Onderzoeksprogramma Methodologie 2025-2030

December 2024

Inhoudsopgave

1. Speerpunten methodologisch onderzoek 2025-2030

Het CBS streeft naar betrouwbare, gedetailleerde, hoogfrequente en fenomeen-gerichte statistieken, die inspelen op de behoefte van de samenleving. Die samenleving verandert continu en dus moet ook het CBS meebewegen. Tegelijkertijd wil het CBS de kosten en belasting voor de samenleving zo laag mogelijk houden. Dit vereist continue innovatie. Methodologisch onderzoek is een belangrijk ingrediënt in die innovatie.

In dit document beschrijven we de onderzoeksplannen op het terrein van methodologie (vanaf nu het 'onderzoeksprogramma methodologie') voor de periode 2025 – 2030. Deze plannen worden uitgevoerd door de sector 'Research & Development', zoveel mogelijk in samenwerking met andere afdelingen binnen het CBS. De onderzoeksplannen zijn tot stand gekomen op basis van bevindingen van het onderzoeksprogramma methodologie van de afgelopen vijf jaren en op basis van consultatie van interne en externe stakeholders. Ze zijn ook ingegeven door (recente) maatschappelijke trends zoals die onder andere zijn samengevat in het CBS meerjarenprogramma 2024-2028[1], echter kent het huidige onderzoeksprogramma Methodologie een bredere tijdshorizon tot 2030.

Het onderzoeksprogramma methodologie kent een aantal belangrijke nuance-verschuivingen die zijn vertaald naar speerpunten. Deze speerpunten zijn:

1. **(Representatie)** Behoud van een representatieve afspiegeling van de Nederlandse samenleving onder meer door het centraal plaatsen van berichtgevers/respondenten;

2. **(Data gebruik)** Verbeterde uitbating en integratie van eigen data en data waar het CBS wettelijk toegang toe heeft, rekening houdend met het toegenomen belang van privacy;

3. **(Data naar informatie)** Het aanbieden van relevante informatie aan de samenleving, bijvoorbeeld via versnelling, verdieping en een eigentijdse 'publicatie van' of 'communicatie over' statistieken;

4. **(Artificial Intelligence als tool)** Verantwoorde toevoeging van Artificial Intelligence (AI) als tool in alle fases van het statistisch proces en standaard bedrijfsvoering;

In de volgende paragraven beschrijven we ten eerste de speerpunten in meer detail. Daarna motiveren we de speerpunten aan de hand van een terugblik op het onderzoeksprogramma methodologie 2020- 2025. Ten slotte, in hoofdstuk 2, vertalen we de speerpunten naar concrete thema's. In de appendices relateren we de onderzoeksplannen aan interne en externe visiedocumenten en gaan we in op de organisatie van het onderzoeksprogramma methodologie.

[1] https://www.cbs.nl/-/media/cbs/over-ons/organisatie/cbs-meerjarenprogramma-2024-2028-def.pdf

Speerpunt 1: Representatie

Een complete afspiegeling van de Nederlandse samenleving in CBS-data is een absolute voorwaarde voor betrouwbare statistieken. Het survey klimaat is steeds complexer geworden. Bereidheid om aan enquêtes mee te doen daalt al vele jaren en is op een punt aanbeland waar openlijk wordt gesproken over andere communicatiekanalen dan die beschikbaar zijn in steekproefkaders. Daarnaast is veel meer maatwerk en aandacht voor respondenten een steeds vaker geopperde oplossing. Een tweede oplossingsrichting bestaat uit het meer gebruiken van bestaande bronnen, zoals 'Big data'. Rond dit alles speelt het toegenomen belang van privacy en ethiek. Hoe deze wensen effectief en verantwoord te integreren in benaderstrategieën is een cruciale opgave.

Speerpunt 2: Data gebruik

In het vorige meerjarenprogramma werd reeds de nadruk gelegd op het effectiever en efficiënter uitbaten van databronnen. Dit is noodzakelijk vanuit meerdere perspectieven, bijvoorbeeld voor het verlagen van de belasting voor respondenten maar ook voor de toegenomen aanvullende statistische dienstverlening. De al genoemde toenemende complexiteit in de CBS-waarneming maakt die noodzaak nog groter. De vele databronnen binnen en buiten het CBS bieden voldoende kansen op versnelling en detaillering van statistieken en op meer maatwerk. Tegelijkertijd is behoud van privacy een steeds prominentere randconditie, onder andere via de implementatie van de AVG. Deze tegenstelling vraagt om tactieken en methoden die geavanceerder, maar veilig, integreren van databronnen mogelijk maken.

Speerpunt 3: Data naar informatie

De hoeveelheid statistieken en maatwerk-analyses die het CBS jaarlijks produceert is enorm. Het CBS is een spil in nationale en internationale beleidsvorming. In een schijnbaar complexer en internationaler wordende samenleving is de behoefte aan snellere en meer gerichte analyses en cijfers groot. Statistische output krijgt daarmee nieuwe vormen zoals monitors met een brede set van indicatoren en interactieve zoekmachines. Verder, door complexe fenomenen te beschrijven moet het mogelijk worden om causale effecten meetbaar te maken, bijvoorbeeld door te kijken naar de effecten van beleid. Het goed kunnen blijven uitleggen en in context plaatsen van statistieken nemen daarmee sterk in belang toe. Het vertalen van geavanceerde methoden naar officiële statistiek en het goed uitleggen van resultaten zijn onmisbare onderdelen. Als onderdeel daarvan zal het CBS moeten onderzoeken hoe diens standaard- en maatwerk publicaties worden verwerkt door eindgebruikers. Bijvoorbeeld, kan empirisch onderzoek vragen beantwoorden of CBS publicaties correct geïnterpreteerd worden of als informatief worden ervaren.

Speerpunt 4: Artificial Intelligence als tool

Waar Artificial Intelligence (AI) vroeger vooral voorbehouden was aan wetenschappers of experts bij bedrijven of overheden, heeft de snelle technologische vooruitgang bij grote technologiebedrijven en een levendige open-source community ertoe geleid dat AI toegankelijk werd voor een breed publiek. Deze ontwikkeling biedt zowel kansen als uitdagingen. De toegankelijkheid van AI opent de deur naar innovatie, economische groei, en persoonlijke ontwikkeling, maar roept ook vraagstukken op over privacy, betrouwbaarheid, verantwoordelijkheid en de menselijke rol in een door AI-gedreven toekomst. Het CBS gaat de komende jaren de kansen van AI onderzoeken, bijvoorbeeld op het gebied van AI-automatisering. Tegelijkertijd moet het CBS kritisch kijken naar de risico's van AI op het gebied van betrouwbaarheid en uitlegbaarheid van de uitkomsten van AI-algoritmes. Onderzoek naar standaarden voor AI-methoden voor het verantwoord inzetten van AI in de officiële statistiek of standaard bedrijfsvoering is daarom een cruciaal speerpunt.

Belangrijke input voor het onderzoeksprogramma Methodologie 2025-2030 zijn de bevindingen en aanbevelingen uit het nog lopende onderzoeksprogramma (2020 – 2025). Het onderzoeksprogramma Methodologie 2020 – 2025[2] koppelde haar speerpunten en doelen aan een viertal ontwikkelingen: (1) de data-samenleving, (2) noodzaak tot meer maatwerk in verzameling en combinatie van databronnen, (3) een grotere behoefte aan duiding, en (4) de steeds invloedrijker wordende internationale context. De speerpunten lijken daarmee deels op die voor de aankomende periode. Methodologisch onderzoek laat zich, net als innovatie, nu eenmaal niet vangen in afgebakende vijfjaarperiodes. Sommige delen van het onderzoeksprogramma lopen daarom ook door, maar met belangrijke nuanceverschuivingen. Het onderzoeksprogramma Methodologie 2020 – 2025 kende zes thema's: (1) 'New observing techniques & Data collection', (2) 'Big Data, Data Mining & Artificial Intelligence', (3) 'Data integration', (4) 'Data security', (5) 'Statistical modelling' en (6) 'Complexity science'. In Tabel 1 geven we de belangrijkste resultaten voor de zes thema's. Deze zijn vervat in een groot aantal interne en externe rapporten, artikelen, PhD-proefschriften, master scripties, congrespresentaties, (software) applicaties, proof-of-concepts en implementaties in productie.

THEMA	BELANGRIJKSTE RESULTATEN
NEW OBSERVING TECHNIQUES & DATA COLLECTION	• Breed inzicht in designkeuzes in smartphone-vragenlijstontwerp • Evaluatie van sensordata voor landbouwstatistieken en het ontwikkelen van een meer efficiënte methode van dataverzameling uitgaande van de situatie bij bedrijven. • Doelgroepenbenadering rekening houdend met mode-specifieke meetverschillen • Proof-of-concept smart applicaties voor Budgetonderzoek (BO) en Onderweg in Nederland (ODiN) • Onderzoek naar hot-spots bij bedrijvenwaarneming • Onderzoek naar situatie bij bedrijven in kader van bedrijf centraal
BIG DATA, DATA MINING & ARTIFICIAL INTELLIGENCE	• CBS speelde een leidende rol in ESSnets Big Data I en II en heeft daarin verschillende internationale use cases uitgewerkt • De statistiek platformeconomie • Een start met best practices voor verantwoord gebruik AI-ML • Verschillende tools voor visualisatie en correctie vertekening
DATA INTEGRATION	• Imputatie onder randtotalen, toegepast bij hoogst behaalde opleiding • Corrigeren misclassificaties, toegepast bij toewijzing SBI en webshops en rapport over de 'omgevingswet' • Inzet latent-klassenanalyse, toegepast bij arbeidspositie, verkeersongelukken, huur-koopwoningen en energiestatistieken • Schatten verborgen populaties en uitbreiding zogenaamde multiple-systems-estimation, toegepast bij daklozen
DATA SECURITY	• Risico's van datalekken in AI-ML-methoden • Beveiliging van statistieken op kaarten • Beveiliging van netwerkdata (personen- en bedrijvennetwerk) • Beveiliging meerdimensionale output, toegepast bij Volkstellingen
STATISTICAL MODELLING	• Kleine-domeinschatters voor consistente tijdreeksen met methodebreuken, toegepast o.a. bij onderzoeken Onderweg in Nederland (ODiN), ziekte verzuim en Sociale Samenhang voor het maken van officiële publicaties • Het maken van snellere cijfers voor de GEZO tijdens corona die gecorrigeerd zijn voor het wegvallen van CAPI via tijdreeksmodellen • Correctie voor COVID-19 methodebreuken via tijdreeksmodellen voor de maandcijfers over de Beroepsbevolking en het Consumenten vertrouwen • Integratie van big-databronnen in tijdreeksen, toegepast bij Korte Termijn Statistieken en onderzoek Enquête Beroepsbevolking • Tactieken voor enquêtedrukspreiding bedrijven
COMPLEXITY SCIENCE	• Uitwerking en verbreding van het personen- en bedrijvennetwerk • Grootschalige berekeningen gedaan op de netwerken, zoals bijvoorbeeld het berekenen van individuele segregatie-scores voor de gehele bevolking. • Agent-based modellen voor ontwikkeling van COVID-19 • Begrip van inkomensafhankelijk consumentengedrag via agent-based modellen • Begrip van supply-chain mechanismen

Tabel 1: *Overzicht belangrijkste resultaten Onderzoeksprogramma Methodologie 2020 – 2025.*

[2] https://www.cbs.nl/nl-nl/achtergrond/2020/25/visie-methodologie-onderzoek-2020-2025

Op basis van deze resultaten zijn dit de belangrijkste conclusies en aanbevelingen vanuit het onderzoeksprogramma Methodologie 2020 – 2025:

- ‘Smart surveys’ waren een belangrijk deelthema binnen “New observing techniques & Data collection”. Veldstudies tonen aan dat de toegevoegde waarde van ‘smart surveys’ vooral ligt in verhoogde datakwaliteit en niet in hogere/meer representatieve respons. ‘Smart surveys’ betekenen qua bedrijfscultuur daarnaast een grote omslag in logistiek en infrastructuur. De noodzaak tot het centraal plaatsen van respondenten is gegroeid, maar bemoeilijkt door de steeds verdere beperkingen in kanalen zoals CATI.
- Een kwaliteitsraamwerk voor ‘machine learning’ is opgesteld. Dit raamwerk dient verder uitgewerkt en getoetst te worden aan de hand van relevante toepassingen binnen het CBS zoals ‘budgetonderzoek’ en ‘onderweg in Nederland’.
- Binnen en buiten het CBS is er een sterk groeiende belangstelling voor zogenaamde non-probability-samples (niet-kanssteekproeven). Methoden voor integratie van dergelijke bronnen zijn toegepast binnen het CBS en verder uitgewerkt. De onderliggende aannames zijn lastig te toetsen, maar lijken vaak niet op te gaan. Slimme, gerichte aanvullende waarneming, in de vorm van nieuwe communicatiekanalen, kan mogelijk verbetering brengen.
- Veel meer dan voorheen is het cruciaal om schatingsmethoden achter de hand te hebben die kunnen omgaan met methodebreuken. Deze methoden zijn gebaseerd op tijdreeksen.
- Artificial Intelligence is voor informatiebeveiliging een belangrijk fenomeen waar nog vele open vragen liggen die beantwoord moeten worden.
- Beveiliging van informatie is een breder gebied geworden met de komst van allerlei nieuwe vormen van output en sterkere nadruk op integratie. Synthetische data zijn één van de mogelijk routes die onderzocht werd en nog verder onderzocht moet worden.
- Het personennetwerk is een waardevolle nieuwe bron gebleken waarop allerlei nieuwe toepassingen mogelijk zijn. Deze dienen verder verkend te worden. Verder zijn er inmiddels opleidings-, herkomst en diabetes-segregatiescores ontwikkeld.

