



Dienst Uitvoering Onderwijs
Ministerie van Onderwijs, Cultuur en
Wetenschap

Synthetische Data: van Pilot naar Praktijk

Resultaten en lessen uit de experimentfase 2022-2025

maart 2026



Samen maken we onderwijs en ontwikkeling mogelijk

Dit rapport is een uitgave van:

Afdeling Informatieproducten
Dienst Uitvoering Onderwijs
Team Synthetische Data

maart 2026

Auteursrechten voorbehouden

Alle intellectuele eigendomsrechten met betrekking tot deze uitgave berusten bij Dienst Uitvoering Onderwijs of diens licentiegevers. Behoudens de uitzonderingen van de Auteurswet mag niets uit deze uitgave worden verveelvoudigd of openbaar gemaakt zonder voorafgaande schriftelijke toestemming van Dienst Uitvoering Onderwijs. Aan de totstandkoming van deze uitgave is de uiterste zorg besteed. Dienst Uitvoering Onderwijs is niet aansprakelijk voor eventuele onvolledigheden of onjuistheden in de door haar verstrekte gegevens, al dan niet afkomstig van derden.

Copyright © 2026, Dienst Uitvoering Onderwijs, Nederland, in opdracht van Ministerie van Onderwijs, Cultuur en Wetenschap

Inhoudsopgave

Introductie 4

Resultaten 5

Ingrediënten van onze synthetische data aanpak 6

Toekomst van Synthetische data 8

Introductie

DUO-Informatieproducten

De schat aan onderwijsdata die DUO-Informatieproducten (DUO-IP) beheert kan onder privacywetgeving, zoals de AVG, zelden tot nooit op microniveau gedeeld worden met derden (geaggregeerde data mag wel gedeeld worden). Omdat data op dit niveau niet gedeeld mag worden, komt DUO-IP verder af te staan van het zijn van een transparante overheid. Daarnaast is er vanwege deze privacyregelgeving beperkte mogelijkheid voor externe onderzoekers om onafhankelijk onderzoek te doen op basis van onze data. In 2022 is DUO-IP een pilot begonnen met het gebruik van synthetische data.

Synthetische Data

Synthetische data is kunstmatige data die gebaseerd is op een echte dataset en gecreëerd is met behulp van machine learning technieken. Een synthetische dataset bevat vergelijkbare informatie als de originele dataset op structureel niveau, maar komt op individueel niveau niet overeen. Bij de manier waarop DUO synthetische data genereerd is er geen 1-op-1 relatie tussen echte en synthetische datapunten, en valt de synthetische data niet onder de AVG¹.

Synthetische data kan waardevol zijn in situaties waar data te privacygevoelig is om op persoonsniveau te delen, maar er alsnog inzicht in de data gewenst is. Door het uitvoeren van statistische analyses op de synthetische dataset kan alsnog inzicht verkregen worden, zoals bij de cao-onderhandelingen in onderwijssectoren.

Synthetische data kan waardevol zijn omdat veel informatie uit de originele data behouden blijft, maar het weinig of geen privacy risico's met zich meebrengt. Tijdens de pilot zijn verschillende experimenten gedaan met synthetische data.

De resultaten hiervan zijn als volgt:

1. We kunnen bruikbare synthetische tabulaire datasets maken.
2. We leveren synthetische datasets aan partners.
3. We hebben de bestaande techniek verbeterd.
4. We hebben een ethische, juridische en technische privacy aanpak voor synthetische data ontwikkeld en in gebruik genomen.
5. We hebben een kennisnetwerk voor synthetische data opgezet.

In dit bevindingsverslag wordt een overzicht gegeven van de belangrijkste bevindingen en lessen uit de experimentperiode (2022–2025).

¹Zie de memo Privacy en Juridische Aspecten Synthetische Tabulaire Data.

Resultaten

1. We kunnen bruikbare synthetische tabulaire datasets maken.

DUO-IP is in staat om goede synthetische tabulaire data te maken voor verschillende analysedoeleinden. Deze data is bruikbaar wanneer exacte getallen niet belangrijk zijn, maar (gedetailleerde) patronen wel, bijvoorbeeld bij data die gebruikt wordt ter informatievoorziening, beleidsverkenning of trendanalyse.

