STATISTIEK
VLAANDEREN

“Uitdagingen bij datakoppelingen voor statistisch en survey-onderzoek”

Nederlandstalig Platform voor Survey-Onderzoek (NPSO), 11 februari 2019, Brussel

Dagvoorzitter: Marc Callens - Statistiek Vlaanderen

Verslag: Dries Verlet -Statistiek Vlaanderen

Inleiding

Startpunt en ideaalbeeld is het aan elkaar koppelen van surveygegevens, administratieve gegevens en big data, deze samenbrengen in “data lakes” en daaruit extra informatie puren. Er zijn evenwel nog een aantal hindernissen te nemen tussen droom en daad: normering vanuit ICT-Informatiebeveiliging, bescherming van de privacy (AVG-verordening) en bescherming door Statistiekwetgeving, …

Maar er dienen zich ook diverse beloftevolle ontwikkelingen aan: toepassing van het FAIR-principe (alle digitale bronnen moeten Findable, Accessible, Interoperable en Reusable zijn), koppeling door trusted third parties, data-virtualisatie technieken en statistische experimenten die van het koppelen van databanken stilaan een dagelijkse realiteit maken.

Dit thema werd tijdens deze NPSO-bijeenkomst verder uitgediept in de volgende presentaties:

- Data integratie uitdagingen in het consumenten en media-onderzoek, Ludo Daemen - Nielsen;
- Het effect van differential privacy op sociaal-wetenschappelijk onderzoek, Daniel Oberski - Universiteit Utrecht;
- Personal Health Train: analyseren van gedecentraliseerde data, Lianne Ippel - Universiteit Maastricht;
- Uitdagingen bij datakoppelingen voor statistisch en surveyonderzoek, Eric Schulte Nordholt - CBS;
- Koppeling van administratieve data in het kader van de Belgische census, Pieter Dewitte - Statbel;
- An introduction to the Open Data Infrastructure for Social Science and Economic Innovations, Tom Emery - ODISSEI.

1. *Data integratie uitdagingen in het consumenten en media-onderzoek.*

Ludo Daemen - Nielsen

Bio: Ludo Daemen is momenteel verbonden aan VP Global R&D, Nielsen DataScience. Na studies sociologie en computerwetenschappen werkte Ludo een tijdje aan het departement sociale wetenschappen van de KU-Leuven. Sindsdien werkt hij bij Nielsen in een reeks van functies die steeds in het gebied lagen van wat nu 'Data Science' genoemd wordt.

In zijn presentatie vertrekt Ludo Daemen van een aantal uitdagingen die gesteld worden door veranderingen in de samenleving op het vlak van gebruik van informatie, consumentengedrag, enzovoort. Hij bekijkt de uitdagingen vanuit de bril van een data-scientist in het consumenten en media-onderzoek bij Nielsen. Hierbij benadrukt hij dat ze focussen op het eigenlijke gedrag van mensen, niet hun waarden, opinies of attitudes. Is data-integratie de sleutel tot succes? Het antwoord op deze uitdagingen is een cocktail van meer externe data, data-integratie en modellering. Op zijn beurt roepen deze oplossingen nieuwe vragen op.

Sleutelelement is alvast het vertrouwen dat men al dan niet heeft in de gebruikte methodologie en data. In deze context wijst Ludo erop dat in de media vaak geen feitelijke gegevens beschikbaar zijn en indien die er al zijn, zitten die verdeeld over verschillende actoren. Hoewel men vanuit de mediasector resoluut inzet op data-integratie, worden er vraagtekens gezet bij big data als wondermiddel. Het meet vaak niet wat het beoogt te meten. Daarnaast wijst Ludo ons ook op de hoge techniciteit van het integreren van data en de afhankelijkheid van andere (commerciële) actoren. Bij de analyse komt verder naar voor dat men in vele gevallen steunt op modellen waarachter veel veronderstellingen schuil gaan.

2. Het effect van differential privacy op sociaal-wetenschappelijk onderzoek.

Daniel Oberski - Universiteit Utrecht

Bio: Daniel Oberski is hoofddocent aan de afdeling methoden en statistiek van de Universiteit Utrecht. Hij werkte vijf jaar voor de ESS, was mede-oprichter van de ESRA, en deed onderzoek naar het schatten en corrigeren van meetfouten in registerdata met latente variabele modellen. Momenteel richt hij zich op het raakvlak tussen sociaal-wetenschappelijke onderzoeksmethoden en data science.