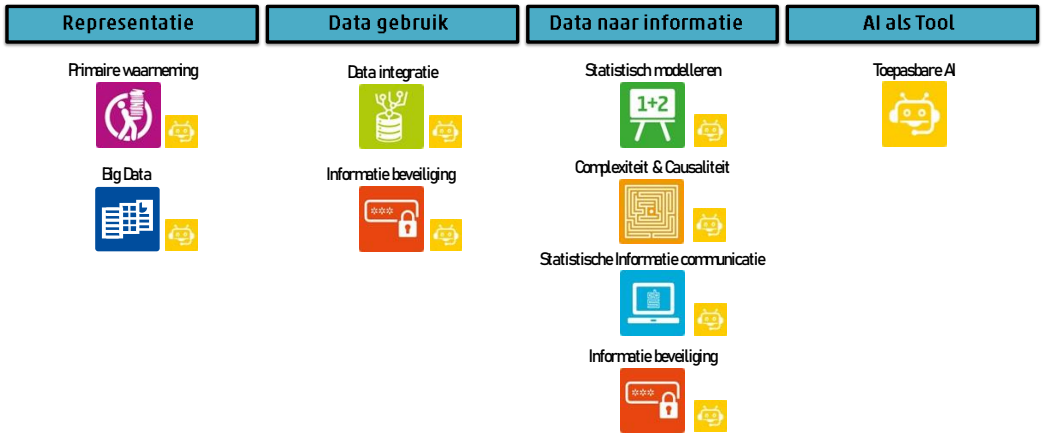
Een deel van de bestaande thema’s heeft met deze ervaringen geleidelijk een her-prioritering doorgemaakt. In onderzoek naar waarneming wordt meer de nadruk gelegd op de berichtgever/respondent. Artificial Intelligence, met in het bijzonder ‘machine learning’, is een groot onderzoeksgebied geworden. In de verkenning van nieuwe data-integratiemethoden staan niet-kanssteekproeven meer centraal. De toepassingen in de officiële statistiek van geavanceerde methoden, zoals statistische modellen of methoden uit de complexiteits-wetenschap, zijn verbreed. Ook is informatie beveiliging breder en belangrijker geworden.

Naast de zes thema’s was er een (buiten methodologie om) onderzoeksthema “Data querying & processing” dat verbonden was met het onderzoeksprogramma methodologie. Dit thema richtte zich sterk op uittesten van architectuur en tooling om die zodanig te optimaliseren dat ook nieuwe methoden goed verankerd zouden kunnen worden in bestaande processen ook al zouden die grotere eisen stellen aan computer-rekenkracht of -geheugen. Daarnaast participeerde het in de ontwikkeling van ‘mobile device’ applicaties voor zogenaamde ‘smart surveys’. Het thema had gedurende de periode 2020 – 2025 last van onderbezetting en is geleidelijk aan gestopt.

2. Van speerpunten naar onderzoeksthema's

Net als in het voorgaande onderzoeksprogramma, zal ook het programma voor 2025 - 2030 opgebouwd worden uit onderzoeksthema's. De onderzoeksthema's zijn inhoudelijk gekaderd en nemen of een deel van het statistisch proces voor rekening of een deel van de methoden. Figuur 1 geeft een schematisch overzicht van hoe de inhoudelijke thema's zijn verdeeld over de speerpunten. De thema's worden getrokken door één of meerdere methodologen, de zogenaamde thema-trekkers. Deze personen zijn tevens de ambassadeurs van het thema. Van thema-trekkers wordt verwacht dat ze op de hoogte zijn van externe (wetenschappelijke) ontwikkelingen en van de interne vraag vanuit statistische divisies (voor diens specifieke onderzoeksonderwerp). We zullen nu de speerpunten in iets meer detail toelichten. In Hoofdstuk 2 gaan we vervolgens in meer detail in op de onderzoeksthema's.

Het speerpunt **Representatie** wordt bovenal geadresseerd in thema's 'Primaire waarneming' en 'Big data' die daarin twee oplossingsrichtingen vertegenwoordigen: versterking en verfijning van de eigen waarneming en inclusie van bestaande 'Big data' bronnen. Het thema 'Primaire waarneming' doet onderzoek naar het verbeteren van enquêtering waarbij de respondent (personen, huishoudens of bedrijven) centraal komt te staan. Dit omvat onder andere onderzoek naar het efficiënt aansluiten op data bronnen bij de respondent en onderzoeken hoe verschillende type respondenten te benaderen en bevragen.
De data uit primaire waarneming en uit administratieve registers kunnen in principe goed aangevuld worden met behulp van nieuwe databronnen die openbaar op het internet beschikbaar zijn: zogenaamde 'Big data' bronnen (denk daarbij aan ongestructureerde tekst-, beeld- en sensor-data). Meestal zijn er dan extra bewerkingen en analyse nodig. Juist in het kader van officiële statistiek, waar lange-termijn consistentie als kwaliteitseis essentieel is, betekent dit dat de meeste extern beschikbare methoden of technieken echt aanpassing en extra onderzoek behoeven voordat ze voor het CBS geschikt zijn.
Aan "Primaire waarneming", met name in surveys die functies van 'smart devices' inzetten, kan mogelijk AI toegevoegd worden om respondenten/berichtgevers te helpen. Het is van belang dat dit verantwoord, begrijpelijk en veilig gebeurt. Het thema "Toepasbare AI" zorgt voor de noodzakelijke 'best practices'.



Figuur 1: *Een schematische weergave van speerpunten naar onderzoeksthema's, waarbij we de thema's indelen op basis van diens belangrijkste onderzoeksonderwerp. Het is echter mogelijk dat thema's ook onderzoeksvragen van andere speerpunten onderzoeken. Het thema 'Toepasbare AI' werkt aan (generieke) standaarden en richtlijnen voor het verantwoord inzetten van AI en ziet de overige thema's als afnemers (met potentiële use-cases).*

Speerpunt **Data gebruik** is breed en divers. Dit speerpunt wordt onder andere belegd in het thema “Data integratie”. Het CBS beschikt over veel data, verkregen van bedrijven, burgers, overheden en andere organisaties. Los van elkaar beschouwd hebben de variabelen die op verschillende manieren worden verzameld (uit enquêtes, ‘Big data’ maar ook uit registraties) zeker wel waarde, maar de volle benutting van de informatie die in data besloten is, kan alleen worden bereikt als de data geïntegreerd kan worden. Het combineren van data op basis van imperfecte of onvolledige koppelvlakken, en het schatten van de bijbehorende onzekerheden, is zeker niet een al opgelost probleem, vooral wanneer een deel van die data uit niet-traditionele bronnen komt. Dit thema onderzoekt methoden om data te integreren, evidente fouten en ontbrekende waardes te corrigeren en om kwaliteit van geïntegreerde data te toetsen.
Naast dit thema is ook voor het thema “Statistisch modelleren” data integratie een belangrijk speerpunt. Waar thema “Data integratie” databronnen in de eerste plaats als gelijkwaardig ziet, zoekt thema “Statistisch modelleren” vooral naar databronnen die ingezet kunnen worden in de verwerking en versterking van een hoofd-databron. Hierbij komen veelal tijdreeksmethoden aan de orde. Deze twee thema’s komen overeen met verschillende methodologische disciplines maar werken nauw samen.
Voor thema’s “Primaire waarneming” en “Complexiteit en causaliteit” is het combineren van bronnen geen hoofddoel, maar wel een middel. Surveys zijn aanvullend op beschikbare databronnen. Dieper begrip van verbanden is noodzakelijk voor integratie van databronnen.
De data die het CBS verzamelt en samenstelt, beschouwt het CBS als publiek goed. Er lopen verschillende initiatieven voor het verbeteren van de toegankelijkheid van CBS data. In het onderzoeksprogramma wordt hier aan gewerkt via onderzoek naar methoden die de vindbaarheid van CBS data verbeteren. De keerzijde van een open en toegankelijk CBS, is dat de kans op onthulling of data lekkage toeneemt. Vertrouwelijk omgaan met gegevens van individuele eenheden is één van de voorwaarden voor een betrouwbaar CBS. Het CBS is internationaal voortrekker van onderzoek op het gebied van informatiebeveiliging, maar door de veranderde wereld (zowel technologisch als cultureel) zijn er nieuwe uitdagingen ontstaan. De komende jaren wordt onderzoek gedaan om de huidige standaard van informatiebeveiliging te handhaven, te verbreden naar nieuwe vormen van output en naar manieren om (geïntegreerde) data veilig te delen (bijvoorbeeld via synthetische data of ‘Data Spaces’).

Speerpunt **Data naar informatie** wordt bovenal belegd in de thema’s “Statistisch modelleren”, “Complexiteit en causaliteit” en “Statistische Informatie Communicatie”.
Het doel van statistische modellen is om gedetailleerde, hoogfrequente en fenomeen-gerichte statistieken (zoals over de economie en het klimaat) te maken.
Gewoonlijk publiceert het CBS dergelijke statistieken los van elkaar, zonder rekening te houden met de onderlinge interacties tussen deze fenomenen. Het thema “Complexiteit en causaliteit” richt zich op methoden die in staat zijn om statistieken samen te stellen die fenomenen beschrijven als een complex systeem, waarbij hun interacties worden meegenomen. Deze thema’s richten zich dus primair op meer duiding (context) bij statistieken.
Daarnaast is er noodzaak om inzicht te krijgen in wijze waarop gebruikers onze communicatie over statistieken interpreteren, gebruiken en waarderen. In een tijd waarop informatie en nieuws snel geconsumeerd worden, bijvoorbeeld via sociale media, is er een kans dat berichtgeving verkeerd geïnterpreteerd wordt. De grote diversiteit aan gebruikers van CBS-cijfers (van studenten, burgers tot beleidsbepalers) maakt de communicatie van statistische informatie uitdagend. Het CBS zal daarom onderzoek moeten doen naar hoe diens standaard en maatwerk publicaties worden verwerkt door eindgebruikers. Empirisch wordt onderzocht hoe deze communicatie (web-artikelen, visualisaties of dashboards op de CBS-website) door

gebruikers wordt ervaren en geïnterpreteerd. Dit is belegd in een nieuw onderzoeksthema “Statistische Informatie Communicatie”.
Thema “Informatie beveiliging” treedt opnieuw op als bewaker van privacy bij het publiceren van of communiceren over statistieken, waarbij het risico op gegevensonthulling wordt afgewogen tegen het behoud van informatiewaarde.

Tenslotte het speerpunt **Artificial Intelligence als tool**. Recente ontwikkelingen in Artificial Intelligence (AI) maken AI methoden potentieel interessant voor gebruik in de officiële statistiek. AI wordt beschouwd als een veelzijdig hulpmiddel dat bij veel verschillende type toepassingen in het gehele statistisch proces (of in de standaard bedrijfsvoering) kan worden ingezet. Echter, bij veel van die toepassingen leven soortgelijke (generieke) vragen over bijvoorbeeld de kwaliteit en uitlegbaarheid van de uitkomsten van AI modellen. AI-methoden worden nu eenmaal nog minder goed begrepen en bieden minder kwaliteitsgaranties vergeleken met standaard statistische methoden die routinematig gebruikt worden bij statistische bureaus. Onderzoek naar standaarden voor AI methoden dat kan leiden tot het verantwoord inzetten van AI in de officiële statistiek (of standaard bedrijfsvoering) wordt nu gecentraliseerd in een nieuw onderzoeksthema “Toepasbare AI”.

