramatisch kleiner is geworden.

Er geldt: hoe nauwkeuriger we stratificeren (en dus hoe hoger de verdeling van onze data correleert met die van het CBS/ADS), hoe meer deelnemers we verliezen. Dat zien we in Figuur 7.



Figuur 7. De correlatie tussen de verdelingen volgens de enquête en volgens CBS/AD Statistiek is uitgezet tegen het aantal deelnemers. De hoogste correlatie wordt bereikt voor 470 (Nederland), 454 (Vlaanderen) en 23 (Brussel) deelnemers.

Aanvankelijk stijgt de correlatie met het stijgen van het aantal deelnemers totdat de hoogste correlatie is bereikt. Vervolgens blijft de correlatie voor Vlaanderen (en in mindere mate voor Nederland) nog even constant. Daarna daalt de correlatie bij het stijgen van het aantal deelnemers.

Het verliezen van veel deelnemers is niet goed, omdat er dan veel subgroepen zijn die geen of weinig deelnemers bevatten. Het is daarom beter om een balans te vinden in het aantal deelnemers enerzijds en de nauwkeurigheid anderzijds. Of anders gezegd: we moeten enerzijds een acceptabele correlatie bereiken en anderzijds zoveel mogelijk het verlies van deelnemers beperken. Er is mij niet bekend of voor nauwkeurigheid een soort van minimum waarde is vastgesteld. Maar als we uitgaan van een minimum correlatie van r=0.85, maken we een veilige keuze. Dan is R^2 nog altijd 0.7225 of 72.25%. We krijgen dan de volgende resultaten:

	correlatie *r*	aantal deelnemers
Nederland	0.854	3003
Vlaanderen	0.862	3419
Brussel	0.852	113

Voor Nederland blijven alle deelnemers behouden. Voor Vlaanderen blijft 81% van de deelnemers behouden, en voor Brussel is dit 83%.

Met deze stratificatie hebben we m.i. voldoende recht om te stellen dat onze enquêtegegevens de Brusselse, Vlaamse en Nederlandse samenlevingen vertegenwoordigen. Echter met één kanttekening. Omdat de enquête in het Nederlands was, ontstaat automatisch een soort filtering: we hebben uitsluitend deelnemers die het Nederlands in ieder geval zodanig beheersen dat ze een enquête in het Nederlands kunnen invullen. De verdelingen van het CBS/ADS omvatten echter *iedereen*, dus ook mensen die helemaal geen Nederlands begrijpen. Tegen die verdelingen zijn onze enquêtegegevens gestratificeerd. Naar ik vermoed is dit voor Nederland het minst problematisch, en voor Brussel veruit het meest problematisch. Ik denk echter dat we er niets aan kunnen doen, dit is het best mogelijke.

Ondervertegenwoordigde groepen

Bijgevoegd bij dit document is een Excel-document: `ondervertegenwoordigd.xls`. Dit document bevat drie tabbladen, één voor Nederland, één voor Vlaanderen en één voor Brussel. Elk tabblad geeft voor het betreffende land of gewest een overzicht van de groepen die in de stratificatie ondervertegenwoordigd zijn.

Van links naar rechts vinden we in iedere tabel eerst kolommen voor *Geslacht*, *Leeftijdsgroep*, *Geboorteland* en *Provincie*, de vier variabelen die voor de stratificatie gebruikt werden.

Daarna volgt een kolom *Aantal* die het aantal mensen geeft volgens het CBS of AD Statistiek. De kolom *Enquête* geeft de aantallen volgens onze enquête.

De kolom *Optimaal* geeft het aantal mensen dat we voor iedere groep zouden moeten hebben om een optimale stratificatie te krijgen, d.w.z. een stratificatie waarbij de aantallen perfect correleren met de aantallen volgens het CBS of AD Statistiek, terwijl we tegelijk zo dicht mogelijk bij de oorspronkelijke aantallen proberen te blijven, d.w.z. de aantallen in de kolom Enquête. De correlaties zijn r=0.9997266 (Nederland), r=0.9999214 (Vlaanderen) en r=0.995774 (Brussel).

