De data-agenda van de overheid dient zich ook op de data zelf te richten

Data worden steeds belangrijker als basis voor beleidsbeslissingen, maar er is nog te weinig aandacht voor het verzamelen en toegankelijk maken van kwalitatief goede data. Hoe kan dat beter?

IN HET KORT

- Data zijn niet-rivaliserend en privacygevoelig en het is onmogelijk om alle mogelijke toepassingen in een contract vast te leggen.
- Hierdoor komt de data-infrastructuur op de markt onvoldoende tot stand en zijn investeringen van de overheid noodzakelijk.
- Bij het investeren moet de overheid inzetten op de kwaliteit van data, de toegang tot datareeksen en de borging van privacy.

MICHIEL BIJLSMA
Hoofd Financiële markten en finance bij SEO Economisch Onderzoek

BAS VAN DER KLAAUW
Hoogleraar aan de Vrije Universiteit

MARIKE KNOEF
Hoogleraar aan de Universiteit Leiden

De auteurs zijn lid van de Gebruikersraad Microdata Services van het Centraal Bureau voor de Statistiek

De rol die data in de maatschappij spelen, is de laatste jaren groter geworden en zal in de toekomst nog meer toenemen. Zowel overheden als bedrijven kunnen zich met behulp van data beter informeren, wat de kwaliteit van beleid en beslissingen ten goede komt. Op steeds meer plekken worden er data verzameld, en de technieken om data te analyseren worden steeds geavanceerder.

De inzichten die data genereren, creëren waarde. De overheid is met behulp van data beter in staat om effectief beleid vast te stellen en uit te voeren. Voorbeelden waarin data een rol kunnen spelen zijn: mensen op de arbeidsmarkt helpen, het ontwerpen en verbeteren van de infrastructuur door het in goede banen leiden van verkeersstromen, de sociale dienstverlening, het investeren in de verduurzaming van wijken, en het verbeteren van een integraal gezondheidsbeleid. In de huidige COVID-19-periode is het nogmaals duidelijk geworden hoe afhankelijk overheden zijn van goede data, en hoe beperkend het is als deze data niet voorhanden zijn. Denk aan locatiedata die het mogelijk maakt om de effectiviteit van contactbeperkende maatregelen te bepalen of gegevens over ziekenhuisopnames, bezettingsgraad van de intensive care en sterfteoorzaak, waar in het begin van de pandemie het zicht op ontbrak.

Beleidsmakers zijn vaak overtuigd van het nut van een gedegen empirische beleidsanalyse, en noemen regelmatig het belang van een datagestuurde besluitvorming. Ze hebben veel aandacht voor alle activiteiten rondom data, zoals big data, machine learning en AI-technieken. Maar voor de basis – namelijk de data zelf – is de aandacht geringer.

De overheid moet ambitieuzer worden wat betreft het verzamelen en toegankelijk maken van kwalitatief goede data. In dit artikel onderbouwen we de rol die de overheid heeft met betrekking tot data, door uit te leggen waarom de markt op dit punt faalt en te weinig investeert in kwalitatief goede datareeksen die over lange tijdsperioden lopen, en door meerdere partijen gebruikt kunnen worden om analyses te maken die het publieke belang kunnen dienen.

Karakter van data

Data verschillen in drie opzichten van 'normale' waren, zoals consumptiegoederen of diensten. Ten eerste zijn data niet-rivaliserend, wat betekent dat dezelfde data in meerdere toepassingen tegelijkertijd gebruikt kunnen worden zonder dat dit belemmeringen oplevert voor gebruik door een andere partij (Jones en Tonetti, 2020).

Ten tweede is de waarde die met data gecreëerd wordt moeilijk te bepalen. Welke nieuwe inzichten data-analyse kan leveren, of zelfs hoe je de data precies moet analyseren, is bijvoorbeeld vaak niet bekend, en ook toekomstige gebeurtenissen hebben invloed op wat je met de data kan of wil doen. Dat maakt het lastig om contractueel vast te leggen hoe deze waarde gedeeld zal worden met degene die de data heeft verzameld en die ze beheert (Hart, 2017). Dit probleem van niet-contracteerbaarheid wordt nog groter als uit verschillende bronnen afkomstige data gecombineerd worden, en deze bronnen ook door verschillende partijen worden beheerd.

