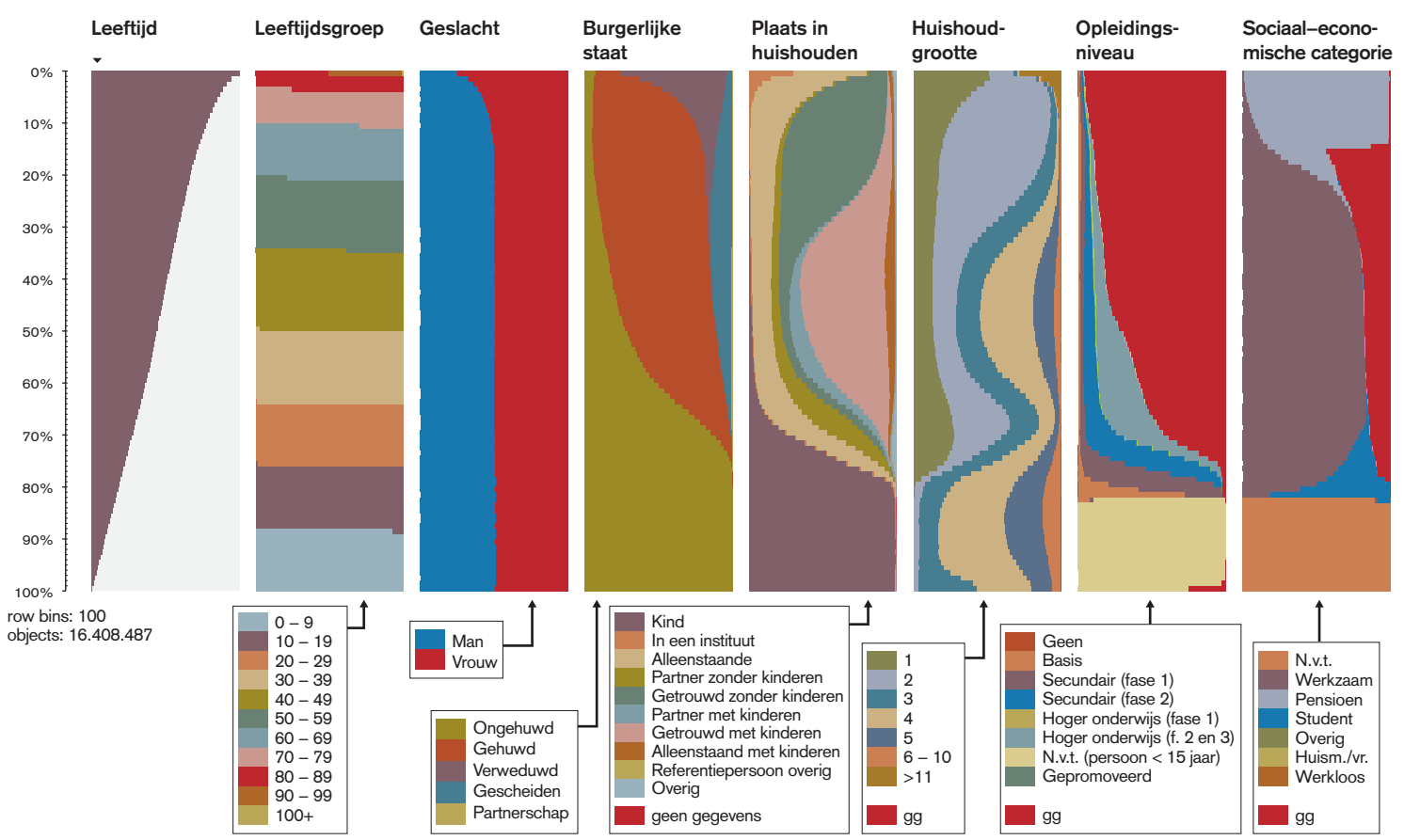

Deze figuur laat een Tableplot zien van de gegevens van 16,5 miljoen Nederlanders voor acht variabelen in het 2008-proefbestand van de Virtuele Volkstelling. Doordat van onder naar boven op leeftijd is gesorteerd, zien we duidelijk de verbanden tussen leeftijd en geslacht, burgerlijke staat, plaats in het huishouden, huishoudgrootte, opleidingsniveau en sociaal-economische categorie. Zo is te zien dat vrouwen langer leven (bovenste deel 3e kolom), een deel van de bevolking nooit trouwt (lichtgroen patroon in de 4e kolom) en dat veel van de jonge volwassenen die het huis verlaten eerst op zichzelf gaan wonen (groen patroon in de 5e kolom). Daarnaast is te zien dat de gezinsgrootte (6e kolom) op twee momenten duidelijk verandert: rond de leeftijd van 25 jaar (de leeftijd dat jongeren uit huis gaan) en rond de leeftijd van 45-50 jaar (de leeftijd van de ouders van de jongeren die het huis verlaten). Het patroon onderaan de kolom geeft de toename in de gezinsgrootte weer door de geboorte van een kind of een broertje of zusje.

Complete dataset kleurrijk weergeven in één figuur

Snel inzicht in de kwaliteit van gegevens en mogelijke dataproblemen met behulp van een Tableplot. Het CBS heeft er inmiddels ervaring mee. ***Piet Daas en Eric Schulte Nordholt***

Het Centraal Bureau voor de Statistiek (CBS) gebruikt voor de statistiekproductie steeds vaker administratieve bestanden die grote hoeveelheden gegevens bevatten. Denk hierbij aan bronnen zoals de Gemeentelijke Basisadministratie, Inschrijfgegevens werkzoekenden (van het UWV WERKbedrijf) en de Polisadministratie (van het UWV).