Onze methode werkt in het bijzonder goed bij:

- Middelgrote datasets (ongeveer 1000 tot 10 miljoen regels);
- Variabelen met een beperkt tot middelgroot aantal categorieën;
- Data waarin patronen en samenhangen belangrijker zijn dan de exacte waarden.

Het is lastig om met de door ons ontwikkelde techniek in de volgende situaties kwalitatief goede synthetische data te genereren:

- Kleine datasets (minder dan 1000 regels);
- Kleine subgroepen binnen de dataset;
- Veel unieke records of afwijkende records (bijvoorbeeld outliers die sterk afwijken van de rest van de populatie).

Daarnaast is het niet mogelijk om met de door ons ontwikkelde techniek synthetische data te genereren voor de volgende situaties:

- Hele grote datasets (meer dan 10 miljoen regels);
- Op basis van tijdreeksen, tekst- of beelddata;
- Hele relationele databases t.b.v. softwaretesten.

Veel van deze complexere toepassingen zijn overigens kansrijk om met andere technieken te benaderen.

2. We leveren synthetische datasets aan partners.

Team synthetische data genereert structureel synthetische datasets voor interne en externe partners. Na afronding van een opdracht ontvangt de partner de

synthetische data en bijbehorende documentatie. Partners hebben daarnaast altijd de mogelijkheid om hun analyses te laten verifiëren op de echte data.

We evalueren de samenwerking en het gebruik van synthetische data samen met de afnemer. Partners geven aan de synthetische data waardevol te vinden en in te zetten voor hun onderzoeks- en beleidsvraagstukken.

3. We hebben de bestaande techniek verbeterd.

Tijdens de pilot hebben we onze techniek aanzienlijk verbeterd. Waar het eerst maar mogelijk was om variabelen tot en met 20 categorieën te synthetiseren, kunnen we dat nu met een groter aantal. Dankzij een coderingsmethode is het nu mogelijk om tot circa 200 à 300 categorieën (zoal scholen of opleidingen) te synthetiseren. Daarnaast hebben we een automatische optimalisatiemethode ontwikkeld, waardoor de synthese consistentere en met hogere kwaliteit kan worden uitgevoerd.

4. We hebben een ethische, juridische en technische privacy aanpak voor synthetische data ontwikkeld en in gebruik genomen.

Onze privacyaanpak bevat drie lagen, die in het volgende hoofdstuk worden uitgelicht:

- a. Juridische positie
- b. Ethische controle vooraf
- c. Statistische controle achteraf

5. We hebben een kennisnetwerk voor synthetische data opgezet.

Synthetische data is een techniek die niet alleen bij DUO, maar Rijksbreed potentie heeft. Daarom is het belangrijk om kennis te delen met collega's binnen de overheid en daarbuiten. Omdat de technologie nog volop in ontwikkeling is, zoeken we actief de samenwerking met partijen die zich bezig houden met synthetische data. Tijdens de pilot is een kennisnetwerk opgezet waarin tientallen geïnteresseerde organisaties kennis uitwisselen op bijeenkomsten en samenwerkingen opzetten.

Ingrediënten van onze synthetische data aanpak

Open source code: *synthpop*

De ontwikkeling van synthetische data wordt op dit moment uitgevoerd in de programmeertaal R. Voor het genereren van synthetische data is het van belang dat er volledige toegang is tot zowel R als RStudio. Beiden zijn gratis toegankelijk. Binnen de huidige aanpak wordt gebruikgemaakt van de open source R-package *Synthpop*². Voor het gebruik van deze package is geen grote reken-capaciteit vereist. Extra rekenkracht kan echter handig zijn wanneer je optimalisaties op de bestaande package wilt uitvoeren. De keuze voor open source brengt zowel voordelen als nadelen met zich mee. Het open karakter van de package maakt het mogelijk om de code inzichtelijk te houden en aan te passen aan eigen wensen, maar zorgt er ook voor dat er geen garantie is op langdurig onderhoud of ondersteuning vanuit een officiële partij.

NB: Op dit moment is de lange termijn borging van *Synthpop* onzeker. Het package wordt actief onderhouden door de originele makers, maar er is geen ondersteunende organisatie. DUO is van plan zich in te zetten om bij te dragen aan de borging.