Het probleem van niet-contracteerbaarheid verlaagt de prikkel om te investeren in dataverzameling en -beheer. De niet-contracteerbaarheid zorgt ervoor dat de positieve externe effecten van data niet-internaliseerbaar zijn door het vastleggen van eigendomsrechten (Bessen en Maskin, 2009).

Een derde belangrijke eigenschap van data is dat ze privacygevoelig kunnen zijn (Acquisti et al., 2016). Privacygevoelige data bevatten informatie over persoonlijke kenmerken of voorkeuren, die individuen en bedrijven vaak alleen willen delen met bedrijven of instanties die ze vol-

doende vertrouwen. Niet alle data zijn overigens privacygevoelig – geaggregeerde data over bijvoorbeeld economische ontwikkelingen of geografische kenmerken hebben hier geen last van.

Marktfalen

Vanwege deze drie eigenschappen van data kan er een aantal vormen van marktfalen optreden. Er kan eenvoudig een vorm van natuurlijk monopolie ontstaan. Het opzetten van een datafaciliteit heeft immers hoge opstartkosten en vaste kosten, maar lage marginale kosten. Niet alleen dalen de gemiddelde kosten naarmate er meer data worden toegevoegd, ook wordt de dataverzameling steeds waardevoller wanneer er meer data gecombineerd kunnen worden.

Verder zijn er extra kosten verbonden aan het betrouwbaar, representatief en voldoende informatief maken van data. Het monitoren van de kwaliteit van data is echter moeilijk, wat pas duidelijk wordt bij gebruik. Een gebruiker kan dus niet vooraf verifiëren hoe 'goed' de data zijn. Hierdoor kan het bekende *lemons problem* (Akerlof, 1970) optreden en kan een goede kwaliteit uit de markt gedrukt worden door een slechtere kwaliteit.

Ook het goed borgen van de privacy in het databeheer is kostbaar. Vele individuen en bedrijven zullen alleen bereid zijn om informatie te delen als ze de beheerder van de data voldoende vertrouwen. De standaardoplossing van een kwaliteitskeurmerk werkt in het geval van privacywaarborging niet goed, omdat het zowel vooraf als achteraf moeilijk te controleren is hoe partijen met data omgaan. Het is dus heel lastig voor bedrijven om op een geloofwaardige manier een reputatie op te bouwen dat er zorgvuldig wordt omgegaan met privacygevoelige informatie.

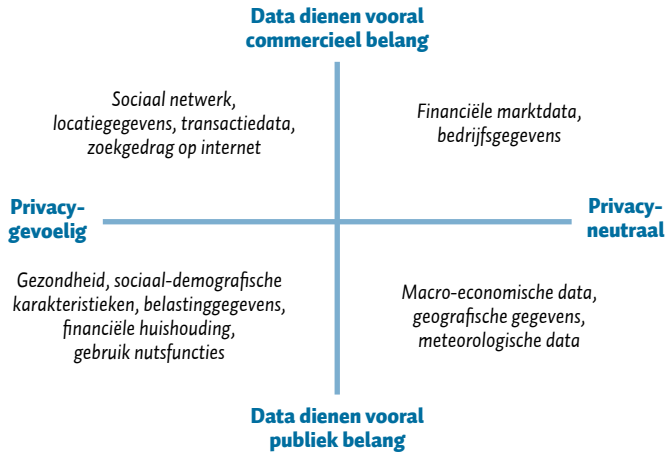
Tot slot is de opbrengst van een initiële data-analyse vaak onzeker, maar door de kennis uit de initiële analyse is het voor vervolganalyses veel beter te voorspellen wat de waarde zal zijn. Een private aanbieder van (unieke) data heeft marktmacht, en kan proberen om deze waarde 'af te romen'. Tegelijkertijd zorgt de niet-contracteerbaarheid ervoor dat afromen niet kan door derden via contracten in staat te stellen eigen data-analyses te maken. De eigenaar van data is daardoor te terughoudend met het ter beschikking stellen van die data aan derden. Het gevolg is dat een deel van de data-analyses die wel waarde hebben, niet gemaakt worden.