Alle speerpunten dekken meerdere disciplines en meerdere delen van het statistisch proces. Om die redenen worden de speerpunten in meerdere thema’s behandeld. Ze zorgen daar ten opzichte van de periode 2020 - 2025 ook voor belangrijke verschuivingen in focus. In de volgende paragrafen geven we per thema meer detail over de niche, de urgentie en de methodologische uitdagingen. De in de thema’s genoemde onderzoeksvragen vormen geen uitputtende lijst van al het potentiële onderzoek en zijn slechts ter illustratie. Gegeven de spreiding van speerpunten over thema’s is doorlopende afstemming en uitwisseling onmisbaar.

2.1. Primaire waarneming

In de komende jaren zal het CBS-data blijven verzamelen via enquêtering om te voldoen aan de statistische behoeften van de samenleving. Echter zijn de afgelopen jaren de responspercentages gedaald en is er een dringende behoefte om de uitvraag bij bedrijven en personen te minimaliseren. Dit doet het CBS door te onderzoeken hoe verschillende respondentengroepen efficiënt kunnen worden benaderd en door nieuwe technologieën te verkennen die het aanleveren van gegevens aan het CBS eenvoudiger en sneller kunnen maken.

Niche en urgentie

Het CBS maakt statistieken op basis van registers en eigen primaire waarnemingen (voornamelijk op basis van enquêtes). Het CBS gebruikt registers waar mogelijk, maar voor een aanzienlijk deel van de informatiebehoefte is primaire waarneming nodig door middel van enquêtes. Ter illustratie, per jaar worden circa 30 persoonsenquêtes uitgevoerd waar meer dan anderhalf miljoen mensen voor worden benaderd en worden er circa 70 bedrijfsenquêtes uitgevoerd waarvoor meer dan één miljoen vragenlijsten naar bedrijven worden gestuurd. Ook in de toekomst zal een belangrijk deel van de benodigde data rechtstreeks verzameld worden met primaire waarneming om te blijven voldoen aan onze leveringsverplichtingen richting o.a. Eurostat. Primaire waarneming is en blijft dus een belangrijke dataverzamelingsmethode voor het CBS. Echter, de trends van dalende responscijfers bij personenwaarneming en de wens van bedrijven om efficiënter aan het CBS te kunnen rapporteren zet zich door. Continue verbetering en vernieuwing van primaire waarnemingsmethoden zijn daarom urgent. Daartoe moeten we ons steeds weer aanpassen aan de veranderende maatschappij. De maatschappij krijgt een steeds meer diverse samenstelling (in afkomst, taal en geletterdheid, maar ook wat betreft technische vaardigheden) en tegelijkertijd staat men minder open voor betrokkenheid bij overheidsorganisaties. Voor bedrijven heeft het correct, volledig en tijdig invullen van CBS-vragenlijsten niet altijd hoge prioriteit ook al zijn de meeste bedrijfsenquêtes verplicht en wordt er strenger gehandhaafd.

Het onderzoek dat binnen dit thema gedaan wordt, heeft betrekking op (1) het optimaliseren van de aansluiting tussen bedrijfsadministraties en CBS uitvraag, c.q. aansluiting bij de leefwereld van personen en huishoudens (het correct in kaart brengen van die situatie is een onlosmakelijk onderdeel van het onderzoek), (2) hoe de (verschillende typen) respondenten vervolgens te benaderen, en tenslotte (3) hoe de rapportering naar het CBS in te richten. Rapportering kan met vragenlijsten of op basis van technisch geavanceerde methoden en technologieën, zoals apps, sensoren, en ‘system-to-system’ datacommunicatie. Deze nieuwe technologieën bieden mogelijkheden om enerzijds de kans op responderen te verhogen (doordat bijvoorbeeld het invullen van een vragenlijst makkelijker wordt) en anderzijds om meer data, met een hoger detailniveau, sneller te verzamelen, waarbij de respondent niet altijd meer vragen over moeilijk te meten concepten hoeft te beantwoorden omdat de informatie direct uit de nieuwe technologieën te halen valt.

Bovenstaande onderzoeksrichtingen sluiten aan bij de recentelijk opgestelde visies voor primaire waarneming: één voor personen en huishoudens (in 2023, bekend als “Visie Personenwaarneming”) en één voor bedrijven (in 2022, bekend als “Bedrijf Centraal”). Hierin staat, meer dan in het verleden, niet meer het CBS maar de respondent centraal.

Belangrijke criteria voor respondenten bij het rapporteren aan het CBS, zijn:

- De rapportage is efficiënt en gebeurt op een veilige manier,
- Het is duidelijke welke gegevens worden gevraagd, waarom, en wat men in de nabije toekomst nog van het CBS kan verwachten, en
- Het is duidelijk wat het CBS is;

Dit alles gebeurt met waardering voor de (inspanning van de) respondent. Daarbij geldt natuurlijk wel de randvoorwaarde dat het CBS op efficiënte wijze tijdig kan blijven voldoen aan de informatiebehoefte van de maatschappij.

Het onderzoek van de afgelopen jaren binnen dit thema heeft, voor zowel personen als bedrijfswaarneming, geresulteerd in onder andere de optimalisering van de benaderstrategie (doelgroepenbeleid), het invoeren van 'CAWI-only' en 'mixed-mode designs', het smartphone first herontwerpen van de lay-out van de vragenlijsten, en het verkennen van de mogelijkheden en valkuilen van het gebruik van sensoren, apps, en het automatisch koppelen van vragenlijsten aan bedrijfssystemen (bijvoorbeeld voor 'system-to-system' data communicatie).

Uitdaging

We zien op dit moment nog meerdere uitdagingen bij primaire waarneming die we nu opsplitsen voor enerzijds de personen- en anderzijds bedrijvenwaarneming.

Bij personenwaarneming zien we wat betreft de benaderstrategie de participatie in CBS-vragenlijsten van moeilijk bereikbare groepen zoals jongeren en mensen met een niet-Nederlandse achtergrond als uitdaging. Wat betreft de vormgeving van vragenlijsten is de verscheidenheid in techniekadaptatie een probleem. Er zijn diverse nieuwe mogelijkheden zoals het gebruik van apps en sensoren. Er is niet één techniek die een duidelijke voorkeur heeft onder respondenten. Dit maakt een combinatie van verschillende techniekadaptaties voor vragenlijsten noodzakelijk. Bovendien vraagt het gebruik van deze nieuwe technieken een behoorlijke investering. Technologische ontwikkelingen maken het noodzakelijk om continu de vragenlijsten te blijven ontwikkelen om tegemoet te komen aan de wensen en gebruiken van respondenten. De wens om moeilijk bereikbare groepen te bereiken samen met de technologische ontwikkelingen maken het doorgronden van de situatie bij de respondent een essentieel onderdeel in het onderzoek van primaire waarneming.

Voor bedrijven is bekend dat het ophalen van gegevens uit hun administraties en het berekenen van de gevraagde gegevens, veel tijd kost waarbij meerdere personen betrokken (kunnen) zijn. Daarbij speelt dat de basisgegevens in hun administraties niet altijd aansluiten bij de gegevens die in vragenlijsten worden gevraagd. Daarnaast kunnen bedrijven veel verschillende vragenlijsten krijgen, en gevraagd worden om te rapporteren op momenten dat ze de gegevens (nog) niet hebben. De uitdaging bestaat eruit om hiervoor efficiënte, indien mogelijk geautomatiseerde oplossingen te vinden, die aansluiten bij de interne bedrijfsprocessen. Het niveau van de respons is voor bedrijfswaarneming een minder groot issue, al blijft het een uitdaging om met bestaande benaderstrategieën het na te streven responspercentage te halen. Een uitdaging in de communicatie met bedrijven is om hen te vertellen wat het CBS is en waarom het CBS al die detailgegevens verzamelt.

Waar tot voor kort het CBS als uitgangspunt werd genomen voor waarnemingsmethodes, komt nu met de nieuwe visies de respondent centraal te staan. In plaats van bestaande CBS-methodes aan te passen aan de situatie van de respondent, onderzoeken we nu

waarnemingsmethodes die de respondent als uitgangspunt nemen De verwachting is dat met deze transitie nieuwe oplossingen voor het themagebied in beeld komen.

Dit alles samenvattend roept een drietal hoofd onderzoeksvragen op.
De eerste onderzoeksvraag richt zich op de respondent (personen en bedrijven). Wat is de situatie bij de respondent? Met daaraan gerelateerd de vragen: Welke specifieke homogene (doel)groepen van respondenten kunnen we onderscheiden? Wat zijn de redenen voor non-respons of late respons? Hoe werkt het rapporteren aan het CBS (en andere instanties) vanuit het perspectief van respondenten, bijvoorbeeld te onderzoeken met paradata?
De tweede onderzoeksvraag richt zich op de benadering van respondenten. Uitgaande van de opgedane kennis en inzichten van de eerste onderzoeksvraag, kunnen we nieuwe oplossingen bedenken om personen en bedrijven (per doelgroep) te benaderen? En hoe effectief zijn deze nieuwe oplossingen? Wat zijn de effecten op respons en datakwaliteit?
De derde onderzoeksvraag betreft het rapporteren aan het CBS. Welke nieuwe technologieën zijn er en hoe kunnen we deze gebruiken ter vervanging van of geïntegreerd in vragenlijsten? Gecombineerd met de opgedane kennis en inzichten van de eerste onderzoeksvraag, kunnen we dan nieuwe rapportage-oplossingen voor traditionele vragenlijsten (per doelgroep) ontwikkelen? Is een nieuwe aanpak in de praktijk efficiënt voor respondenten (uitgaande van de eigen interne situatie en rapportageprocessen)? Is het inzetten van nieuwe technologieën een verbetering t.o.v. de eerdere aanpak? En wat zijn uiteindelijk de effecten op respons en datakwaliteit, en op bestaande CBS-methodes en processen?

2.2. Big data

Geschat wordt dat zo'n 90% van alle momenteel beschikbare gegevens in de wereld gedurende de afgelopen 2 jaar zijn gegenereerd. Veel van deze nieuwe, grote, datasets zijn het gevolg van de interactie tussen computersystemen en personen of bedrijven. Dergelijk gegenereerde data worden 'Big data' genoemd en zijn potentieel interessant voor de officiële statistiek. Het gebruik van 'Big data' is echter niet eenvoudig vanwege de specifieke eigenschappen van (de gegevens in) die bronnen. Ze zijn immers niet met het doel van statistiek maken verzameld. Het onderzoeksthema 'Big data' bestudeerd methoden om dergelijke data optimaal te benutten voor gebruik binnen de officiële statistiek.

Niche en urgentie

Gegevens worden vaak met behulp van enquêtes verzameld. Deze aanpak is tijd- en arbeidsintensief. Het hergebruiken van data die reeds door anderen zijn verzameld, aangeduid als secundaire data, kan hiervoor een goed alternatief zijn. Enkele voordelen zijn de lage kosten en de relatief grote hoeveelheden (vaak ook historische) data die kunnen worden verkregen zonder extra administratieve lastendruk voor bedrijven of personen. Registers zijn een bekend voorbeeld van secundaire data. Registers zijn tabelvormige, gestructureerde datasets die doorgaans door andere overheidsinstanties worden verzameld en uitermate geschikt zijn voor (her)gebruik door het CBS. Enkele voorbeelden van registers zijn: financiële gegevens over personen, verzameld door de Belastingdienst en gegevens over onroerend goed (bijv. gebouwen, grond, etc.) verzameld door het Kadaster.

Naast registers zijn er, in onze moderne wereld, veel nieuwe, grote datasets beschikbaar die mogelijk relevant kunnen zijn voor de officiële statistiek. Voorbeelden van dergelijke 'Big data' bronnen zijn afbeeldingen (zoals satelliet- en luchtfoto's), teksten (zoals die op webpagina's en in bedrijfsrapporten) en sensormetingen (zoals verkeers- en ruimtelijke-data). De hoeveelheid en diversiteit aan 'Big data' is de afgelopen jaren explosief toegenomen. Nadeel is echter dat deze gegevens vaak ruis bevatten, vluchtig en ongestructureerd zijn, zeker in vergelijking met registergegevens. Dit leidt tot de belangrijke onderzoeksvraag of het mogelijk is 'Big data' als vervanging of aanvulling in te zetten voor de gegevensbehoefte van het CBS.

Dit onderzoeksthema richt zich dan ook op de belangrijkste vragen voor het optimaal benutten van dergelijke data voor officiële statistieken, namelijk (1) het conceptualiseren van de statistische informatie die in 'Big data' aanwezig is, (2) het corrigeren van fouten in 'Big data' (zoals selectiebias, meetfouten, etc.), en (3) het efficiënt extraheren van informatie uit grote data bronnen en het efficiënt verwerken daarvan.