Techniek: classificatie- en regressiebomen

Voor het synthetiseren van data gebruiken we het R-package *Synthpop*, waarbij CART (Classification and Regression Trees) als default methode wordt gebruikt. De eerste variabele wordt hierbij gesampeld en daarna worden de variabelen één voor één gesynthetiseerd gebaseerd op de relatie met de voorgaande variabele(n). De keuze voor CART is zowel gebaseerd op eigen ervaring als op de wetenschappelijke literatuur, waar het consistent wordt aangewezen als de methode die de best lijkende synthetische data genereerd³.

²Nowok, B., Raab, G. M., & Dibben, C. (2016). *synthpop*: Bespoke creation of synthetic data in R. *Journal of statistical software*, 74, 1-26.

³Fössing, E., & Drechsler, J. (2024). 'An Evaluation of Synthetic Data Generators Implemented in the Python Library Synthcity'. *Privacy in Statistical Databases*, 178-193.

Technische kennis en domeinkennis

Om data op een verantwoorde en bruikbare manier te synthetiseren, is zowel domeinkennis van de betreffende data als technische kennis van *Synthpop* essentieel. Tijdens de experimentfase heeft het team Synthetische Data ervaring opgedaan met het werken met de package, en zo een goed globaal begrip ontwikkeld van het gehele syntheseproces: van voorbereiding van de data tot evaluatie van de output.

Eigen package: *DUOpop*

Voor gebruiksgemak en herbruikbaarheid heeft het team bovendien een eigen R-package ontwikkeld bovenop *Synthpop*, genaamd *DUOpop*⁴. De package automatiseert veelvoorkomende stappen in het syntheseproces, zoals het inladen van de data, het instellen van syntheseparameters, en het exporteren van resultaten en documentatie. Dit verhoogt de reproduceerbaarheid, verlaagt de instapdrempel voor gebruik en maakt het makkelijker om een gestandaardiseerde werkwijze te hanteren.

Operationalisering

Elke synthese wordt afgelopen volgens een vast stappenplan. Dit begint met een intake met de klant. Naar aanleiding daarvan vullen we een projectplan in waar verwachtingen van partijen helder in worden vastgelegd. Rond dezelfde tijd wordt er een standaardleveringsovereenkomst getekend en gaan we aan het programmeren. Na de programmeringstappen wordt er een synthetische dataset met bijbehorende interactieve documentatie opgestuurd. Enige tijd later evalueren we samen met de klant de levering.

⁴<https://github.com/DUOpopstar/DUOpop>

Bruikbaarheidsbeoordeling

Om de bruikbaarheid van de synthetische data te beoordelen wordt voornamelijk naar een specifieke metriek (de S_pMSE score⁵) gekeken. Deze metriek bekijkt voor elke *combinatie van (twee) variabelen* of de samenhang in de echte en synthetische data vergelijkbaar is. Daarnaast worden plots gecreëerd die de synthetische en echte data plotten, en kijken we naar hoe vaak bepaalde combinaties van categorieën van variabelen voorkomen in de synthetische en originele data. In de regel maken we synthetische data waarbij de verdelingen van variabelen en combinaties van variabelen goed overeen komen. Wel is de hoeveelheid statistische onzekerheid in de synthetische data inherent wat groter⁶. Daarom bieden we ook altijd verificatie aan op de echte data.

Privacy: juridische positie

Voor de juridische positie van synthetische data is in overleg met de central privacy office een memo⁷ opgesteld. In deze memo wordt gesteld dat binnen DUO de resterende privacyrisico's van synthetische data vergelijkbaar zijn met die van geaggregeerde data, mits afkomstig uit onze aanpak.

Privacy: ethische controle vooraf

De ethische check vooraf is bedoeld om te bepalen of de oorspronkelijke dataset überhaupt geschikt is om in gesynthetiseerde vorm te delen. Daarbij wordt beoordeeld of de data te gevoelig is, bijvoorbeeld vanwege kleine aantallen, bijzondere persoonsgegevens of specifieke combinaties van kenmerken. Daarnaast wordt er gekeken naar het doel van het onderzoek waarvoor de data ge-

bruikt wordt en het bijbehorende maatschappelijke belang. NB: In principe zou deze check voor elke publicatie gedaan kunnen worden.