Markt versus overheid

Als gevolg van dit marktfalen zal de markt sommige datadiensten nooit kunnen leveren. De markt heeft een prikkel om data te genereren als ze hier een financieel belang bij heeft, en internaliseert niet de positieve externe effecten van datadelen en goed databeheer. Ook privacyzorgen kunnen marktpartijen vaak niet zelfstandig oplossen. Figuur 1 classificeert (niet uitputtend) verschillende soorten van data langs deze twee assen.

De horizontale as representeert de dimensie 'privacy' die raakt aan de karakteristiek van de data. Sommige data zijn privacygevoelig, andere niet. De Algemene verordening gegevensbescherming (AVG) biedt houvast en maakt daarbij onderscheid tussen persoonsgegevens en bijzondere persoonsgegevens. Bijzondere persoonsgegevens zijn

Verschillende typen data FIGUUR 1



ESB

herleidbaar tot een individu, en gaan bijvoorbeeld over gezondheid, migratieachtergrond, godsdienst, strafrechtelijk verleden of seksuele geaardheid. Niet alle data zijn echter privacygevoelig, denk aan data over prijzen op financiële markten, of data over inrichting van de publieke ruimte.

De verticale as representeert het doel waarvoor data worden verzameld en gebruikt. In de commerciële sector is dit voornamelijk om winst te maken. Winst geeft een prikkel tot het efficiënt- en klantgericht-zijn, maar data kunnen ook gebruikt worden om het publieke belang te dienen. Soms gaan die twee wel hand in hand, maar vaak ook niet vanwege het karakter van data: niet-rivaliteit, niet-contracteerbaarheid en privacy-gevoeligheid. Ook kan het zo zijn dat de overheid de controle wil houden over waardevolle datasets die ook gebruikt kunnen worden voor commerciële doeleinden, om ervoor te zorgen dat de baten die met de publieke data gegenereerd worden tevens terugvloeien naar de algemene publieke middelen. Denk bijvoorbeeld aan gegevens van de Belastingdienst of het Kadaster. Belangrijke terreinen van overheidsbeleid waarbij data een centrale rol spelen zijn: zorg, onderwijs, arbeidsmarkt, woningmarkt, milieu, ruimtelijke ordening, belastingen en mobiliteit.

De rol van de overheid moet het grootst zijn in het kwadrant linksonder: bij het verzamelen, beheren, verwerken en (onder specifieke voorwaarden) toegankelijk maken van privacy-gevoelige data die vooral voor publieke doelen wordt gebruikt. Eén onderzoek met goede data dat leidt tot beleidsverbetering kan de samenleving miljoenen euro's besparen. Omdat commerciële partijen niet verdienen aan deze maatschappelijke waarde, hebben ze geen prikkel om dergelijk onderzoek te realiseren. De markt heeft daar bij uitstek niet de prikkels voor om kwalitatief goede data te leveren, terwijl het combineren van dergelijke data met andere data, waar een groot deel van de waarde ligt, grote zorgvuldigheid vereist vanuit privacy-perspectief.

Ook vanuit het oogpunt van de betrouwbaarheid van onderzoeksresultaten is in dit kwadrant een grote rol van de overheid bij het beheren en verzamelen van data te verdedigen. Onderzoek op basis van private (niet-publieke) data heeft vaak het probleem dat derden de onderzoeks-

resultaten niet eenvoudig kunnen reproduceren omdat ze geen toegang hebben tot de brondata. Repliceerbaarheid is een belangrijke kwaliteitsmaat bij empirisch onderzoek.