Eén van de grootgebruikers van administratieve bestanden is Eric Schulte Nordholt en zijn Virtuele-Volkstellingteam. De volkstelling wordt namelijk tegenwoordig virtueel uitgevoerd. Nadat in 1971 de laatste traditionele Volkstelling in Nederland werd gehouden, waarbij iedere inwoner een vragenlijst moest invullen, worden nu grote hoeveelheden gegevens uit allerlei administratieve bronnen en reeds uitgevoerde enquêtes gecombineerd (zie E-data&Research 3, jaargang 3). Controle van de correctheid van de gelegde koppelingen en de resulterende set van gegevens is belangrijk maar lastig. De Volkstelling brengt namelijk gegevens in allerlei bestanden bij elkaar en is een belangrijk ijkpunt in de vergelijking van gegevens tussen landen, met name binnen Europa. Het feit dat de gekoppelde bestanden een aanzienlijke omvang hebben, maakt controleren extra lastig. Voor de Virtuele Volkstelling van 2011 moet gedacht worden aan een dataset van ongeveer 50 variabelen voor zo'n 17 miljoen personen! Dat zijn enorm veel gegevens.

Alles in één

Om de kwaliteit van dergelijke grote bronnen goed en snel inzichtelijk te maken, is recentelijk door de CBS-onderzoekers Martijn Tennekes, Piet Daas en Edwin de Jonge een visualisatiemethode ontwikkeld. De methode, gebaseerd op een aanpak die oorspronkelijk door Duitse onderzoekers is bedacht, maakt het mogelijk de inhoud van een geselecteerd aantal variabelen in een groot databestand in één figuur volledig weer te geven. Een dergelijk figuur wordt in de literatuur een *Tableplot* genoemd. Belangrijkste verbeteringen die de CBS'ers aan het originele ontwerp hebben toegevoegd, zijn: het weergeven van de gebieden waar gegevens ontbreken, de mogelijkheid om de getoonde schaalverdeling en het aggregatieniveau van de gegevens aan te passen en het kunnen gebruiken van kleurpaletten. Daarnaast is de ontwikkelde software als *open source* beschikbaar (als 'R-package') waardoor het door iedereen kan worden gebruikt en (waar nodig) kan worden uitgebreid en aangepast.