De uitkomst van de evaluatie deelt de dataset in een van vier categorieën in:

1. De data is te gevoelig en wordt niet gedeeld;
2. De data mag gedeeld worden, maar alleen onder een leveringsovereenkomst;
3. De data kan gepubliceerd worden, maar uitsluitend met een explain (toelichting op de methodiek en beperkingen);
4. De data kan zonder aanvullende voorwaarden als open data gepubliceerd worden.

Privacy: statistische controle achteraf

Bij het synthetiseren van data worden theoretische privacyrisico's zoals attributie, groepsinformatie en heridentificatie sterk verminderd. Directe en indirecte identifiers worden aangepast of verwijderd, waardoor personen niet of nauwelijks herleidbaar zijn. Dit vergroot de privacy. Na het genereren van de synthetische data wordt achteraf een technische controle uitgevoerd (o.a. op basis van een recent gepubliceerd wetenschappelijke consensus artikel⁸). Daarbij wordt geëvalueerd in hoeverre deze privacyrisico's succesvol zijn verminderd.

⁵Snoke, J., Raab, G. M., Nowok, B., Dibben, C., & Slavkovic, A. (2018). General and specific utility measures for synthetic data. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 181(3), 663-688.

⁶Decruyenaere et al. (2023). Synthetic Data: Can we Trust Statistical Estimators? <https://arxiv.org/abs/2312.07837>

⁷<https://nicpet.pleio.nl/news/view/of05fb6e-eaaf-46e3-bbcd-eafdaqb12160/synthetische-tabulaire-data-bij-duo-privacyvriendelijke-datasets-voor-onderzoek-en-testen>

⁸Pilgram, L., Dankar, F. K., Drechsler, J., Elliott, M., Domingo-Ferrer, J., Francis, P., Kantarcioglu, M., Kong, L., Malin, B., Muralidhar, K., Myles, P., Prasser, F., Raisaro, J. L., Yan, C., & El Emam, K. A Consensus Privacy Metrics Framework for Synthetic Data. <http://arxiv.org/abs/2503.04980>

Toekomst van Synthetische data

DUO-IP blijft data synthetiseren en delen met derden. Na de experimentfase heeft DUO-IP een vaste werkwijze ontwikkeld voor het genereren en leveren van synthetische datasets. Deze werkwijze wordt inmiddels structureel toegepast voor nieuwe of terugkerende afspraken met partners en sluit aan bij de missie van DUO-IP om informatieproducten over het onderwijs aan te bieden.

Verdere onderzoeken lopen de komende jaren door om onze werkwijze continu te verbeteren, met aandacht voor de klant. We werken aan de optimalisatie van het algoritme en de verbetering van het testing framework. Het doel hiervan is om het synthesesproces sneller in te kunnen zetten bij nieuwe use cases. Hiernaast streven we er middels de analyse van de gebruikerservaring naar om de belangen van onze klanten te overwegen en steeds meer vertrouwen te kweken in onze werkwijze. Uiteindelijk moet dit leiden tot een werkwijze die te repliceren is en bovenal toepasselijk is in andere contexten en in andere organisaties.

Tegelijkertijd draagt DUO-IP bij aan de ontwikkeling van een centrale, overheidsbrede aanpak voor synthetische data, samen met meerdere rijksorganisaties. Het onderwerp synthetische data heeft van de Programmeringsraad Generieke Digitale Infrastructuur (PGDI) budget gekregen om te kijken naar opschalingsmogelijkheden. DUO heeft samen met het UWV een coördinerende rol in het consortium van dit opschalingsticket. Daarmee wordt de samenwerking tussen overheidsorganisaties gestimuleerd om de inzet van synthetische data schaalbaar, toegankelijk en efficiënt te maken. Dit budget wordt o.a. besteed aan de ontwikkeling van open source kwalitatief goede software, waaronder doorontwikkeling van *Synthpop*, het opstellen van een generiek juridisch kader voor het gebruik van synthetische data en het bereiken van consensus over het gebruik van synthetische data.

Kortom, na een succesvolle pilot op het gebied van synthetische data positioneert DUO zich als voorloper: met een beproefde werkwijze die binnen de overheid kan worden gedeeld. Daarmee treedt DUO op dit onderwerp op als drijver van innovatie binnen de Rijksoverheid.

Colofon

Auteurs	Team Synthetische Data
Afdeling	Informatieproducten
Contact	synthetische.data@duo.nl
Datum	maart 2026