Ministerie van Financiën

Handreiking data-analyse voor beleid

Operatie inzicht in kwaliteit

In 6 stappen naar beleidsrelevante data-inzichten

Inzicht in Kwaliteit

Inleiding

Deze handreiking gaat in op het gebruik van data-analyse in de beleidscyclus en is gericht op het ondersteunen van Rijksmedewerkers bij het opzetten of verbeteren hiervan. Deze handreiking is mede dankzij de inbreng van collega's van verschillende ministeries en de projectleiders van de initiatieven van de operatie Inzicht in Kwaliteit tot stand gekomen. Met deze initiatieven is binnen een rijk scala aan beleidsterreinen ervaring opgedaan met het gebruik van innovatieve methoden en werkwijzen om vooraf, tijdens en achteraf optimaal inzicht te krijgen in de kwaliteit van beleid. Deze handreiking is bedoeld voor beleidsmedewerkers die gebruik willen maken van data in het beleidsproces.

Wat verstaan we onder data-analyse?

Het gebruik van data(-analyse) om over beleid te informeren is niet nieuw. Wel wordt data steeds toegankelijker en makkelijker te gebruiken. In het afgelopen decennium hebben de mogelijkheden van data zich op een viertal vlakken razendsnel ontwikkeld:

1. *Productie:* de mogelijkheden om beleidsrelevante datasets te produceren zijn toegenomen door de opkomst van onder meer open data, analyse van tekstdata en de toename van het aantal big-databronnen;
2. *Management:* het opslaan en beheren van data in de Cloud is steeds gemakkelijker en veiliger geworden;
3. *Analyse:* Het aantal open source analysetalen zoals R en Python en bijbehorende algoritmen en analysemethoden is sterk toegenomen;
4. *Presentatie:* door interactieve online dashboards is het steeds beter mogelijk nieuwe resultaten tijdig en periodiek te presenteren.

In deze handreiking richten we ons op de nieuwe mogelijkheden binnen deze vier dimensies en de manieren waarop je deze optimaal kunt benutten om meer inzicht te verkrijgen in beleid.

Wat is het doel van data-analyse in de beleidscyclus?

Data zijn essentieel voor goed onderbouwde besluitvorming. Met hoogwaardige data die de juiste inzichten op het juiste moment verschaffen wordt het ontwerpen, monitoren en evalueren van beleid inzichtelijker en gemakkelijker. Het doel van data-analyse in de beleidscyclus is dus om tijdig de juiste inzichten op te doen om geïnformeerde besluiten te kunnen nemen over beleid. In stap 1 volgen enkele voorbeelden van het gebruik van data in verschillende fasen van de beleidscyclus.

Opleiding Monitoren, evalueren en leren

Vanuit de operatie Inzicht in Kwaliteit is samen met de Rijksacademie de Opleiding Monitoren, evalueren en leren (MEL) ontwikkeld. In deze opleiding leren deelnemers hoe zij monitoring en evaluatie in elke fase van de beleidscyclus kunnen inzetten om tot effectiever beleid en grotere maatschappelijke meerwaarde te komen. Er zijn twee algemene modules, namelijk over evalueren in den brede en over de Strategische Evaluatie Agenda (SEA). Daarnaast is er een viertal verdiepende modules die elk inzoomen op een specifiek type evaluatie en op monitoring. Heb je na het lezen van deze handreiking behoefte aan meer diepgang? Meld je dan aan voor één of meer modules via: https://rijksacademie.nl/opleiding/financieel-management/opleiding-monitoren-evalueren-en-leren.

Stappen

Stap 1 – Verken hoe data een rol kan spelen voor het beleid

Data kunnen in alle fasen van het beleidsproces nuttige inzichten opleveren. Ex ante kunnen data helpen om inzicht te krijgen in de stand van zaken op een bepaald beleidsthema zoals de staat van volksgezondheid en zorg doet en om het effect van verschillende beleidsopties te voorspellen zoals de analyse van de stikstofbron-maatregelen. Door indicatoren samen te stellen die periodiek worden gepresenteerd in een dashboard kan tijdig tussentijds worden gemonitord op bijvoorbeeld het aantal corona-besmettingen, stikstof depositie of de veiligheidssituatie in een bepaalde regio[1]. Voor een deugdelijke ex post analyse is het bovendien nodig om van tevoren variabelen aan te wijzen waarop die evaluatie wordt gebaseerd. Een 0-meting vooraf kan een essentiële bijdrage leveren om achteraf inzicht te krijgen in de effectiviteit van een beleidsinterventie. De gemene deler tussen deze toepassingen is dat een helder beeld van de beleidstheorie nodig is om in kaart te brengen welke data er mogelijk te verkrijgen is over de beoogde impact, outcome, output, activiteiten en input van beleid.