Het kwadrant linksboven bestaat vooral uit data verzameld door grote technologiebedrijven. Combinatie van data vindt veelal plaats binnen het bedrijf. Deze data zijn potentieel relevant voor overheidsbeleid, denk aan gegevens die door big tech of telecombedrijven worden verzameld, zoals locatiedata die relevant zijn voor mobiliteitsvraagstukken. De belangrijke vraag hoe we ervoor kunnen zorgen dat de overheid deze data wanneer dat nodig is kan gebruiken voor publieke doelen, met voldoende waarborgen voor de privacy, is een ander actueel probleem dat een zelfstandige analyse behoeft.

Verzamelen en beschikbaar maken van data is kostbaar, maar wel noodzakelijk

In het kwadrant rechtsboven is er een beperkte rol voor de overheid. De markt heeft zelf voldoende prikkels om dergelijke data te generen en de kwaliteit van die data te borgen. De overheid koopt dergelijke data vaak in op de markt wanneer ze deze nodig heeft.

In het kwadrant rechtsonder fungeert de overheid vaak als een betrouwbare databron, denk bijvoorbeeld aan macrocijfers. De overheid produceert deze data en verstrekt ze vaak gratis aan de private sector als diee dergelijke data nodig heeft. Combineren van dergelijke data via derden met andere data is ook goed mogelijk.

Huidige data-infrastructuur

De Nederlandse overheid is zich bewust van het belang van data. In Nederland ligt de rol van het verzamelen, beheren en borgen van de toegang tot data die in het kwadrant linksonder vallen, voor een groot deel bij het Centraal Bureau voor de Statistiek (CBS). Het CBS verzamelt ze en beheert bestanden over een breed scala aan domeinen, zoals sociale zekerheid, gezondheid, onderwijs, arbeidsmarkt, pensioen, veiligheid en economie. Maar ook organisaties als het Kadaster, het RIVM, de Nederlandsche Bank, het KNMI, Rijkswaterstaat, de Belastingsdienst of Translink verzamelen en beheren data in het publieke belang, die vaak gedeeld worden met of als basis dienen voor statistieken van het CBS.

Een deel van de data die worden verzameld, heeft als primair doel dienstverlening of communicatie van overheden met individuele burgers of bedrijven, denk aan een vooraf ingevuld belastingformulier, of herinneringen bij het verlopen van officiële documenten.

Maar diezelfde data kunnen ook worden gebruikt voor beleidsdoeleinden en wetenschappelijk onderzoek. Met de data van het CBS en andere semi-publieke organisaties worden er dan ook veel maatschappelijk nuttige analyses gemaakt op het vlak van milieu, huizenmarkt, onderwijs, pensioenen, zorg, arbeidsmarkt en ruimtelijke ordening. Dit gebeurt door zowel publieke als private partijen, waarbij de gebruikers uit een breed veld van disciplines komen.

Deze analyses zijn omkleed met voorwaarden zodat privacybescherming is gegarandeerd. Dat gebeurt door pseudonimisering (direct identificerende variabelen vervangen door een betekenisloze sleutel), dataminimalisatie (alleen die data gebruiken die noodzakelijk zijn voor een onderzoek) of anonimiseren (aggregatie van informatie zodat die niet te herleiden is tot een individuele burger of organisatie). De AVG stelt hieraan ook strikte voorwaarden.

Het CBS neemt daarbij ook andere partijen veel werk uit handen, door bijvoorbeeld te voorkomen dat onderzoekers steeds apart data aanvragen bij UWV of de Belastingdienst, of door efficiënte dienstverlening mogelijk te maken via koppeling van verschillende databronnen.

Huidige initiatieven

Er zijn diverse beleidsinitiatieven om data voor publieke doelen beter te ontsluiten. Zo zijn er schaalvoordelen door één centrale partij verschillende databronnen te laten koppelen en aan te laten bieden. Vermeldenswaard is het initiatief ODISSEI *(Open Data Infrastructure for Social Science and Economic Innovations)*, dat tot doel heeft om onderzoekers samen te brengen met de noodzakelijke data, expertise en middelen. Ook hierbij speelt het CBS een belangrijke rol.