De kolom *Stratificatie* geeft de aantallen volgens de daadwerkelijke stratificatie. De kolom *Toevoegen* geeft de aantallen die aan de aantallen in de kolom Stratificatie toevoegd zouden moeten worden om te komen tot de optimale stratificatie zoals gegeven in de kolom *Optimaal*.

Welke talen worden hoevaak gesproken?

In de enquête werd vaak gevraagd naar het gebruik van het Nederlands, waarbij er dan zes keuzemogelijkheden waren: 'Altijd', 'Vaak', 'Soms', 'Zelden', 'Nooit', 'Niet van toepassing'. Als niet voor alle vragen in een blok 'Altijd' gekozen werd, werd gevraagd welke taal men dan nog meer sprak. Men kon een taal selecteren uit een lijst, of men kon kiezen voor 'Een andere taal' of 'Meerdere talen', waarna die ingetypt konden worden.

Enerzijds wordt dus gevraagd naar het gebruik van het Nederlands, en anderzijds wordt gevraagd wat men eventueel naast of in plaats van het Nederlands spreekt. Het mooiste is als men in één grafiek zowel de mate waarin Nederlands gesproken wordt als ook de mate waarin andere talen gesproken worden kan zien. Ik heb daarom de vraag over het gebruik van het Nederlands en de vragen over het gebruik van een andere taal of meerdere talen als volgt in elkaar geschoven:

- Als iemand geen enkele taal opgaf die hij/zij naast of in plaats van het Nederlands spreekt, wordt ervan uitgegaan dat de deelnemer 'altijd Nederlands' spreekt.

- Als iemand opgaf 'Altijd', 'Vaak', 'Soms' of 'Zelden' Nederlands te spreken, en daarnaast een andere taal of een combinatie van andere talen heeft opgegeven (bijvoorbeeld 'Duits en Frans'), is dit in de grafiek/tabel terug te vinden als `ook ...' (in ons voorbeeld: 'ook Duits en Frans'). Een label in een grafiek/tabel dat begint met 'ook' betekent dus: naast het Nederlands wordt óók ... gesproken.

- Als iemand opgaf 'Nooit' Nederlands te spreken, en vervolgens een andere taal of een combinatie van andere talen heeft opgegeven (bijvoorbeeld `Duits en Frans'), is dit in de grafiek/tabel terug te vinden als 'altijd ...' (in ons voorbeeld: 'altijd Duits en Frans').

- Als bij de vraag naar het gebruik van het Nederlands 'Niet van toepassing' was gekozen, wordt de opgave in het geheel niet in de grafiek/tabel verwerkt.

In een ander type vraag waarin ook naar het gebruik van het Nederlands gevraagd wordt, zijn er vijf keuzemogelijkheden, namelijk: 'Uitsluitend Nederlands', 'Nederlands en een of meerdere andere talen', 'Uitsluitend een of meerdere andere talen dan het Nederlands', 'Dat weet ik niet' en 'Niet van toepassing'. Als niet voor alle vragen in een blok 'Uitsluitend Nederlands' gekozen werd, kon men vervolgens ook hier een taal selecteren uit een lijst, of kiezen voor 'Een andere taal' of 'Meerdere talen', waarna die ingetypt konden worden.

De vraag over het gebruik van het Nederlands en de vragen over het gebruik van een andere taal of meerdere talen zijn op vergelijkbare wijze in elkaar geschoven als hierboven beschreven:

- Als iemand geen enkele taal opgaf die hij/zij naast of in plaats van het Nederlands spreekt, wordt ervan uitgegaan dat sprake is van 'Uitsluitend Nederlands'.

- Als iemand opgaf 'Nederlands en een of meerdere andere talen' te spreken, is dit in de grafiek/tabel terug te vinden als 'ook ...'.

- Als bij de eerste vraag naar het gebruik van het Nederlands 'Niet van toepassing' of 'Dat weet ik niet' gekozen was, wordt de opgave in het geheel niet in de grafiek/tabel verwerkt.