Deze presentatie gaat over Differential privacy, een nogal strikte definitie van privacy en de mogelijke gevolgen ervan voor de praktijk van het sociaal-wetenschappelijk onderzoek. Differential privacy impliceert dat publieke data beschermd worden door ze te onderwerpen aan verstoring, het toevoegen van ruis. Voordeel is alvast dat er minder koppelingsaanvallen mogelijk zijn en dat er minder privacygevoelige informatie uit de data te halen is, terwijl de data wel informatief blijven. Er zijn evenwel ook een aantal nadelen te melden. Zo zijn minder koppelingen mogelijk en heeft men veel grotere databanken nodig. Verder is het op die manier ook onmogelijk om outliers te analyseren. Is op die manier minder data beschikbaar? Dat kan men ook anders zien: misschien is op die manier meer mogelijk in vergelijking met de situatie waarin geen ruis wordt toegevoegd of data helemaal niet beschikbaar worden gesteld.

Het toevoegen van ruis aan de data kan de respondenten een gevoel van vertrouwen geven. Hoewel het ook omgekeerd kan werken. Je ontneemt de respondenten ook hun “agency”.

3. Personal Health Train: analyseren van gedecentraliseerde data.

Lianne Ippel - Universiteit Maastricht

Bio: Lianne Ippel is als Postdoctoraal onderzoeker verbonden aan het Institute of Data Science van de Universiteit Maastricht. Ze behaalde haar PhD graad aan de Universiteit Tilburg voor haar thesis "Multilevel Modeling for Data Streams with Dependent Observations", waarvoor ze in 2018 de Best Thesis Award won op de General Online Research conference in Keulen. Haar onderzoeksgebied concentreert zich momenteel op het gebied van ethisch en verantwoordelijk gebruik van Machine Learning modellen op het vlak van response style, measurement invariance en missing data.

In haar presentatie gaat Lianne dieper in op het principe van de Personal Health Train. Personal Health Train is gebaseerd op een innovatief koppelingsparadigma. Gewoonlijk brengt men de decentrale databanken samen om ze te koppelen. In de Personal Health Train laat men de databases waar ze zijn en stuurt men de analyses er naar toe.

Mogelijkheden, maar ook moeilijkheden van deze nieuwe werkwijze illustreert Lianne aan de hand van het FAIRhealth project, waar een "koppeling" wordt gemaakt tussen medische data uit de zogenaamde Maastricht Studie en meer algemene data van het Centraal Bureau voor de Statistiek.

In de analyse van gedecentraliseerde databanken combineert men drie invalshoeken: het technische (wat kan de infrastructuur aan), het inhoudelijke (ethische/juridische) en het methodologische. De invalshoeken komen ook naar voor in de benadering vanuit de "Personal Health Train". Personen en organisaties behouden de controle op de data die men al dan niet op de rails zet. Niettemin, creëert men een omgeving die toelaat om data uit verschillende instanties en in een verscheidenheid aan vormen te linken aan elkaar en te delen. Basisidee is: wetenschap versnellen door samenwerking.

In de discussie komt ook naar voor dat het niet altijd wenselijk is om de bewaring van data te leggen bij die actoren die ook instaan voor duiding en analyse van deze data.

4. Uitdagingen bij datakoppelingen voor statistisch en survey-onderzoek, casus Nederland.

Erik Schulte Nordholt - CBS

Bio: Na Wiskunde in Utrecht en Econometrie in Rotterdam te hebben gestudeerd ging Erik bij het CBS werken in 1992. In eerste instantie werkte hij voor de afdeling statistische methoden, daarna voor de divisie Sociale Statistieken. In 1995 was hij gedetacheerd bij Eurostat en in 2006 bij Statistics New Zealand. Eric is bij het CBS adviseur statistische beveiliging en verantwoordelijk voor de volkstelling.

In zijn presentatie gaf Erik een overzicht van de administratieve registers waarover het CBS beschikt en welke wensen er vanuit praktijk, beleid en wetenschap nog leven. Aan de hand van de casus "Nederlandse Volkstelling" illustreerde hij o.m. het koppelingsproces en de bescherming van de privacy in de praktijk. Zo worden verschillende types registers aan elkaar gekoppeld. Een aantal knelpunten bij het organiseren van zo'n virtuele volkstelling hebben onder meer betrekking op de moeilijkheid om beroep en opleidingsniveau in kaart te brengen op grond van registerdata, vooral beroep blijkt een harde noot om kraken. Daarnaast is er eveneens de afhankelijkheid van de registerhouders wat de beschikbaarheid en tijdigheid betreft.