Een relevant initiatief is de *Data Agenda Overheid*, onderdeel van de Nederlandse Digitaliseringsstrategie 2.0 (Digitaal, 2019; 2020). Doel van de Data Agenda Overheid is om "in samenwerking het delen, combineren en de analyse van data te bevorderen". De Data Agenda Overheid heeft als uitgangspunt dat data nooit een doel op zich vormen, maar een middel, en bevat geen verplichtende kaders. Concrete producten bestaan vooral uit kaders, vormen van kennisdeling en projecten waarbij data geanalyseerd of gebruikt worden.

De data-agenda bevat vijf actiepunten (problemen oplossen via datagestuurd werken, aandacht voor wetgeving en publieke waarden, overheidsdata kwalitatief verbeteren en efficiënter benutten, kennis over datagestuurd werken verzamelen en delen, en dit investeren in mensen, organisatie en cultuurverandering), maar hier ontbreekt misschien wel de belangrijkste: het verzamelen en op een veilige manier toegankelijk maken van de (privacygevoelige) microdata zelf.

De digitaliseringsstrategie en data-agenda besteden vooral aandacht aan datadelen *tussen* bedrijven, en het toegankelijk maken van niet-privacygevoelige data *over* overheidsinstanties. Daarnaast is er geen aandacht voor datakwaliteit en consistentie over de tijd heen van de datareeksen.

Nieuw beleidsagenda

Waar moet de overheid in het kwadrant linksonder nu vooral aandacht aan besteden om de gevolgen (te kleine dataverzameling en van onvoldoende kwaliteit, te weinig hergebruik, beperkte prikkels voor de markt om privacy te waarborgen) van het marktfalen dat het karakter van data

met zich meebrengt aan te pakken? De volgende drie aspecten moeten centraal staan om de microdata van de overheid beter toegankelijk te maken en de kwaliteit te waarborgen.

Ten eerste moet in de visie van de overheid de verzameling en de kwaliteit van data meer prioriteit krijgen. Dat betekent dat data betrouwbaar, representatief en voldoende informatief dienen te zijn. Als data niet aan deze voorwaarde voldoen, dan geven statistische analyses slechts een beperkt antwoord of bezit het antwoord een grote foutmarge. Daarbij heeft de samenleving baat bij lange datareeksen die consistent zijn over de tijd, want alleen hierdoor wordt het leren van ontwikkelingen mogelijk. Daarnaast moeten data op zo'n manier bewaard worden dat de uitgevoerde analyses gerepliceerd kunnen worden. Dit betekent dat data in hun oorspronkelijk vorm bewaard moeten blijven.

Het realiseren van deze voorwaarden is kostbaar, maar wel noodzakelijk. Aandacht voor verzamelen, onderhouden en beheren van data sneeuwt gemakkelijk onder in de beleidsdiscussies over innovatie en digitalisering. Daarnaast zou er ook meer aandacht kunnen komen voor het zichtbaar maken van de maatschappelijke waarde van het gebruik van de data die door het CBS en andere (semi-) overheidsinstellingen verzameld worden.

Ten tweede moet men datareeksen breed kunnen gebruiken om voor de samenleving nuttige analyses te maken. Hoe meer en hoe vaker bepaalde data geanalyseerd worden, hoe robuuster de uitkomsten worden, en hoe meer we leren welk beleid er wel en welk niet werkt. De publieke opbrengsten van datagestuurd beleid zijn substantieel, wat een investering van de overheid rechtvaardigt. Hierbij past een *lump-sum*-financiering vooraf beter dan een financiering onder het motto 'de gebruiker betaalt'. Waar het de publieke opbrengsten betreft, moet een financiering door de gebruiker vooral als doel hebben om te prikkelen tot een efficiënt gebruik van faciliteiten, niet om de kosten te dekken van het verzamelen, bewaren en beschikbaar stellen van data.