Een Tableplot maakt patronen in de verdeling van variabelen en de samenhang tussen variabelen zichtbaar. Maar ook het ontbreken van gegevens is duidelijk. Zo laat de laatste kolom, sociaal-economische categorie, zien dat gegevens ontbreken (rode kleur) van personen die niet studeren, niet in loondienst werken en niet gepensioneerd zijn (zoals zelfstandigen). Deze informatie is in geen enkele publieke administratie te vinden en wordt met steekproeven verzameld, wat de hoeveelheid beschikbare informatie beperkt. Opvallend in kolom 8 is verder dat een aanzienlijk deel van de ouderen boven de 65 jaar nog blijkt te werken (de groene uitloper in bovenste deel van 8e kolom).

Sorteervariabele vereist

Alle genoemde bevindingen zijn, na enige oefening, direct uit het figuur te halen. Met meer traditionele methoden, zoals staafdiagrammen en tabellen, lukt dat niet zo snel en gemakkelijk. Zeker niet bij grote bestanden. Met behulp van een Tableplot kan dat wel. Andere toepassingen van Tableplots door de CBS-onderzoekers laten zien dat dergelijke figuren ook als hulpmiddel voor de controle van de interne verwerking van gegevens te gebruiken zijn. De betrokken CBS'ers verwachten dan ook dat er op het CBS in de nabije toekomst veel vaker van Tableplots gebruik zal worden gemaakt. Momenteel worden BTW-gegevens van de Belastingdienst en de Polisadministratie (loon- en uitkeringsgegevens) van het UWV op deze manier bekeken. Een dergelijke aanpak is ook voor andere onderzoekers van grote databestanden nuttig. Belangrijkste eis voor een succesvolle toepassing is wel dat in de dataset een geschikte sorteervariable, zoals leeftijd (bij persoonsgegevens) en omzet (bij bedrijfsgegevens), aanwezig is.

http://cran.r-project.org/web/packages/tabplot
http://cran.r-project.org/web/packages/tabplot GTK

KORT

Wetenschapper mag zelf e-boeken kiezen

Leden van de Koninklijke Bibliotheek kunnen sinds kort 200.000 wetenschappelijke e-boeken lenen. De KB sloot daartoe een contract met het Australische Ebook Library. Publicaties worden *patron driven* aangeboden: de bibliotheek betaalt pas wanneer het boek ook daadwerkelijk wordt geraadpleegd. Zo wordt de collectie wetenschappelijke monografieën van de KB niet langer *just-in-case* maar *just-in-time* en op basis van concrete behoeftes van onderzoekers opgebouwd. Die moeten wel inloggen om bij het aanbod te komen. (SC)

CATCHPlus: van demo tot software

Eind juni loopt CATCHPlus af. In CATCHPlus zijn enkele CATCH-demo's doorontwikkeld tot software die breed bruikbaar is in de erfgoedwereld. Musea en archieven krijgen bijvoorbeeld toegang tot een Art Recommender en software die automatisch trefwoorden suggereert aan documentalisten. De ontwikkelde tools en diensten zullen eind november gepresenteerd worden tijdens een congres, dat in samenwerking met DEN georganiseerd zal worden. Meer informatie over dit congres is binnenkort te vinden op www.catchplus.nl. (ER)

Data Portal maakt zoeken eenvoudiger

Met de DANS Data Portal is het mogelijk om datacollecties van zo'n 75 (inter)nationale organisaties te doorzoeken. Het gaat om omvangrijke collecties van organisaties zoals: het Dataverse Network, Google Public Data Explorer, World Bank, EUROSTAT, Medieval Song Network, European Union Statistics on Income and Living, Geodan, Educatief Gisportaal, Archiefwebsites Nederlandse Politieke Partijen, Werkgelegenheidssysteem LISA, en Europese en Amerikaanse data-archieven.

De data zijn afkomstig uit de volgende wetenschappelijke disciplines: geesteswetenschappen, archeologie, ruimtelijke wetenschappen engedrags- en maatschappijwetenschappen. De portal kan worden uitgebreid met sites door een mail te sturen naar info@dans.knaw.nl met de link naar de site. (HB)