Uitdagingen

Het werken met 'Big data' biedt kansen voor de officiële statistiek. Echter, daar waar er veel ervaring is met registerdata - op het CBS en bij andere officiële statistische bureaus – is het gebruik van 'Big data' nog geen routine. De eigenschappen van dergelijke bronnen verschillen fundamenteel met die van data verzameld met behulp van primaire waarneming en wijken vaak ook af van registerdata. 'Big data' bronnen worden namelijk niet specifiek voor de officiële statistiek gegeneerd en zijn het gevolg van de interactie tussen computersystemen en personen en bedrijven, waardoor i) de kwaliteit slecht kan zijn, ii) de inhoud relatief snel kan veranderen over de tijd en iii) er een vertaalslag noodzakelijk kan zijn om van ruwe gegevens naar de gewenste statistische concepten te komen. Dit maakt kwaliteitsonderzoek, op verschillende niveaus, noodzakelijk. Er is een dringende behoefte aan het ontwikkelen van methodologie om op een betrouwbare manier informatie uit 'Big data' bronnen te halen. Hierbij kan veel geleerd worden van de ervaringen met het gebruik van registergegevens op het CBS. De volgende problemen worden herkend bij het gebruiken van 'Big data' voor de officiële statistiek.

Ten eerste, omdat 'Big data' door anderen (vaak private bedrijven) zijn verzameld ontbreekt vaak gedetailleerde kennis over het dataverzamelingsproces (het bron-niveau) en ook over de aanwezige concepten en hun definities (het metadata-niveau) is weinig bekend. Voor registers is hiervoor een kwaliteitskader opgesteld, iets dergelijks is voor 'Big data' nog niet volledig uitgewerkt. Enkele onderzoeksvragen zijn dan ook: hoe meet je de statistische concepten betrouwbaar in 'Big data'? Hoe kun je de validiteit en betrouwbaarheid van metingen in 'Big data' vaststellen? Omdat gebleken is dat de inhoud en samenstelling van 'Big data' bronnen relatief snel kan veranderen, wat tot zogenaamde 'concept-drift' kan leiden, is het belangrijk methoden te ontwikkelen om deze drift te detecteren en ervoor te corrigeren.

Ten tweede, omdat 'Big data' fouten kunnen bevatten is het noodzakelijk deze te herkennen en ervoor te corrigeren. Ook kunnen er gegevens ontbreken (bijvoorbeeld doordat sensoren minder goed of zelfs helemaal niet meer functioneren). Hoe signaleren we deze fouten en hoe corrigeren we daarvoor? Een ander belangrijk punt is het corrigeren voor selectiviteit van 'Big data'. Omdat officiële statistieken altijd over een bepaalde populatie worden gepubliceerd, is onderzoek naar de populatiesamenstelling van (de eenheden in) 'Big data' bronnen een belangrijk onderzoeksonderwerp. Hierbij moet niet alleen gedacht worden aan de representativiteit van de eenheden, maar ook aan het eenduidig identificeren van die eenheden. Dit maakt het trekken van steekproeven niet triviaal. Ook zijn de grote hoeveelheden data in dergelijke bronnen uitermate geschikt voor onderzoek naar bijzondere groepen van eenheden. Kortom, enkele onderzoeksvragen binnen dit onderdeel zijn: wat zijn de eenheden in 'Big data' en hoe corrigeren we voor de selectiviteit van deze eenheden?

Ten derde, 'Big data' bronnen bevatten niet alleen numerieke gegevens maar vaak ook teksten, afbeeldingen of ruimtelijke informatie. Uit deze vormen van 'data' kunnen voor het CBS bruikbare gegevens worden geëxtraheerd. Dit is een relatief nieuw terrein binnen de officiële statistiek, wat het ontwikkelen van nieuwe methodologie noodzakelijk maakt.
Daarbij bevatten veel 'Big data' bronnen zeer grote hoeveelheden gegevens en het is (meestal) noodzakelijk deze gegevens binnen de beveiligde omgeving van het CBS te verwerken en te analyseren. Daarvoor dienen er efficiënte methoden ontwikkeld te worden die dit mogelijk maken. Deze behoeften en eigenschappen zorgen ervoor dat dit werk zich bevindt op het grensvlak van methodologie en IT.

Enkele onderzoeksvragen binnen dit onderdeel zijn: hoe worden ‘Big data’ verzameld en opgeslagen bij het CBS? Welke geavanceerde methoden (bijvoorbeeld uit het werkveld van de ‘Artificial Intelligence’) kunnen worden gebruikt voor het analyseren van en het extraheren van informatie uit ‘Big data’? Denk daarbij aan convolutionele netwerken voor het analyseren van beelden (zoals bijvoorbeeld luchtfoto’s). Enkele relevante onderzoeksvragen zijn: welke methoden lenen zich goed voor het analyseren van beelddata? Welke voorbewerkingsstappen van beelddata zijn noodzakelijk alvorens de data geanalyseerd kan worden? Kunnen we luchtfoto’s gebruiken om bijvoorbeeld het bodemgebruik te herkennen?

Waarschijnlijk de grootste bulk van secundaire data betreft zogenaamde ‘vrije tekst’ (denk aan vacatures, social media berichten of websites van bedrijven). Het CBS heeft afgelopen jaren onderzoek gedaan naar Natural Language Processing methoden, met recentelijk een studie naar de potentie van Large Language Models (LLMs) voor het analyseren van vrije tekst. Kunnen LLMs betrouwbaar statistische classificaties uitvoeren op basis van vrije teksten?
Tenslotte is er nog een schat aan sensor data mogelijk interessant voor de officiële statistiek, zoals aardobservatie data en wegsensoren. Hoe verwerken we dergelijke ‘big data’ op efficiënte manier en wat zijn potentiële toepassingen van aardobservatie binnen de officiële statistiek?

2.3. Data integratie

Bij het samenstellen van statistieken is het niet altijd mogelijk om dit op basis van één enkele bron te doen. Het combineren van verschillende gegevensbronnen (verkregen uit enquêtes, registers of 'Big data') biedt veel meer mogelijkheden. Het integreren van zulke databronnen is echter niet eenvoudig. Dit thema onderzoekt methoden voor het integreren van (micro)data en kwaliteitsmaten voor deze (geïntrigeerde) data. Daarnaast wordt er onderzoek gedaan naar integratie op het (macro)niveau van statistieken. Hierbij worden methoden ontwikkeld om de consistentie te waarborgen tussen statistieken die vergelijkbare concepten behandelen, maar op basis van verschillende databronnen zijn samengesteld.

Niche en urgentie

De afgelopen decennia is het aantal beschikbare databronnen flink toegenomen. Deze data kan verschillende vormen aannemen zoals administratieve-, tekst-, beeld- en sensor-data. Statistieken maken uitsluitend op basis van een enkele van bron is niet altijd reëel, bijvoorbeeld omdat in zo'n bron delen van de populatie ontbreken (selectiviteit) of omdat de gemeten variabelen onvoldoende aansluiten bij de beoogde variabelen (meetverschillen). Echter, door al deze verschillende databronnen te integreren is er veel meer mogelijk. Zo kunnen we mogelijk nieuwe statistische informatie genereren, nieuwe variabelen berekenen, meer gedetailleerde en tijdige output realiseren en bovendien voor subpopulaties die nu nog (deels) buiten de waarneming vallen. Daarnaast kan het gebruikt worden om kwaliteit van de bronnen te toetsen. Binnen de statistische afdelingen van het CBS is er de laatste jaren ook in toenemende mate belangstelling om gebruik te maken van combinaties van bronnen bij het maken van statistieken. Het doel van de afdelingen is om informatie uit meerdere beschikbare databronnen en enquêtes te benutten om zo veel mogelijk dekking te krijgen van de fenomenen die deze statistieken beschrijven. Voorbeelden hiervan zijn de energiestatistieken, zorgcijfers, arbeidsmarktstatistieken en het in kaart brengen van multiproblematiek. Er is ook steeds meer aandacht voor de kwaliteit van de bronnen en het proces van het combineren. Dit thema onderzoekt methoden en kwaliteitsmaten voor het integreren van data. We kijken naar vijf groepen van uitdagingen die we geordend hebben van waarneming naar output. Het gaat om problemen die te maken hebben met (1) de beschikbare bronnen aan waarnemingskant, (2) koppelen van bronnen, (3) gaafmaak- en imputatiemethoden, (4) het maken van schattingen, en (5) onzekerheidskwantificatie.

Uitdagingen

De eerste (groep) uitdagingen richt zich op wat er nodig is aan de waarnemingskant om te bevorderen dat het CBS zoveel mogelijk gebruik kan maken van al beschikbare databronnen. Dit onderdeel is nieuw ten opzichte van het eerdere onderzoeksprogramma. Voor het CBS is niet altijd inzichtelijk welke informatie en databronnen er allemaal al beschikbaar zijn. Dat inzicht is een cruciale randvoorwaarde voor het herkennen van kansen op het gebied van data integratie. Kunnen we, bijvoorbeeld op basis van meta-data, automatisch afleiden welke data integraties op verschillende aggregatieniveaus mogelijk en nuttig zijn? Uit zo'n analyse zou naar voren kunnen komen dat bepaalde output alleen gemaakt wordt wanneer er bepaalde aanvullende informatie beschikbaar is om 'tekortkomingen' op te vangen in de al beschikbare bronnen. Hoe richt je enquêtes in die bedoeld zijn als aanvulling op al beschikbare informatie? Deze onderzoeksvraag willen we samen met thema 'Primaire waarneming' gaan bekijken.

Een tweede uitdaging is dat sommige bronnen lastig of onmogelijk te koppelen zijn op microniveau. Dat kan bijvoorbeeld komen doordat ze geen unieke koppelsleutels hebben (probabilistisch koppelen), of omdat het gaat om twee bronnen met weinig overlappende eenheden (statistisch koppelen), of omdat de eenheidstypen van de twee bronnen onderling afwijken. Dat laatste is bijvoorbeeld het geval bij het identificeren en koppelen van websiteadressen aan het bedrijvenregister. Hoe kunnen we zulke databronnen integreren? Het gaat hier om verbeteringen en uitbreidingen van methoden ten opzichte van het eerdere onderzoeksprogramma.

De derde groep van uitdagingen heeft ermee te maken dat er duidelijk onjuistheden in de beschikbare waarden van databronnen kunnen zitten en dat er ontbrekende waarden kunnen zijn. Binnen het thema richten we ons op het verbeteren en uitbreiden van automatische gaafmaak- en imputatiemethoden. We kijken naar nieuwe mogelijkheden om handmatig gaafmaken te vervangen door automatisch gaafmaken (ook bij een enkele bron). Nieuwe onderzoeksonderwerpen zijn het automatisch gaafmaken over bronnen heen waarbij er verbanden tussen variabelen zijn, en het (tegelijk) imputeren van meerdere variabelen. Daarnaast speelt dat de bestanden/databronnen waar we ons op richten bij het gaafmaken nu soms groter zijn dan ze vroeger waren. Kunnen bijvoorbeeld methoden uit het werkveld van Artificial Intelligence fouten in databronnen automatisch opsporen en corrigeren?

De vierde uitdaging is dat we bij het maken van schattingen op basis van geïntegreerde databronnen vaak rekening moeten houden met bepaalde beperkingen zoals meet-, selectie – of koppelfouten in bronnen. Dit onderzoeksthema onderzoekt correctiemethoden voor bovenstaande beperkingen. Deze methoden worden onderzocht op zowel microniveau (data) als op macroniveau (statistieken). Een onderdeel hiervan is onderzoek naar nieuwe methoden om niet-kanssteekproeven en kanssteekproeven te combineren.

De vijfde uitdaging richt zich op het kunnen kwantificeren van de onzekerheid (vertekening en variantie) van de output ten gevolge van de al genoemde beperkingen van de bronnen en ten gevolge van modelfouten. Doel hiervan is niet om een absoluut getal te kunnen geven van vertekeningen door alle mogelijke fouten, want dat is onmogelijk. Het gaat erom te bepalen hoe effectief bepaalde data integratie methoden zijn in termen van de nauwkeurigheid van de behaalde uitkomsten. Soms is het wenselijk meerdere methoden te vergelijken en te bepalen of de één tot nauwkeuriger resultaten leidt dan de andere. Tot nu toe hebben we vooral gekeken naar het effect van een enkele foutsoort op de nauwkeurigheid, nu gaan we ook kijken naar effecten van meerdere foutsoorten tegelijk zoals meet- en selectiefouten omdat in praktijk bij data integratie vaak meerdere fouten tegelijk een rol spelen.

