Simulatiedataset HO

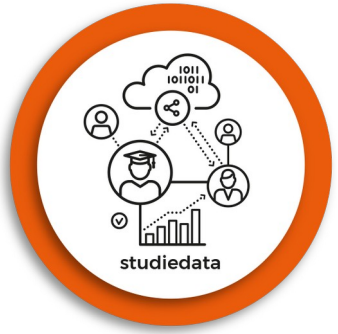
Door de Zone Studiedata

Een simulatiedataset voor uitwisseling van studiedata-analyses tussen onderwijsinstellingen en onderzoek naar studiedata door studenten

Versnellingsplan Onderwijsinnovatie met ICT

studiedata

28-10-2020, v2020.06



Inhoudsopgave

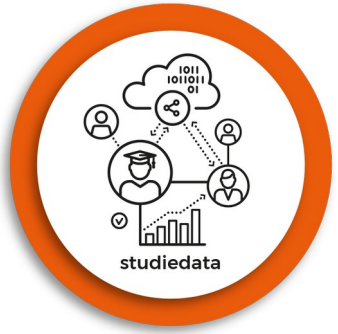
Leeswijzer

Dit document bevat een toelichting op de ontwikkeling van een simulatiedataset voor het SURF Versnellingsplan onder Creative Commons licentie (CC by 4.0) en gaat in op de technische onderbouwing en methodiek van de bouw.

Uitvoering
Dominique van Deursen, EUR, projectleider
Jurriaan Janssen, VU, ontwikkelaar
Katja van der Perk, VU, ontwikkelaar

Contact
Dominique van Deursen, EUR, d.l.vandeursen@eur.nl
Theo Bakker, VU, t.c.bakker@vu.nl

Inhoudsopgave



Introductie

Verscheidene onderwijsinstellingen zijn ervaren met het werken met studiedata. Het onderwijs wordt verbeterd door nieuwe inzichten, analyses en feiten, maar onderlinge kennisdeling is gering. **De uitwisseling van analyses en algoritmes tussen hoger onderwijsinstellingen is door de karakteristieken van de data niet mogelijk, waardoor kennisdeling en samenwerking tussen de instanties wordt bemoeilijkt.**

Sinds de inwerkingtreding van de Algemene Verordening Gegevensbescherming is het uitvoeren van onderzoek onlosmakelijk verbonden met het waarborgen van privacy. **Om de balans tussen de beschikbaarheid van studiedata voor onderzoekers en de noodzakelijke privacy-reguleringen ter bescherming van studenten te bewaren, is er behoefte aan een oplossing.**

Een dataset die beschikbaar is voor onderzoekers of in data- of statistiek-onderwijs voor studenten, geschikt is voor exploratieve data-analyse, het testen van algoritmes en het uitvoeren van statistische toetsen, maar óók de privacy van studenten waarborgt lijkt ambivalent. Daarom hebben het VU Analytics team en het BI Competence Center van de EUR een simulatiedataset ontwikkeld in opdracht van de zone Studiedata van het Versnellingsplan Onderwijsinnovatie met ICT. Dit is een dataset met gegenereerde gegevens waaruit dezelfde statistische inferenties kunnen worden ontleend als uit originele data behorend bij een universiteit, maar geheel anoniem en onherleidbaar zijn naar persoonsgegevens.