Bronnen:
- IAK
- Van beleidscyclus naar datacyclus
- Toolbox datagedreven werken

Stap 2 – Regel de randvoorwaarden

Om effectief aan de slag te kunnen gaan met de beschikbare data moeten een aantal randvoorwaarden worden ingevuld. Het is niet altijd nodig dat dit volledig binnen de eigen organisatie gebeurt, externe organisaties kunnen helpen door de benodigde data of expertise in te brengen. In essentie zijn de randvoorwaarden voor een succesvolle data-analyse in een beleidstraject terug te brengen tot:

1. *Infrastructuur:* de verzamelde data moet opgeslagen worden en geanalyseerd met geschikte tools. In het geval van open data is het mogelijk gebruik te maken van het rijke open source aanbod aan databases en analysetalen. Bij vertrouwelijke data is het vereist dat je ministerie beschikt over mogelijkheden tot opslag en analyse binnen een beveiligde digitale omgeving. Vraag aan de CIO binnen jouw organisatie wat er mogelijk is.
2. *Data governance:* Zeker in het geval van vertrouwelijke data is het belangrijk dat er zorgvuldig wordt omgesprongen met data zodat privacy en informatiebeveiliging geborgd zijn en de data geen verwerkingsfouten bevat. De ethische data-assistent (Deda) helpt je goed zicht te krijgen op alle ethische aspecten van verantwoord datagebruik.
3. *Vaardigheden:* Om de data vervolgens te kunnen analyseren zijn analysevaardigheden vereist. Indien deze niet binnen je afdeling aanwezig zijn kun je mogelijk beroep doen op een centraal analyseteam binnen je organisatie of bij een uitvoerings-organisatie van jouw departement. Ook is het mogelijk een data-trainee aan te nemen via het Rijks i-traineeship of een beroep te doen op het Rijks ICT-gilde.

Bronnen:
- De ethische data-assistent (Deda)
- Rijks i-traineeship
- Rijks ICT-gilde

Stap 3 – Identificeer en verzamel relevante data

Wanneer de uitdaging helder in kaart is gebracht is het tijd om de relevante data te verzamelen die meer inzichten kunnen bieden. Stap 1 uit de handreiking “in 6 stappen naar een monitor” kan hierbij helpen. Zoals in de introductie benoemd, is er een toename van het aantal databronnen en manieren om data te verzamelen. Naast de meer traditionele dataverzamelingsmethoden zoals interviews, focusgroepen en enquêtes zijn er nu steeds meer nieuwe mogelijkheden om beleidsrelevante datasets samen te stellen. Denk hierbij aan:

- *Open data:* in toenemende mate wordt data openbaar beschikbaar gesteld. Op data.overheid.nl tref je bijvoorbeeld meer dan 20.000 verwijzingen naar openbare datasets van de Nederlandse overheid. Ook de documenten op Rijksoverheid.nl worden bijvoorbeeld als open data beschikbaar gesteld.
- *Textmining:* In de beleidswereld ligt een enorme schat aan informatie verborgen in onder meer kamerstukken, rapporten, evaluaties en rapportages. Het is echter niet altijd makkelijk dit gestructureerd te ontsluiten. Textmining maakt het mogelijk om bijvoorbeeld de belangrijkste termen uit een document te halen, trends in woordgebruik over de tijd te signaleren, relaties tussen termen te vinden en grote hoeveelheden documenten te categoriseren op basis van de inhoud.
- *Big-databronnen:* de term ‘big data’ wordt op verschillende manieren gebruikt maar verwijst doorgaans naar grote en weinig gestructureerde data. Denk hierbij aan satellietdata, sensordata en grote hoeveelheden afbeeldingen. Met behulp van luchtfoto’s kun je bijvoorbeeld benaderen hoeveel zonnestroom er in Nederland geproduceerd wordt.

Bronnen:
- Toolbox - dataverzamelingsmethoden
- https://data.overheid.nl/
- https://opendata.cbs.nl/statline#/CBS/nl/
- Tekstmining voor beleidsmakers

[1] Pagina 16.