2.4. Statistisch modelleren

De traditionele manier waarop statistieken bij het CBS worden samengesteld zorgt vaak voor een hoog abstractieniveau en vertraging in het publiceren van resultaten. Het opstellen van statistieken op basis van statistische modellering kan deze tekortkomingen verhelpen. Bovendien zijn deze technieken bijzonder geschikt om nieuwe fenomenen te analyseren. Zo is het mogelijk om iteratief concepten en definities van nieuwe fenomenen vast te stellen en te toetsen met statistische modellen. Dit thema heeft dus twee hoofddoelen: ten eerste het verbeteren van statistische modelleringstechnieken voor het creëren van meer gedetailleerde en actuele statistieken die betrouwbare metingen van bepaalde fenomenen mogelijk maken; en ten tweede het toepassen van deze modellen op complexe fenomenen, met een specifieke focus op economie, welzijn en klimaat.

Niche en urgentie

Veel statistieken op het CBS zijn nog steeds gebaseerd op de traditionele steekproeftheorie. Traditioneel zijn deze schattingsmethoden gebaseerd op het kans mechanisme van het steekproefontwerp. Hierbij wordt vaak gebruik gemaakt van hulpinformatie waarvan de verdelingen in de doelpopulatie bekend zijn uit bijvoorbeeld registraties. Deze schattingsmethodieken worden aangeduid als 'design-based' (ook wel directe schatters genoemd), omdat schattingen alleen gebaseerd worden op steekproefwaarnemingen uit een bepaalde deelpopulatie en verslagperiode.

Directe schatters hebben een grote mate van onzekerheid als de steekproefomvang afneemt en kunnen niet goed omgaan met selectie- en meetfouten. Uitkomsten raken hierdoor vertekend, in toenemende mate door de steeds lager wordende responspercentages. Daarnaast zijn er systematische verschillen in de uitkomsten van een onderzoek ten gevolge van het implementeren van een andere veldwerkstrategie, die worden aangeduid als methodebreuken. Een ander nadeel van de huidige manier van statistieken maken is dat het proces van steekproeftrekken, data verzamelen, verwerken en publiceren tijdrovend en kostbaar is. De klassieke kwaliteitsmaat voor statistieken is de betrouwbaarheid gemeten via de steekproeffout. Echter, tijdigheid, vergelijkbaarheid met uitkomsten uit het verleden en gedetailleerde uitsplitsingen zijn voor veel gebruikers minstens zo belangrijk. Wanneer gedetailleerde uitsplitsingen of schattingen gemaakt moeten worden voor korte referentieperioden is de steekproef vaak te klein om directe schatters te kunnen gebruiken.

Dit onderzoeksthema richt zich op het ontwikkelen van schattingsmethoden die veel sterker gebaseerd zijn op een statistisch model dan de traditionele directe schattingsmethoden. Een belangrijke voorbeeld zijn de zogenaamde Kleine-domeinschatters (KDS). KDS maken het mogelijk om gedetailleerde schattingen te maken die nauwkeuriger zijn dan de directe schatters. Daarnaast is het met KDS mogelijk om statistieken te versnellen en draagt daarmee bij aan één van de strategisch doelen in het Meerjarenplan van het CBS. Het maken van voorlopige schattingen voor een doelvariabele gedurende of vlak na de referentieperiode wordt in de literatuur als 'nowcasten' aangeduid. Verder kunnen modelgebaseerde schattings-methoden ingezet worden om beter te corrigeren voor selectieve non-respons. Methoden die ontwikkeld zijn voor 'informative sampling' en 'non-probability sampling' kunnen vaak beter corrigeren voor de selectiviteit van de toenemende non-respons dan de tot nu toe gebruikte weegmethoden. Met het wegvallen van een register met telefoonnummers dat gebruikt wordt

bij enquêtes is dit onderwerp extra relevant. Tenslotte bieden modelgebaseerde methoden veel mogelijkheden om gebruik te maken van nieuwe databronnen, waarbij survey data nog steeds de primaire bron zijn en nieuwe databronnen als hulpinformatie worden gebruikt. Hierdoor blijft het afbreukrisico van een statistiek beperkt, omdat het CBS nog altijd in controle is over de beschikbaarheid van de primaire data.
KDS hebben de afgelopen jaren zijn meerwaarde bewezen. Zo zijn er methoden ontwikkeld voor het maken van maand en jaarcijfers voor de beroepsbevolking die in productie zijn genomen. Dit geldt ook voor de maandcijfers van het consumenten vertrouwen. Via werk voor derden zijn KDS ontwikkeld voor het maken van trends over mobiliteit, ziekteverzuim en sociale samenhang. Op basis van deze methoden worden momenteel officiële statistieken geproduceerd. Er zijn ook diverse methoden ontwikkeld om methodebreuken te kwantificeren en om hiervoor te corrigeren. Deze technieken zijn in diverse herontwerpen toegepast voor het kwantificeren van methodebreuken. Voorbeelden zijn het effect van diverse herontwerpen van de Enquête Beroepsbevolking op maand en kwartaalcijfers over de beroepsbevolking.

Een ander toepassingsgebied waar het inzetten van schattingsmethoden gebaseerd op een statistisch model essentieel is, is het meten van nieuwe fenomenen en het vullen van lacunes in bestaande (economische) statistieken. Voor veel nieuwe fenomenen zijn er geen of onvolledige (steekproef)gegevens beschikbaar en is een modelmatige aanpak noodzakelijk voor het maken van goede schattingen. Ten tweede vereist het verankeren van deze nieuwe fenomenen in onze statistieken vaak een herziening en herontwerp van de bestaande concepten. Met een modelmatige aanpak is het mogelijk om de nieuwe concepten en definities toe te passen en te testen. We modelleren nieuwe fenomenen, in het bijzonder vraagstukken rond het meten van economie en klimaat. Enkele concrete voorbeelden van nieuwe fenomenen zijn de platform-economie, free services (diensten die gratis zijn, maar waar grote geldstromen achter zitten), arbeid en digitalisering, digitalisering en marktconcentratie, waarde van data, en waarde van (digitale) bedrijven. Ook richten we ons op maten voor brede welvaart en de zogenaamde Sustainable Development Goals (SDG) van de Verenigde Naties. Bij de brede welvaart kijken we verder dan klassieke economische maatstaven en is het doel indicatoren hiervoor in samenhang te meten. Verder wordt onderzoek gedaan rond de European Green Deal, waarbij we in dit thema in het bijzonder werken aan het meten van de klimaatimpact op de economie.

Ook voor het beter meten van bestaande fenomenen kunnen met name statistische modellen uitkomst bieden, zoals bij economische concepten zoals inflatie en prijsindexcijfers. In plaats van enquêtes worden ook hier steeds vaker zeer gedetailleerde transactiedata en scannerdata gebruikt. Het betekent dat prijsindexmethoden moeten worden toegepast die specifiek bedoeld zijn voor ‘Big data’ bronnen. In dit thema doen we onderzoek naar prijsindexmethoden en hun implementatie in productie.

Samenvattend richt dit onderzoeksthema zich op de volgende deelgebieden, te weten (1) onderzoek naar Kleine-domeinschatters voor meer gedetailleerde, snellere schattingen, (2) schatten en corrigeren voor methodebreuken, (3) het ‘nowcasten’ van statistieken, (4) het inzetten van nieuwe databronnen bij schattingen, (5) correctie voor non respons en (6) het meten van nieuwe fenomenen, in het bijzonder rond de economie, welvaart en klimaat.

Uitdagingen

Wat betreft de zes deelgebieden, zien we naar de toekomst toe meerdere uitdagingen.
De gevestigde literatuur over Kleine-domeinschatters (KDS) houdt te weinig rekening met de manier waarop officiële statistieken worden gemaakt, namelijk dat het gaat om herhaaldelijk uitgevoerde steekproefonderzoeken waarop meerdere outputtabellen worden gebaseerd. De vraag is hoe aan de hand van tijdreeksmodellen informatie uit voorgaande referentieperioden en andere domeinen kan worden gebruikt in KDS. Om ervoor te zorgen dat gepubliceerde tabellen zoveel mogelijk numeriek consistent zijn met elkaar is het noodzakelijk en efficiënt om zoveel mogelijk outputtabellen uit één model af te leiden in plaats van één afzonderlijk model voor iedere outputtabel te ontwikkelen. In de literatuur worden methoden beschreven die varianties van de directe schattingen als input gebruiken, maar vooralsnog lijken de resultaten bij KDS instabiel. Dit thema onderzoekt of we de uitkomsten stabiel(er) kunnen maken, bijvoorbeeld door varianties gladde te maken via variantiefuncties?

Wat betreft methodebreuken, constateren we dat de reeds ontwikkelde methoden voor het kwantificeren van methodebreuken vooral goed werken op hoge aggregatieniveaus. Om methodebreuken uit te splitsen naar deelpopulaties kunnen deze technieken worden gecombineerd met de modellen die voor KDS zijn ontwikkeld. In het kader van de komende SBI-herziening (classificatie van bedrijven) is er behoefte aan het terugleggen (‘backcasten’) van de nieuwe indeling. Hoe kunnen we op een verantwoorde manier methodebreuken op een hoog detail niveau detecteren, bijvoorbeeld door bestaande methoden te combineren met KDS?

Wat betreft het “nowcasten” van statistieken zijn nu multivariate, structurele tijdreeksmodellen (STM) ontwikkeld waarbij de doelreeks en hulpreeksen in één model worden gecombineerd. Informatie uit de hulpreeksen wordt gebruikt door in het model de correlatie tussen de trend- en seizoencomponenten van de doelreeks en de hulpreeksen te modelleren. Het standaard lineaire STM veronderstelt dat deze correlaties tijdsonafhankelijk zijn. Dit is een sterke veronderstelling die in de praktijk niet opgaat. Deze methoden kunnen dus niet omgaan met structurele veranderingen in de (cor)relatie tussen een doel- en hulpreeks. Dit thema onderzoekt hoe we correlaties tussen doel- en hulpreeks dynamisch kunnen modelleren? Andere issues met het hiervoor genoemde multivariate STM is hoe om te gaan met grote hoeveelheden hulpreeksen en hoe reeksen die op verschillende frequenties worden waargenomen het beste kunnen worden gecombineerd in één model. Numerieke consistentie tussen schattingen voor verschillende tabellen speelt ook een rol bij het multivariaat STM. Er zal daarom worden onderzocht hoe op basis van één multivariaat STM numerieke consistentie tussen tabellen kan worden afgedwongen door restricties aan het model op te leggen. Dit zal worden toegepast op de kwartaalcijfers van de economische groei.

Het vierde onderzoeksgebied betreft het gebruik van nieuwe databronnen zoals bijvoorbeeld luchtfoto’s, sociale media platforms, internet, maar ook registers. Het gebruik van dit soort databronnen voor het maken van statistieken brengt diverse risico’s met zich mee, zoals nauwkeurigheid, beschikbaarheid en vergelijkbaarheid door de tijd. Deze risico’s kunnen worden beperkt door deze informatie te gebruiken als hulpinformatie in statistische modellen waar survey data nog steeds de primaire databron is, of door deze informatie te gebruiken om het steekproefontwerp of de weging van een survey te optimaliseren.

Het vijfde onderzoeksgebied betreft het meten van nieuwe en bestaande fenomenen, in het bijzonder rond de economie, welvaart en klimaat. De economie verandert continu, en daardoor ontstaan er nieuwe fenomenen die we zo goed mogelijk in kaart willen brengen, en uiteindelijk

in onze statistieken willen opnemen. Dat kan betekenen dat een onderdeel van onze bestaande statistieken anders gemeten moet worden, of dat er nieuwe indicatoren ontwikkeld moeten worden. Allereerst speelt daarbij een meetprobleem: wat is de definitie van het nieuwe fenomeen en lenen de bestaande cijfers zich om dit fenomeen hieruit af te leiden?
Daarnaast moet een model ontwikkeld worden dat dit fenomeen zo goed mogelijk meet. Ieder nieuw fenomeen is anders en idealiter ontwikkelen we hier een generieke aanpak voor.
Naast het verbeteren van onze reguliere economische statistieken werken we aan publicaties die verder kijken dan klassieke economische maatstaven zoals het bruto binnenlands product. In de monitor brede welvaart worden veel indicatoren samengebracht en in samenhang gepresenteerd. Niet alle indicatoren hebben dezelfde tijdigheid, en daarom moeten sommige indicatoren 'genowcast' worden. Een andere vraag is hoe we trends in deze indicatoren meten en weergeven. De trendmethode moet enerzijds de onderliggende lange termijn ontwikkeling weergeven, maar anderzijds ook actuele inzichten geven in de meest recente ontwikkeling. Hier spelen vragen over de techniek, maar ook conceptuele vragen (wat willen we weergeven en hoe willen we indicatoren van het CBS en andere landen onderling vergelijken?)
Binnen dit onderzoeksthema werken we in het bijzonder aan vragen rond klimaatimpact- en adaptatie. We ontwikkelen modellen die de relatie tussen klimaat en economie proberen te meten. Bijvoorbeeld, hoe kunnen we het effect kwantificeren van weersextremen op de economie in een bepaalde verslagperiode? Hoe heeft klimaatverandering door de jaren heen de economie beïnvloed? Gezien de hoeveelheid weersdata die beschikbaar zijn, zijn er raakvlakken met onderzoeksvragen in het thema 'Big data'.