Fries/Frysk

Sommige sprekers van het Fries gaven 'Fries' op, en anderen 'Frysk'. 'Frsyk' is vervangen door 'Fries' zodat die twee in de verdere verwerking niet onderscheiden worden.

Dialect

Onder onderdeel 1 wordt onder andere gevraagd of de moedertaal van de deelnemers Nederlands is (B10). Als dat niet het geval is wordt gevraagd wat de moedertaal dan wel is (B10a) en hoe goed de deelnemer het Nederlands beheerst (B10b). Soms wordt bij B10a een Nederlands dialect opgegeven. Omdat we echter Nederlandse dialecten ook beschouwen als 'Nederlands', heb ik de opgaven bij B10a semi-automatisch gecontroleerd en gecorrigeerd. Als een Nederlands dialect wordt opgegeven, wordt de opgave bij zowel B10a als B10b verwijderd, en wordt 'Nee' onder B10 gewijzigd in 'Ja'.

Aan het begin van onderdeel 2 leest de deelnemer onder andere: "*Nederlands mag u in deze enquête opvatten als een paraplubegrip dat alle mogelijke taalvarianten omvat – van dialect tot standaardtaal, van informeel tot formeel Nederlands."* In de onderdelen 2 en later wordt vaak gevraagd of de deelnemers naast of in plaats van het Nederlands 'een andere taal' en 'meerdere talen' spreekt. Onder 'een andere taal' en onder 'meerdere talen' worden -- ondanks de melding aan het begin van onderdeel 2 -- veelvuldig dialecten opgegeven. Ik heb daarom een correctie in de data doorgevoerd, zodanig dat opgaven van Nederlandse dialecten verwerkt worden als 'Nederlands' indien uitsluitend een Nederlands dialect is opgegeven. Wanneer een dialect in combinatie met één of meer talen is gegeven, is niet ingegrepen in de data. Een combinatie wordt herkend door een komma, een slash '/', het woordje 'en' of het woordje 'of' in de opgave.

Er zijn 166 variabelen die opgeschoond werden, namelijk 83 waar gevraagd wordt naar 'een andere taal', en 83 waar gevraagd wordt naar 'meerdere talen'. Op basis van de 166 variabelen heb ik een inventarisatie gemaakt, en daarin de keywords gevonden die verraden dat de deelnemer een dialect heeft opgegeven. Heeft een deelnemer een dialect opgegeven, dan heb ik de opgave gewist. Als er dan verder geen taal of talen opgegeven zijn die naast het Nederlands gesproken worden, wordt bij de vraag naar het gebruik van het Nederlands 'Altijd' ingevuld.

Maar ik weet niet welke

Bij onderdeel 7 wordt gevraagd naar de talen die deelnemers om zich heen gesproken horen worden. Als de deelnemer gevraagd wordt de talen op te geven die hij en zij naast of in plaats van het Nederlands nog meer hoort, zijn onder meer de volgende twee opties mogelijk:

Een andere taal, maar ik weet niet welke
Meerdere talen, maar ik weet niet welke

Deze labels zijn vrij lang en passen daardoor niet goed in de grafieken en tabellen. Wanneer in de grafieken dan alleen de eerste 17 karakters weergegeven worden, zien we alleen het onderstreepte gedeelte, wat verwarrend zou zijn. Om die reden zijn ze gewijzigd in:

Een onbekende taal
Onbekende talen

Resultaten

De resultaten zijn te vinden in twee documenten, namelijk `grafieken.pdf` en `tabellen.pdf`. Beide documenten bevatten precies dezelfde resultaten, er is alleen verschil in de manier van weergave. `Grafieken.pdf` bevat grafieken, en `tabellen.pdf` bevat de tabellen die met de grafieken in `grafieken.pdf` corresponderen.

De resultaten vormen de uitwerkingen van de 'Vragen te beantwoorden' die opgesteld zijn door Marten en aangevuld door Fieke.