Stap 4 – Leer de data kennen en prepareer de data

De data science tegenhanger van de welbekende beleidscyclus is de "CRoss Industry Standard Process for Data Mining" kortweg CRISP DM. Dit model beschrijft in zes fasen hoe je een data-analyse project kunt plannen, organiseren en implementeren. Hoewel deze niet één op één toepasbaar zijn op een beleidscasus, zijn de stappen die beschrijven hoe je de data leert kennen en prepareert wel relevant. Het belang van deze stappen en de tijd die hiervoor nodig is, is in grote mate afhankelijk van de kwaliteit van je data-bronnen. Voor bijvoorbeeld data afkomstig van het CBS kun je deze stappen relatief snel doorlopen, maar voor data die je hebt verzameld met bijvoorbeeld scraping of die direct afkomstig is vanuit administratieve systemen is het belangrijk een aantal acties te doorlopen:

- Beschrijf de data die je hebt verzameld: over welke variabelen beschik je en wat geven de waarden van een variabele weer?
- Controleer de datakwaliteit: missen er gegevens en zijn er onverklaarbaar hoge of lage waarden?
- Schoon de data op: verwijder de variabelen en observaties die niet relevant zijn voor de beleidsvraag die je probeert te beantwoorden.
- Visualiseer de data en identificeer mogelijke verbanden: maak bijvoorbeeld enkele box-plots en histogrammen om de zien hoe de observaties verdeeld zijn en spreidingsdiagrammen om te zien of er mogelijke verbanden zichtbaar zijn tussen variabelen.
- Creëer nieuwe variabelen op basis van de beschikbare data: vaak kun je door verschillende waarden slim te combineren tot indicatoren komen die meer zeggen. Bijvoorbeeld door het budget van een gemeente te delen door het aantal inwoners, of de CO_2-emissie door het aantal auto's.

Bronnen:
- CBS publicatie over data visualisatie

Stap 5 – Analyseer met de juiste techniek

Nu je over een schone dataset beschikt met de benodigde data start je met de daadwerkelijke analyse. In deze fase kijk je eerst naar welke analysemethoden het meest geschikt zijn voor het vraagstuk. De geschikte methode is onder meer afhankelijk van het type data waarover je beschikt en de gewenste inzichten. Hieronder volgt een aantal veel voorkomende analysevraag-stukken met suggesties voor bijpassende analysemethoden.

- *Een verband aantonen:* om te toetsen of er een verband bestaat tussen twee of meer variabelen is regressieanalyse een veelgebruikte methode.
- *Een trend signaleren:* Om trends over de tijd te signaleren is tijdreeksanalyse een veelgebruikte methode.
- *Toekomstige waarden voorspellen:* om toekomstige waarden te voorspellen of bevindingen te extrapoleren zijn regressie-analyses ook populair. Er zijn echter steeds meer machine learning technieken beschikbaar die het mogelijk maken nog accuratere voorspellingen te doen. Het CPB gebruikt deze modellen bijvoorbeeld om werkloosheidsramingen te ondersteunen. Keerzijde van machine learning modellen is dat ze vaak minder transparant en moeilijker te interpreteren zijn.
- *Observaties categoriseren:* Om observaties te categoriseren is logistische regressie een voor de hand liggende methode. In het geval van tekst data is het met topic modelling mogelijk passages en documenten te categoriseren.

Sinds het verschijnen van open source analysetalen zoals R en Python is het niet langer vereist om softwarepakketten aan te schaffen om bovenstaande analyses uit te voeren. Zowel R en Python zijn gratis beschikbaar.

Bronnen:
- Zie het fiche over regressieanalyse (O13) in de toolbox beleidsevaluaties
- Machine learning casus met uitleg en voorbeelden
- R
- Python

Stap 6 – Interpreteer en rapporteer

Wanneer de analyse beleidsrelevante resultaten oplevert is het tijd om over de resultaten te rapporteren. Waar traditioneel zulke resultaten worden opgeleverd in een statisch rapport is het door interactieve online dashboards steeds beter mogelijk nieuwe resultaten tijdig en periodiek te presenteren. Een bekend voorbeeld van een dashboard dat periodiek wordt bijgewerkt is het corona dashboard. Indien het dashboard gevoelige data bevat die alleen bedoeld zijn voor geautoriseerde medewerkers kan het dashboard ook alleen binnen een beveiligde omgeving beschikbaar gesteld worden aan geselecteerde gebruikers. Mocht er binnen jouw organisatie nog geen dashboard software beschikbaar zijn dan is in het geval van openbare data een open source dashboard het overwegen waard. Zorg er goed voor dat er in- of bij het dashboard voldoende informatie beschikbaar is over de duiding van de resultaten en eventuele aannames die ten grondslag liggen aan de analyse. Indien de datasets die gebruikt zijn voor de analyse geen persoonsgegevens of andere vertrouwelijke informatie bevatten is het belangrijk deze data ook pro-actief beschikbaar te maken als open data zodat burgers, bedrijven en mede-overheden hier hun voordeel mee kunnen doen.

Bronnen
- https://coronadashboard.rijksoverheid.nl/
- Open source dashboards
- https://coronarekening.rekenkamer.nl/coronarekening/