Behalve de omvang (volume) van bepaalde delen van de economie is het voor een goed zicht op economische ontwikkelingen ook belangrijk inzicht te hebben in hoe prijzen zich ontwikkelen (de zogenaamde 'prijzenstatistieken'). Door de veranderende economie en het opkomen van nieuwe fenomenen, treden er regelmatig problemen op rond het meten van deze prijsontwikkelingen en het opnemen van deze nieuwe fenomenen. Denk hierbij aan nieuwe productgroepen die meegenomen moeten worden of aanpassingen van de methode bij een (energie)crisis. De vraag is daarom hoe we de nauwkeurigheid van de prijsindices op korte en lange termijn waarborgen. Een concreet probleem is het beperken van de vertekening bij het aan elkaar knopen van korte indexreeksen.

2.5. Complexiteit en causaliteit

Het CBS streeft ernaar een statistische bijdrage te leveren aan grote maatschappelijke uitdagingen zoals wonen, criminaliteit, klimaat en duurzaamheid. Dergelijke fenomenen zijn van nature complex en onderling verbonden. Gewoonlijk publiceert het CBS-statistieken los van elkaar, zonder rekening te houden met de onderlinge interacties tussen deze fenomenen. Dit onderzoeksthema richt zich op methoden die in staat zijn om statistieken samen te stellen die fenomenen beschrijven als een complex systeem, waarbij hun interacties worden meegenomen. Daarnaast worden methoden onderzocht die causale verbanden tussen de verschillende fenomenen kunnen beschrijven.

Niche en urgentie

De behoefte aan kwantitatief inzicht in de huidige complexe samenleving vraagt om nieuwe methodologische instrumenten. Het thema 'Complexiteit en causaliteit' voorziet daarin. Sinds de oprichting van het CBS in 1899 is de wereld veel ingewikkelder en meer verknoopt geworden. De moderne maatschappij kan gezien worden als een verzameling van complexe systemen waarbij de verschillende actoren (zoals personen, bedrijven, instellingen, voertuigen, etc.) meer met elkaar in contact staan dan ooit. Met andere woorden, sociale, maatschappelijke, logistieke en economische processen grijpen meer dan voorheen op elkaar in en zijn met elkaar verbonden. Het analyseren en beschrijven van de fenomenen die voortkomen uit de interactie van deze complexe systemen vraagt van het CBS extra inzet.
Eén van de inhoudelijke strategische doelen van het CBS voor de komende jaren, zoals geformuleerd in het CBS -eerjarenprogramma 2024–2028, is het leveren van een statistische bijdrage aan grote maatschappelijke opgaven zoals wonen, ondermijnende criminaliteit, klimaat, globalisering en duurzaamheid (zoals hierboven beschreven). Dit soort maatschappelijke fenomenen zijn zonder uitzondering complex: er is sprake van verschillende actoren en interactie tussen verschillende fenomenen. In de analyses van het CBS worden die interacties nu doorgaans niet expliciet meegenomen. Voor beleidsmakers en journalisten, belangrijke afnemers van CBS-statistieken, zou het weglaten van deze interacties een probleem kunnen vormen. Ter illustratie, cijfers over fietsongevallen naar leeftijd en cijfers over vergrijzing per regio en mobiliteit worden gepresenteerd in afzonderlijke StatLine tabellen. Het wordt hieruit niet duidelijk of een potentiele toename van verkeersdoden op de fiets te wijten is aan vergrijzing van de bevolking, dat fietsen onveiliger geworden is of dat ouderen meer fietsen. Voor beleidsmakers is het van wezenlijk belang of beleidsterreinen elkaar tegenwerken of juist versterken, hoe ze elkaar beïnvloeden en wat de kwantitatieve aspecten daarvan zijn. Voor journalisten is het ook van belang om cijfers in een juiste context te plaatsen. Het CBS kan deze afnemers ondersteunen door meer context bij statistieken te kwantificeren en bekende en aanwezige causale relaties te toetsen en beschrijven.
De wetenschappelijke discipline die zich bezighoudt met het analyseren en modelleren van complexe systemen heet 'Complexity Science'. Methoden voor het analyseren van complexe systemen zijn grofweg te groeperen in de volgende drie werkgebieden: 'Network Science' (NWS), 'Agent Based Modeling' (ABM), en 'Dynamic Systems' (DS).
Ten eerste, bij NWS ligt de nadruk op de (evoluerende) structuur van het systeem, zoals sociale netwerken, verkeer, transport en productieketens. Wat zijn belangrijke elementen en verbindingen/interacties in het systeem, hoe beïnvloeden elementen elkaar en wat heeft dat voor gevolgen? Bijvoorbeeld, hoe afhankelijk is het Nederlandse bedrijfsleven van gas of andere importgoederen? Hoe werkt dat door in de Nederlandse productieketen? Welke economische

sectoren zijn gevoelig omdat er weinig leveranciers zijn? Een reeds succesvolle toepassing van deze techniek zijn de Personen- en Bedrijvennetwerken die de afgelopen jaren zijn ontwikkeld en die zowel intern (bij statistische afdelingen, o.a., in de vorm van dashboards en artikelen over opleidingsniveau- en herkomstsegregatie) als extern (academia) veelvuldig gebruikt worden.
En ten tweede, bij ABM ligt de nadruk op het individuele gedrag van de elementen. Welk gedragsregels leiden tot ander systeemeigenschappen? Bijvoorbeeld, leidt een ZZP-belastingvoordeel tot meer ZZP'ers en ten koste van wat?
Ten derde, DS legt de nadruk op veranderende eigenschappen van het systeem: hoe beïnvloedt het systeem zichzelf? Bijvoorbeeld, wordt in een buurt met veel zonnepanelen ook sneller overgegaan op zonnepanelen door anderen (zelf versterkend effect). Hoe beïnvloeden onderdelen (groepen van elementen) van het systeem elkaar?

Het onderzoeksthema Complexiteit en Causaliteit heeft als doel om methoden te ontwikkelen en onderhouden die (1) leiden tot een uitbreiding van de output van het CBS met statistieken die fenomenen van Nederland beschrijven als een complex systeem, (2) verder inzicht geven in de causaliteit bij complexe mechanismen ter ondersteuning van de duiding van gepubliceerde statistieken en (3) het uiteindelijk ook mogelijk maken voor (interne en externe) onderzoekers om beleid te evalueren door middel van met CBS-cijfers-gekalibreerde scenario's.

Uitdagingen
Het onderzoekthema Complexiteit en Causaliteit heeft een goede basis opgebouwd in de afgelopen jaren, met als succesvolle toepassingen de eerdergenoemde personen- en bedrijvennetwerken. Er liggen voor het CBS echter nog meerdere uitdagingen en kansen.
Wat betreft het publiceren van statistieken op basis van complexe (netwerk) systemen, zijn er verschillende aanknopingspunten. Ten eerste, het combineren van steekproefdata met netwerkstructuren (uit het werkgebied 'Network Science') is nog niet of nauwelijks onderzocht. Is een bedrijvennetwerk te combineren met bestaande enquêtes op een dusdanige manier dat dit extra informatie oplevert? Is een personennetwerk te gebruiken als een steekproefkader voor onderzoek naar bijvoorbeeld de invloed van het sociale netwerk van jonge ouders op hun keuze voor kinderopvang? Om dat te kunnen bepalen, kan je een steekproef trekken, maar dan wel een steekproef waarbij de kans groot is dat je slechts een deel van het netwerk van (sommige) jonge ouders waarneemt. Het is op dit moment onduidelijk hoe je dat op een methodologisch verantwoorde wijze doet.
Ten tweede, de personen- en bedrijvennetwerken zijn grotendeels gebaseerd op registerdata waarbij de data niet altijd toereikend en/of compleet is. Zo worden bijvoorbeeld relaties tussen entiteiten in een netwerk deels geschat of afgeleid. Wanneer mogen we afleiden (en met welke zekerheid) dat bedrijven met elkaar handelen zodat daar goede statistische conclusies uit getrokken kunnen worden? Hoe weten we of personen sociaal gezien met elkaar in contact staan? Ten derde, zoals in de inleiding werd benoemd, is de huidige maatschappij complex en gaat deze veel verder dan alleen bedrijven en personen die onderling met elkaar interacteren. Het CBS kan onderzoek doen naar meer netwerkstructuren zoals transport- en energienetwerken. Welke bij het CBS beschikbare databronnen en technieken lenen zich voor het implementeren van deze nieuwe netwerken? En wat zijn de toepassingen van deze nieuwe netwerken in de context van de strategische doelen uit het meerjarenplan? Ten vierde, de bestaande netwerken (en de potentieel nieuwe) netwerken beschrijven interacties tussen specifieke type actoren (bijvoorbeeld, personen, bedrijven of verkeer), echter werd juist betoogd dat de maatschappij bestaat uit interacties tussen allerlei typen actoren. Deze interacties worden op dit moment nog niet gemodelleerd. Kortom, naast het verbeteren van

bestaande en implementeren van nieuwe netwerken, liggen er ook kansen bij het combineren van verschillende typen netwerken. Echter, methoden voor het combineren van verschillende type netwerken, zoals het bedrijven-, wegennetwerk- en personennetwerk, staan in de kinderschoenen. Tenslotte zien we als vijfde uitdaging het afleiden van robuuste en zuivere statistieken op basis van netwerken. De complexiteit zit hier in de onzekerheid over de netwerkstructuur zelf, alsmede nauwkeurigheid van de statistieken zelf. Hoe leiden we de statistieken af uit netwerk/complexe systemen, en zijn de onzekerheidsmarges te bepalen?

Ten behoeve van een juiste duiding van gepubliceerde, complexe, statistieken zien we het aantonen van causale complexe inferenties als belangrijke uitdaging. Hoe modelleer je causale relaties in complexe systemen? Hoe combineer je onafhankelijk gemeten data voor verschillende fenomenen op een verantwoorde wijze tot een causaal model?

Voor het derde onderzoeksdoel, het ondersteunen van beleidsevaluaties, zien we ook meerdere uitdagingen. Het doorrekenen van verschillende scenario's kan met technieken uit het werkgebied 'Agent Based Modeling' (ABM). Echter, het inrichten van ABM-modellen is complex. In academia werken ze vaak met kleine, eenvoudige voorbeelden om te kijken wat de effecten zijn van bepaald gedrag. Zoals eerder benoemd, streeft het CBS naar inzichten voor grote maatschappelijke thema's en is het opstellen van een ABM vele malen complexer. Enkele vragen daarbij zijn: hoe ontwerp je ABM voor een zogenaamde Twin-populatie (een digitale representatie van een proces) waarin de gehele Nederlandse bevolking als actor meedraait voor het testen van scenario's (bijvoorbeeld het verspreiden van infectieziekten, of het overgaan tot aanschaf van een elektrische auto)? De uitdaging hierbij is de schaalgrootte en het formuleren van realistische interactie scenario's. Verder, hoe kalibreer je ABM-modellen op basis van een gehele populatie (bijv. personen/bedrijven)? Tenslotte, een relatief nieuwe invalshoek voor dit probleem: gegeven de (macro)uitkomst die het CBS kent, hoe stel je daarmee een ABM-model af? Wat is een plausibel bijbehorend gedrag van agents, gegeven de bekende uitkomst (op basis van CBS-statistieken)?

2.6. Informatie beveiliging

Het vertrouwelijk omgaan met gegevens is wettelijk verplicht volgens nationale en internationale regelgeving en vormt een essentiële voorwaarde voor het CBS om het vertrouwen van de samenleving en beleidsmakers te behouden. Aan de andere kant is het onwenselijk om gegevens volledig af te schermen en ontoegankelijk te maken voor anderen: een van de strategische doelen van het CBS is om de rol te pakken van 'data-hub van Nederland'. Dit onderzoeksthema richt zich op methoden om data te beveiligen en om data veilig te delen met externe partijen, waarbij het risico op gegevensonthulling wordt afgewogen tegen het behoud van informatiewaarde.

Niche en urgentie

Vertrouwelijk omgaan met gegevens van individuele eenheden (personen, bedrijven, huishoudens, instituten, gemeenten, etc.) is een van de voorwaarden voor een betrouwbaar CBS. Niet voor niets staan er in de CBS-wet meerdere artikelen die aangeven hoe het CBS vertrouwelijk met gegevens om moet gaan. Zo staat in artikel 37 van de CBS wet bijvoorbeeld dat uit publicaties van het CBS geen herkenbare gegevens over een afzonderlijk persoon, huishouden, onderneming of instelling ontleend mogen kunnen worden. Ook de Europese GDPR (in Nederland uitgewerkt in de AVG en de uAVG) verplicht ons om nauwgezet met gegevens van individuele personen om te gaan.