Ten derde moet de privacy gewaarborgd blijven. Hoe kun je data op een verantwoorde manier delen en combineren? Een overheidsinstelling als het CBS kan hierbij een centrale rol spelen door een plek te vormen waar er op een veilige manier toegang geboden wordt tot kwalitatief hoogwaardige data voor analyses die het publieke belang dienen. Het CBS zou kunnen fungeren als een *trusted third party*, waar er datasets op een privacy-verantwoorde manier gekoppeld kunnen worden binnen een infrastructuur die tegemoetkomt aan de eerder genoemde privacy-eisen, zoals pseudonimisering, dataminimalisatie en anonimisering van onderzoeksresultaten. Om die centrale rol te spelen, moet aan het CBS worden toegestaan om data te 'hosten' voor anderen die ze niet zelf óók gebruiken (dit mag het instituut momenteel niet).

Conclusie

Het verzamelen, beheren, bewerken en beschikbaar maken van data voor publieke analyses is bij uitstek een publieke taak. Marktpartijen zouden voor dit doel minder data verzamelen en toegankelijk maken dan maatschappelijk gewenst zou zijn. Daarnaast is het risico groot dat één of enkele marktpartijen veel marktmacht krijgen, en dat de kwaliteit en privacywaarborgen van de data onvoldoende hoog zijn.

De overheid kan het maatschappelijk nut makkelijker internaliseren door een goede kwaliteit data toegankelijk te maken voor onderzoekers die zich bezighouden met maatschappelijke vraagstukken. Bovendien kan de overheid er toezicht op houden dat de data niet door bedrijven misbruikt worden om een concurrentievoordeel te behalen of om marktmacht te misbruiken. Ook is de overheid beter in staat dan de markt om de privacy te borgen.

Op dit moment treedt het CBS in Nederland voor veel data op als databeheerder. De microdata-omgeving omvat de privacy-gevoelige beleidsrelevante data uit het kwadrant linksonder, terwijl StatLine de beleidsrelevante niet-privacygevoelige data heeft in het kwadrant rechtsonder. Het CBS verzamelt niet alleen zelf data via enquêtes, maar bewaart ook gegevens uit verschillende administratieve bronnen. Onderzoekers kunnen onder strikte voorwaarden en tegen betaling binnen de CBS-omgeving beschikbare data analyseren. Ook overheden, semi-publieke instellingen en bedrijven kunnen onder voorwaarden gecontroleerd toegang krijgen tot data. Daarnaast wordt alle output door het CBS gecheckt voordat onderzoekers deze in hun rapportage mogen gebruiken en hebben onderzoekers de verplichting om hun rapportages openbaar te maken.

Deze constructie met het CBS als databeheerder zorgt voor een omgeving waarin er steeds meer kennis beschikbaar komt en beleidsbeslissingen steeds vaker datagestuurd zijn. Deze infrastructuur is erg waardevol, zowel voor de maatschappij als voor onderzoekers, overheden en bedrijven die sneller kennis kunnen achterhalen. Dit rechtvaardigt investeringen van de overheid en van onderzoeksfinanciers in de dataomgeving van het CBS en andere onderzoeksinstellingen en organisaties die data verzamelen en beheren in het publieke belang.

Literatuur

Acquisti, A., C. Taylor en L. Wagman (2016) The economics of privacy. *Journal of Economic Literature*, 54(2), 442–492.

Akerlof, G.A. (1970) The market for 'lemons': quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3), 488–500.

Bessen, J. en E. Maskin (2009) Sequential innovation, patents, and imitation. *RAND Journal of Economics*, 40(4), 611–635.

Digitaal (2019) *Nederlandse Digitaliseringsstrategie 2.0*, juli. Te vinden op www.rijksoverheid.nl.

Digitaal (2020) *NL Digitaal: Data Agenda Overheid*. Geactualiseerde versie 2020. Te vinden op www.digitaleoverheid.nl.

Hart, O. (2017) Incomplete contracts and control. *The American Economic Review*, 107(7), 1731–1752.

Jones, C.I. en C. Tonetti (2019) Nonrivalry and the economics of data. *The American Economic Review*, 110(9), 2819–2858.