Typerend voor de Nederlandse volkstelling is dat deze enkel en alleen registers gebruikt, dat deze daardoor relatief goedkoop is en dat men met de eigenlijke koppeling pas begint eens alle data beschikbaar zijn. Hij illustreerde ook hoe de koppeling in de praktijk gebeurt op grond van een centraal koppelbestand waar slechts enkele personen toegang toe hebben. Verder duidde hij eveneens de verschillende instanties die mee waken over het correct en deugdelijk verlopen van de koppelingen.

5. De koppeling van administratieve data in het kader van de Belgische Census.

Pieter Dewitte - Statbel

Bio: Pieter Dewitte is wiskundige van achtergrond (KU Leuven) en momenteel hoofd van de dienst Databanken Burgers bij Statbel (Statistiek België).

De Census is het belangrijkste voorbeeld van een project waarbij gebruik gemaakt wordt van (gekoppelde) administratieve bevolkingsgegevens in België. Pieter overliep de voornaamste registers die worden gekoppeld en hoe men koppelingen kan maken tussen gegevens voor personen en ondernemingen, via KSZ-gegevens. Hij liet ons kennis maken met de diversiteit aan registers waar bv. de gegevens rond onderwijs worden bijgehouden. De complexiteit van de institutionele Belgische setting vertaalt zich in een niet evident landschap waarin data over een verscheidenheid aan thema's of tal van actoren zit verspreid. Ook de bescherming via juridische, technische en organisatorische middelen kwam ter sprake.

6. Open Data Infrastructure for Social Science and Economic Innovations.

Tom Emery - ODISSEI

Bio: Tom Emery is the Executive Director of ODISSEI, the National Data Infrastructure for Social Science . He has a PhD in Social Policy from Edinburgh University and has worked at NIDI since 2013 as Project Manager of the GGP. Since 2016, Tom has been a member of the ODISSEI Management Board and has assisted in the development of the infrastructure.

Achter het acroniem "ODISSEI" gaat een data infrastructuur schuil ten dienste van de sociale wetenschappen op grond van data uit Nederland. Voornaamste doel van ODISSEI is het coördineren van de inspanningen om data te verzamelen en analyseren in Nederland. De "ODISSEI data facility" is het kloppend hart van deze organisatie en omvat 3 essentiële componenten: veilige toegang tot data vanop afstand (secure remote access), een rekenkrachtige computeromgeving (high-computing environment) en het toevoegen van eigen data door onderzoekers. In zijn lezing gaf Tom een overzicht van ODISSEI's data facility en illustreerde dit met verschillende cases. Uit deze cases leren we onder meer de nood aan nog krachtiger computeromgevingen. De hoeveelheid te verwerken informatie is enorm en de krachten dienen gebundeld te worden om die analyses mogelijk te maken.

Discussie

In de slotbeschouwing kwamen nog tal van interessante beschouwingen aan bod. Zo is er de stelling dat men als onderzoeker niet alle data hoeft te zien om de analyses te doen. Als men terugblikt stelt men ook vast dat o.a. door het koppelen (of in ieder geval laten communiceren) van datastromen, er meer gedetailleerde analyses mogelijk zijn, al heeft men hiervoor bijkomende infrastructuur nodig.

De link met de privé komt eveneens aan bod. Binnen deze privé-context merkt men op dat er vaak een grote afhankelijkheid is van data van andere private actoren. Het verlenen van toegang tot data van andere private actoren is geen evident verhaal. Wel wisselt men modellen en analyseresultaten uit, onder meer in het finetunen van de schattingen.

Een element dat in verschillende presentaties naar voor kwam, is dat men ruis toevoegt aan de data om deze anoniemer te maken. Maar is er al niet veel ruis op die data? Moet onze zorg niet in eerste instantie gaan naar het oplossen en voorkomen van de al aanwezige ruis in de data? Immers, hoe zeker is men van de kwaliteit van de data?

In de discussie kwam naar voor dat kwaliteit van de data een aandachtspunt is en moet blijven. Onder meer door het koppelen van registerdata is men in staat om meer gedetailleerde analyses te maken, al moeten we alert blijven voor de kwaliteit van de data. Zo zijn registerdata niet per definitie meer betrouwbaar dan surveydata en ook hier komt het thema van de nood aan duidelijke conceptuele kaders naar voor. Validatiegegevens voor zowel survey als administratieve data blijven noodzakelijk. Belangrijk hierbij is dat data voldoende gedocumenteerd zijn. Dit brengt ons opnieuw bij het principe dat tijdens de studiedag verschillende keer naar boven kwam: het FAIR-principe. Data zijn idealiter Findable (vindbaar), Accessible (toegankelijk), Interoperable (uitwisselbaar) en Reusable (herbruikbaar).