In de CBS-wet staat ook te lezen dat het CBS de publieke taak heeft om statistische gegevens te publiceren over de Nederlandse samenleving en wetenschappelijk en statistisch onderzoek te faciliteren. Bovendien is er een (internationale) beweging die data delen stimuleert tussen organisaties. Met de komst van de European Data Act moet het makkelijker worden om data te delen en gezamenlijk onderzoek te doen, bijvoorbeeld via (inter)nationale 'Data Spaces' (een centrale plaats voor het delen van data tussen organisaties). Data delen is ook onderdeel van de strategische doelen van het CBS (toegang tot data vergroten). Het begrip "data" moet hier in brede zin gezien worden: zowel geaggregeerde data (o.a. via StatLine) als ook microdata (o.a. via aanvullende statistische diensten en via remote access) vallen hieronder.
Het is van belang om de beveiliging van informatie en het nut van informatieverstrekking op een verantwoorde manier tegen elkaar af te wegen. Dit onderzoeksthema houdt zich dan ook bezig met de afweging van het risico op onthulling versus het behoud van informatie. Iedere mate van beveiliging betekent immers tegelijkertijd ook een bepaalde hoeveelheid van informatieverlies.

Het CBS is al jaren voortrekker binnen de statistische bureaus van onderzoek op het gebied van informatiebeveiliging. Sinds de jaren 90 van de vorige eeuw is het CBS de coördinator van vele internationale projecten, deels gefinancierd door Eurostat en/of de Europese Commissie. Via die projecten is het CBS dan ook al lange tijd coördinator van een Europees 'Center of Excellence on Statistical Disclosure Control'. Een belangrijk product van die projecten is de software die de Europese standaard is geworden voor statistische bureaus om statistisch beveiliging toe te passen op hun publicaties. Een aantal van de geïmplementeerde methoden zijn door het CBS ontwikkeld. Op deze manier wordt het onderzoek op het gebied van de informatiebeveiliging door de afdeling Methodologie van het CBS niet alleen binnen de eigen organisatie gebruikt, maar ook bij vele internationale instellingen en statistische bureaus. Ten slotte is het CBS ook actief lid van de 'Expert Group Statistical Disclosure Control' van Eurostat.

Er ligt dus een stevig fundament wat betreft informatiebeveiliging, maar door een snel veranderende wereld (technologisch, cultureel) zijn er nieuwe uitdagingen voor toekomstig onderzoek. Het onderzoek binnen dit thema heeft tot doel om informatiebeveiliging op een hoog niveau te houden, gegeven de veranderende omgeving. Dit betekent concreet dat het onderzoeksthema onderzoek doet naar (1) het beveiligen van gecombineerde en nieuwe typen informatie, (2) het veilig delen van informatie zowel extern als intern en (3) bijdragen aan de onderbouwing van beleid op het gebied van informatiebeveiliging.

Uitdagingen

Onderzoek binnen dit thema heeft tot doel om informatiebeveiliging op een hoog niveau te houden, en dit kent meerdere uitdagingen.

Ten eerste, door toenemende mogelijkheden om data en informatie te combineren wordt ook het risico op onthulling van individuele informatie vergroot. Nieuwe soorten data en informatie (denk aan de eerder benoemde netwerkdata, ongestructureerde data, 'non-probability samples', 'Big data', visualisaties, etc.) kunnen tot nieuwe soorten output leiden. Hoe kunnen we de veranderende onthullingsmogelijkheden het hoofd blijven bieden? Hoe beveiligen we dit soort nieuwe typen data en welke methoden lenen zich hiervoor?
Ook zijn begrippen als onthulling en privacy niet altijd duidelijk of eenduidig gedefinieerd voor dergelijke nieuwe soorten data en informatie. Over welke eenheden gaat de ongestructureerde data? Welke populatie zit achter de 'non-probability samples'? Hoe definieer je onthullingsrisico in netwerkdata? Hoe neem je de complexiteit van netwerkdata mee? Wat betreft netwerkdata wordt er samengewerkt met het thema 'Complexiteit en causaliteit'. Concluderend, het gaat bij dit onderzoeksgebied niet alleen om te onderzoeken wat nu nog niet mogelijk is (en dat in de toekomst wel mogelijk te maken), maar ook om te onderzoeken hoe bepaalde bestaande activiteiten mogelijk kunnen blijven.

Ten tweede, de trend om informatie te delen met externe onderzoekers is niet nieuw. Echter moet dit mogelijk blijven binnen de wettelijke kaders en veranderende omstandigheden. Gegeven het feit dat er steeds meer mogelijkheden komen om bestaande beveiligingsmethoden te omzeilen of aan te vallen, moeten de informatiebeveiligingsmethoden daarop reageren.
Nieuwe methoden en technieken kunnen ook bijdragen aan het (her)ontwerpen van onze interne statistische processen. Ontwerpen en testen van (delen van) productiestraten vindt vaak plaats in een ontwikkelomgeving waar geen "echte" data mogen staan. Echter, "echte" data hebben karakteristieken die gefabriceerde data niet hebben.
Vragen die dan ook spelen zijn: hoe kunnen we op een veilige manier samenwerken met externen waarbij informatie en data op een veilige manier gecombineerd worden? Hoe kunnen we bruikbare en veilige data genereren die zonder problemen gedeeld kunnen worden met derden? En hoe kunnen vervolgens dergelijke databestanden gebruikt worden bij het ontwerpen en testen van productiestraten?

Ten derde, indirect kan onderzoek binnen dit thema ook bijdragen aan onderbouwing van (nieuw) beleid aangaande informatiebeveiliging. Het huidige beleid heeft deels een subjectief karakter, gebaseerd op ervaring, expert kennis en inschattingen. Dit thema kan bijdragen aan het explicieter en objectiever maken van het beleid. Ook de veranderende omgeving kan en zal van invloed zijn op het beleid. Om die veranderingen in het beleid goed te kunnen duiden, is onderzoek nodig. Ten slotte speelt de manier waarop gecommuniceerd wordt over de statistische beveiliging een rol in dit thema. Hoe leg je voor verschillende doelgroepen uit dat de

gebruikte methoden voldoende beveiliging bieden? Hoe leg je uit aan verschillende doelgroepen hoe ruis toevoegen kan helpen bij het waarborgen van privacy?
Tenslotte, informatiebeveiliging heeft een brede scope en heeft in principe een link met alle andere thema's uit het CBS-onderzoeksprogramma. Alles krijgt op een bepaald moment (indirect) met informatiebeveiliging te maken. Al is het maar via het 'privacy by design' principe dat in de AVG/GDPR wordt genoemd. Dat gezegd, dit onderzoeksthema richt zich alleen op bovenstaande methodologische aspecten en gaat bijvoorbeeld niet in op IT-technische aspecten. Voornamelijk bij onderwerpen als 'Secure Multi Party Computation' en 'Federated Learning' spelen IT-technische/implementatie aspecten een belangrijke rol. Deze IT-technische component valt echter buiten de scope van dit onderzoeksthema.

2.7. Statistische informatie communicatie

Het CBS doet er alles aan kwalitatief goede statistieken te produceren: in elke stap van het productieproces staat kwaliteit bovenaan. De laatste stap in dat proces is het publiceren van statistische informatie op de CBS-website: de communicatie over onze statistische informatie aan een algemeen publiek. Ook hier wordt veel zorg aan besteed. Dit onderzoeksthema richt zich niet op de output zelf, de statistische informatie, maar op de manier waarop het CBS hierover communiceert: de statistische informatie communicatie. In het bijzonder worden methoden ontwikkeld om de interpretatie, gebruik en waardering van onze communicatie door gebruikers empirisch te onderzoeken. Ook worden kwaliteitscriteria voor interpretatie en begrip ontwikkeld. Op basis van verkregen inzichten kan vervolgens bekeken en getoetst worden hoe de interpretatie en begrip vergroot kan worden. Ook wordt onderzocht hoe concepten als onzekerheid op een begrijpelijk wijze overgebracht kunnen worden. Dit onderzoek is daarmee aanvullend op het onderzoek dat reeds door CCN gedaan wordt naar esthetische en 'usability & user-experience' aspecten van onze communicatie. Daar waar raakvlakken zijn, trekken we in dit thema samen op met CCN. Met dit thema geeft het CBS mede richting aan ontwikkelingen in de Europese Unie op dit gebied.

Niche en urgentie

Het CBS publiceert dagelijks statistische informatie. Dat gebeurt in vele vormen: persberichten, Corporate artikelen op de CBS-website, rapporten, tabellen in Statline, etc. Het CBS doet dat niet voor zichzelf, maar ten behoeve van een groot publiek dan wel specifieke gebruikers. Dat kan een geïnteresseerde burger zijn, een student die gegevens zoekt voor een scriptie, een journalist, een onderzoeker, een ambtenaar, een Tweede Kamer lid, etc. Allemaal hebben ze hun eigen wensen, en voor allemaal is het belangrijk dat ze onze publicaties (in welke vorm dan ook) goed kunnen gebruiken en interpreteren.

Het CBS besteedt veel aandacht aan de kwaliteit van de te publiceren statistische cijfers. In alle stappen van het productieproces is kwaliteit een belangrijk thema. Onze dataverzameling-, steekproef-, verwerkings- en schattingsmethodes zijn erop gericht dat onze statistieken 'zuiver en precies' zijn: dat wil zeggen dat ze een fenomeen zo adequaat mogelijk beschrijven, dus met minimale vertekening en ruis. Ook aan de kwaliteit van onze communicatie over statistische informatie wordt veel aandacht besteed. Kwaliteitsaspecten gerelateerd aan de 'user-experience' en andere meer esthetische aspecten van onze output-communicatie krijgen al veel aandacht van CCN middels onderzoek en inhoudelijke expertise op dit terrein. Met het onderzoeksthema "Statistische Informatie Communicatie" breiden we het doen van systematisch, empirisch onderzoek uit naar kwaliteit van deze laatste stap in het productieproces: interpretatie, gebruik en waardering van onze communicatie. We focussen ons daarmee op de inhoud van de communicatie. Dit sluit aan bij de 'European Statistics Code of Practice' (EU 2018). Naast de kwaliteitscriteria voor de statistische informatie op zichzelf (zoals relevantie, tijdigheid, etc.) wordt in Principe 15 (Toegankelijkheid en Duidelijkheid) gesteld dat "... statistics are presented in a clear and understandable form". Wat daarmee wordt bedoeld wordt echter niet aangegeven. Dit onderzoeksthema wil hier antwoorden op geven samen met CCN. Onderzoek naar de communicatie over statistische informatie krijgt in de Europese Unie steeds meer aandacht, wat o.a. blijkt uit congressen op dit gebied.

Het onderzoek in dit thema is drieledig, te weten (1) het operationaliseren van de kwaliteit van statistische informatie communicatie in kwaliteitscriteria voor verschillende gebruikersgroepen (2) methode-ontwikkeling om kwaliteit te onderzoeken, en (3) indien de kwaliteit ontoereikend blijkt, op basis van gedefinieerde kwaliteitscriteria en verkregen inzichten het onderzoeken van manieren om de statistische informatie communicatie aan te passen en te toetsen. De focus die we hierbij aanbrengen is toepassingsgericht: empirisch onderzoek waarin we de gebruiker actief betrekken.

In het verleden is bij het CBS al onderzoek gedaan naar met name grafieken. Beschreven werd welke vormen van grafieken in theorie lastig zijn te interpreteren. Er werden echter geen onderzoeksmethodes beschreven om de kwaliteit van grafieken empirisch te onderzoeken; ook zijn de beschreven criteria niet getoetst op factoren die samenhangen met interpretatie, zoals bijvoorbeeld duidelijkheid en begrijpelijkheid. Het CBS heeft de afgelopen decennia in het thema “Primaire Waarneming” veel ervaring opgedaan met het systematisch empirisch onderzoeken van de begrijpelijkheid en beantwoordbaarheid van vragen in vragenlijsten door respondenten. De ontwikkelde onderzoeksmethodes en de opgedane ervaring worden in dit onderzoeksthema ingezet. Daarnaast kunnen we te raden gaan bij meerdere kennisgebieden die onderzoek doen naar informatieverwerking: welke onderzoeksmethodes worden gebruikt en wat zijn de resultaten van dat onderzoek?

Uitdagingen

Onderzoek naar de kwaliteit van statistische informatie communicatie kent verschillende uitdagingen. Zoals geldt dat onze statistische informatie ‘zuiver en precies’ moet zijn, zouden we bijvoorbeeld ook kunnen stellen dat onze publicaties correct moeten kunnen worden geïnterpreteerd, gebruikt en gewaardeerd.
Over deze drie aspecten kunnen we ons de volgende vraag stellen:

- Interpreteren gebruikers de gepresenteerde informatie op een manier zoals wij dat willen, op een manier die overeenkomt met onze intentie? Als dat zo is zouden we kunnen zeggen dat er sprake van valide communicatie.
- Is de communicatie (qua inhoud en vorm) zodanig gekozen dat de gepresenteerde informatie door een gebruiker is te gebruiken voor het door deze gebruiker beoogde doel? Dan zouden we kunnen zeggen dat de communicatievorm effectief en doeltreffend is.
- Waarderen gaat over een waardeoordeel van de gebruikers: waarderen gebruikers onze communicatie?

Daarnaast zijn de aantrekkelijkheid en toegankelijkheid van een publicatie kwaliteitscriteria die ook expliciet kunnen worden gebruikt: een ‘infographic’ moet ook aantrekkelijk zijn om te bekijken (en daarmee de aandacht trekken); een lange blok tekst is mogelijk niet voor iedereen aantrekkelijk om te gaan lezen. Naar deze aspecten wordt al onderzoek gedaan door CCN. Er zijn dus meerdere kwaliteitscriteria, waarbij we ons er bewust van moeten zijn bepaalde criteria met elkaar op gespannen voet kunnen staan en conflicteren. Het vinden van de juiste criteria (objectieve maatstaven), het conceptualiseren en operationaliseren hiervan, en het toetsen van onze communicatie hieraan is een uitdaging.

Naast het bepalen van kwaliteitscriteria op zichzelf is het een uitdaging om deze empirisch te onderzoeken. Hier gaat het om methode-ontwikkeling. De vraag is: hoe kunnen we bijvoorbeeld onderzoeken of gebruikers een communicatievorm (een visualisatie of een tekst) correct interpreteren? Welke onderzoeksmethoden zijn hiervoor geschikt? Naast kwalitatieve onderzoeksmethoden (diepte-gesprekken en eye-tracking) kunnen ook kwantitatieve methoden ingezet worden, bijvoorbeeld met experimenten.

Samenvattend komen we voor dit thema tot de volgende centrale onderzoeksvragen: Hoe kunnen we de kwaliteit van onze statistische informatie communicatie met betrekken tot interpretatie, gebruik en waardering onderzoeken? Welke kwaliteitscriteria kunnen we hieruit destilleren (voor verschillende doelgroepen van gebruikers)? Welke kennisgebieden zijn hierbij ondersteunend, en wat zeggen die disciplines over de kwaliteit van communicatie over statistische informatie? Welke kennis hebben we al in huis, en welke moeten we ontwikkelen? Tevens doen we onderzoek naar de communicatie van specifieke statistieken en statistische concepten zoals bijvoorbeeld onzekerheid van cijfers gebaseerd op steekproeven en uitkomsten van complexe analyse methoden.

2.8. Toepasbare Artificial Intelligence

Recente ontwikkelingen in Artificial Intelligence (AI) kunnen potentieel interessant zijn voor het gebruik in de officiële statistiek. Het toepassen van AI brengt echter ook risico's met zich mee. AI-methoden zijn minder goed begrepen en bieden minder kwaliteitsgaranties vergeleken met standaard statistische methoden. Het thema 'Toepasbare AI' richt zich op het ontwikkelen van methodologische richtlijnen voor het succesvol inzetten van AI binnen de officiële statistiek. In tegenstelling tot andere onderzoeksthema's is dit thema niet beperkt tot één of enkele stap(pen) in het statistische proces. AI wordt beschouwd als een veelzijdig hulpmiddel dat in het gehele proces kan worden ingezet.

Niche en urgentie

Artificial Intelligence (AI) is een onderzoeksgebied dat zich richt op het ontwikkelen van tools die in staat zijn taken uit te voeren die normaal menselijke intelligentie vereisen, zoals redeneren, probleem oplossen, perceptie en taalbegrip. De oorspronkelijke vraag van AI-onderzoekers was: kunnen we machines laten denken? We zien dat een specifieke techniek, Machine learning (ML) het succesvolst is. ML probeerde oorspronkelijk de vraag te beantwoorden hoe mensen leren. Aan de ene kant werden hiervoor cognitieve modellen ontwikkeld, terwijl aan de andere kant de vraag ontstond hoe hersenen zich organiseren en zo informatie kunnen vasthouden. Dit leidde tot de ontwikkeling van o.a. neurale netwerken. Het huidige veld van de ML heeft nog maar weinig te maken met deze oorspronkelijke ambities. De toepassingen van AI buiten het CBS nemen in een rap tempo toe, en ook binnen het CBS stijgt het aantal initiatieven. Toch merken we dat, binnen de statistiek, het aantal succesvolle toepassingen in productie beperkt is. Veel van de initiatieven lijken hoopgevend, maar uiteindelijk zijn er onzekerheden of aan de kwaliteitseisen van de officiële statistieken wordt voldaan. Het toepassen van AI bij het CBS blijkt dus ingewikkeld. Een probleem is dat de AI vooral aan de kant van de techniek sterk ontwikkelt. Ter illustratie, universiteiten en grote technologiebedrijven publiceren frequent nieuwe technieken en modellen die volgens bepaalde maatstaven steeds beter worden. Echter, deze maatstaven staan niet perse gelijk aan de maatstaven in de officiële statistiek (lees: het verkrijgen van een statistisch zuiver resultaat). Dit is dan ook het expliciete doel van dit thema, het formuleren van methodologische richtlijnen voor het succesvol toepassen van AI in de officiële statistiek door (1) te onderzoeken welke problemen er, vanuit een statistisch oogpunt, ontstaan als we AI inzetten ('normatief kader'), (2) onderzoeken hoe we deze problemen kunnen oplossen ('methodologisch kader') en tenslotte (3) ondersteunende werkzaamheden bij specifieke AI implementaties bij het CBS ('toepassingsgericht'). Het zwaartepunt van het onderzoek in dit onderzoeksthema ligt op het methodologische kader. Met AI wordt al breed geëxperimenteerd binnen het CBS, maar de kennis is momenteel verspreid over verschillende divisies en onderzoeksthema's. Dit onderzoeksthema beoogt een geïntegreerde aanpak om te bepalen wanneer en hoe specifieke AI-technieken verantwoord kunnen worden ingezet binnen de officiële statistiek. We zullen hierbij concrete casestudies gebruiken binnen het CBS om de praktische toepasbaarheid van deze technieken te evalueren. Hierbij ontstaat er uiteraard een wisselwerking tussen Methodologie en de statistische afdelingen.

Uitdagingen

Het succesvol toepassen van AI heeft een aantal uitdagingen die we vanuit drie verschillende perspectieven belichten.
Ten eerste het normatief kader van AI. De centrale vraag hierbij is: Waar moet AI aan voldoen om succesvol binnen de officiële statistieken te worden ingezet? Voor de officiële statistieken komen we dan ook snel bij de 'European Statistics Code of Practice' ('European Statistical System Committee', 2017) uit. Deze 'code of practice' beschrijft de normen en richtlijnen voor de kwaliteit en integriteit van statistische systemen binnen de Europese Unie. Het biedt een raamwerk voor de waarborging van de onafhankelijkheid, betrouwbaarheid, en transparantie van statistische gegevens en processen. Door verschillende groepen werd beschreven wat de gevolgen zijn van dit kwaliteitskader voor de inzet van, specifiek, ML binnen de officiële statistieken. Binnen het CBS heeft dit op het gebied van ML-methodologie geleid tot het 'Total Machine Learning Error Model'. Dit model brengt voor ML in kaart welke fouten er kunnen optreden als we dergelijke methoden toepassen in de officiële statistiek. Ook voor alle andere onderwerpen binnen de AI zouden we, als we een normatief kader hebben, een soortgelijk methodologisch kader kunnen maken.
Bovendien bestaat er een grote overlap tussen de toepassing van dit kwaliteitskader voor AI en de verschillende uitwerkingen van 'Responsible AI' door derden. De overlap bestaat vooral in de focus op het gebied van transparantie, eerlijkheid en verantwoordelijkheid, maar er zijn ook duidelijke verschillen. Zo ligt de focus bij de 'European Statistics Code of Practice' vooral op betrouwbaarheid en kwaliteit van statistieken, terwijl het 'Responsible AI' raamwerk ook op de maatschappelijke en juridische consequenties ingaat.
Enkele vragen die er zijn: wat is de overlap tussen de 'code of practice' en het 'Responsible AI' raamwerken en waar zitten de verschillen? Wat kunnen we hiervan leren met betrekking tot het verantwoord gebruiken van AI binnen de officiële statistieken? En daaruit volgend: aan welke kwaliteitseisen moet een AI-model voldoen om geschikt te zijn voor officiële statistieken?

Ten tweede het methodologische kader van AI. Terwijl het normatieve kader vooral op de randvoorwaarden ingaat, gaat het methodologisch kader veel meer over hoe we deze normen kunnen bereiken. Zoals al eerder gezegd is AI de afgelopen jaren sterk ontwikkeld. Technologisch is er dus ook al veel mogelijk. De (statistisch) methodologische achterstand zal dus moeten worden ingehaald en er zal moeten worden gekeken of de juiste technieken hiervoor beschikbaar zijn. Om dit te bereiken, moeten we onderzoeken hoe we AI-technieken kunnen gebruiken om aan onze eisen te voldoen. Hiervoor denken we aan technieken zoals 'Large Language Models' voor tekstanalyse en informatie-extractie, 'Generative adversarial networks' en 'Generative neural networks' voor bijvoorbeeld imputatie en synthetische datasetgeneratie, en bijvoorbeeld 'Random forests', 'Support vector machines' en 'Neurale netwerken' voor classificatieproblemen. Deze technieken zijn voor een groot deel niet ontwikkeld met het doel een statistisch zuiver resultaat te krijgen. De centrale vraag is dan ook wat is er nodig om AI-modellen aan het normatief kader te laten voldoen en daarmee de kwaliteit te waarborgen? Het methodologische kader gaat in op vragen die betrekking hebben op, onder andere, representativiteit, externe validiteit, verklaarbaarheid en vertekening van modellen. Dit zijn onderwerpen die specifieke onderzoeksvragen opleveren met betrekking tot methodologie en die een betere invulling kunnen geven op de onderwerpen die binnen het kwaliteitskader worden aangestipt. Enkele concrete onderzoeksvragen zijn: hoe ga je om met ongebalanceerde datasets bij een classificatieprobleem? Hoe krijg je 'kansen' uit een classifier en kunnen deze gebruikt worden bij het prioriteren van handmatig werk in een human-machine feedback loop? Hoe verklaar je waarom een 'Classifier' een bepaalde uitkomst geeft?

Verder zien we uitdagingen bij het implementeren van specifieke AI-oplossingen, kortom het daadwerkelijk toepasbaar maken van AI bij het CBS. De toepassingen, zoals ze worden uitgewerkt binnen de verschillende thema's en binnen de statistische afdelingen, zijn uiteindelijk allen onderhevig aan de vragen die betrekking hebben op het normatieve en methodologische kader van het gebruik van AI binnen de officiële statistieken. Zolang we dus niet met zekerheid kunnen zeggen dat de kwaliteit van de te gebruiken modellen hoog genoeg is, zullen al deze onderzoeken niet tot concrete toepassingen leiden. Vanuit het methodologisch kader zullen generieke oplossingen worden ontwikkeld voor het toepasbaar maken van AI. Toch zullen er tijdens de toepassing specifieke vragen opkomen die afhangen van het doel waarvoor AI wordt ingezet. Bijvoorbeeld, het gebruik van AI bij gaafmaken zal andere vragen oproepen dan het gebruik van AI bij schatten of waarnemen. Onderzoek naar toepassingen gebeurt al binnen de verschillende thema's en zal daar nu ook gewoon blijven plaatsvinden, ondersteund vanuit het thema Toepasbare AI. Centrale vraag hierbij is dus: Hoe passen we AI uiteindelijk toe bij het maken en publiceren van statistieken? Verder, wat zijn typische vuistregels bij het inzetten van AI-technieken? Bijvoorbeeld, welke familie van AI-algoritmes passen bij specifieke toepassingen en beschikbare datasets?

Tenslotte, de afgelopen periode is specifiek onderzoek gedaan naar de potentie van 'Large Language Models' (LLMs) in de standaard bedrijfsvoering, bijvoorbeeld voor het laten samenvatten, vertalen of genereren van tekst (of programmeercode) op basis van open-source LLMs die op eigen infrastructuur gebruikt worden. De eerste indrukken waren positief en daarom valt in het toepassingsgerichte kader van dit thema ook de verdere exploratie naar het verantwoord inzetten van AI in de standaard bedrijfsvoering.