Voordelen simulatiedataset

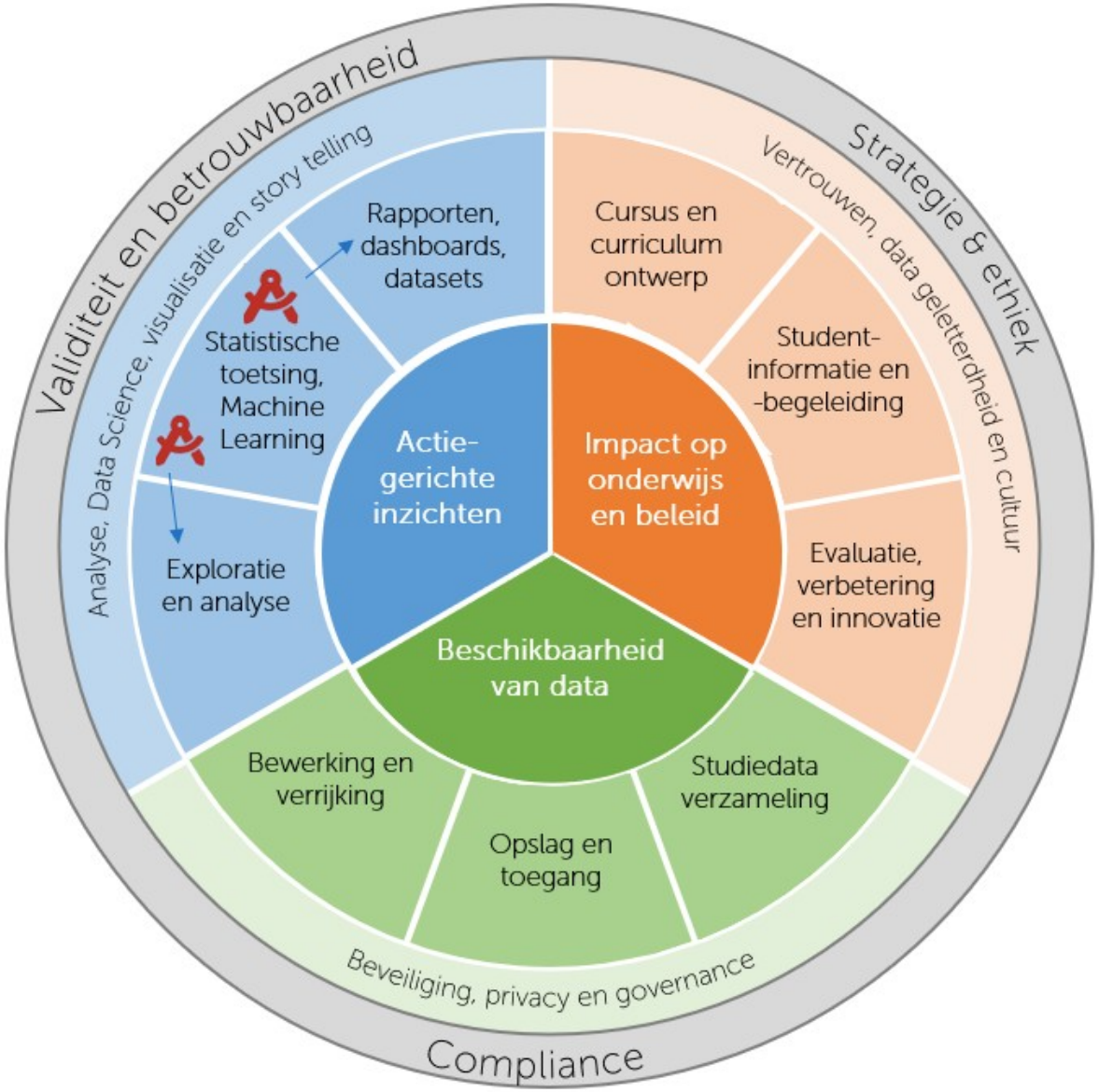
Kennisdeling tussen universiteiten

- Simulatiedata is anoniem en kan tussen onderwijsinstellingen worden gedeeld.
- Simulatiedata is toegankelijk voor studenten en onderzoekers.

Kwaliteit van onderzoek

- Uit simulatiedata kunnen dezelfde statistische resultaten worden ontleend als uit originele data.
- Modellen en algoritmes kunnen in een vroeg stadium worden getest.

Voordelen simulatiedataset

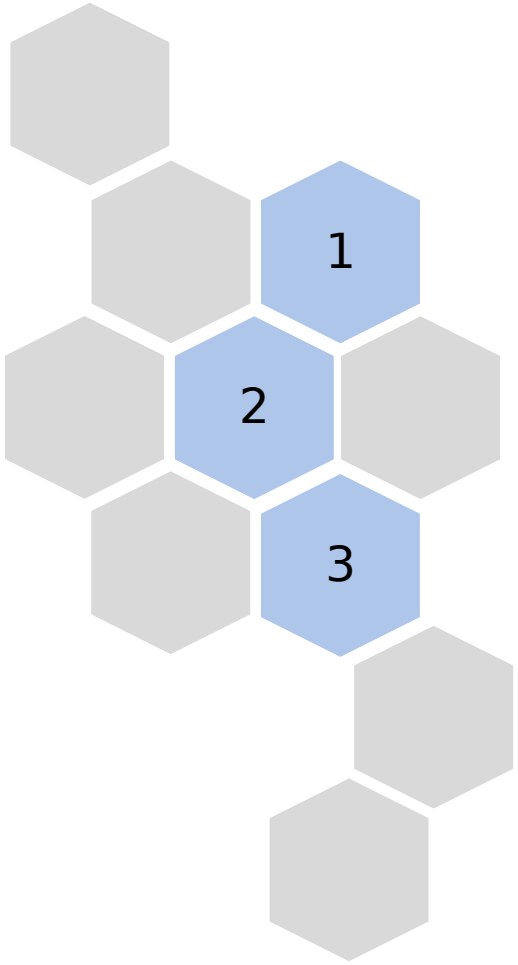


De simulatiedataset draagt bij aan validiteit en betrouwbaarheid van data-analyses op studiedata door:

- Mogelijk maken van kennisdeling met betrekking tot statistische toetsing
- Analyses kunnen tussen onderwijsinstellingen gedeeld worden
- Waar rapporten en dashboards al gedeeld kunnen worden komt er ook een anonieme dataset beschikbaar

Oftewel, inzichten kunnen met elkaar gedeeld worden waar dit voorheen niet kon.

Projectbeschrijving



De simulatiedata leidt tot dezelfde statistische inferenties als de originele data.

Om betrouwbaarheid te waarborgen bevat de simulatiedata inschrijf- en studievoortgangsdata van 25.000 studenten voor een periode van 9 jaar.

Simulatiedata bevat data van een fictieve universiteit: Universiteit van Schipluiden en is gebaseerd op data van de VU, maar geheel gesimuleerd en daarmee in geen enkel opzicht herleidbaar tot individuele studenten van deze universiteit.

Methode

1. Voorbereidingen

In dit stadium wordt de workspace van de gebruiker klaargemaakt voor de simulatie.

Bij het inrichten van de workspace horen:

- Het inladen van later benodigde functies
- Het installeren van de juiste directories
- Het installeren van benodigde packages

Methode

2. Input

In dit stadium wordt de data die gebruikt zal worden in de simulatie ingelezen.
Er is een combinatie van instellingsdata (inschrijvingen en studievoortgangresultaten) gebruikt van de VU.

1. Overeenkomstige opleidingen zijn geclusterd
2. Overige opleidingen zijn gecategoriseerd zodat de originele karakteristieken van opleidingen van de VU onherleidbaar zijn.
3. De simulatiedataset typeert daarmee een generieke universiteit

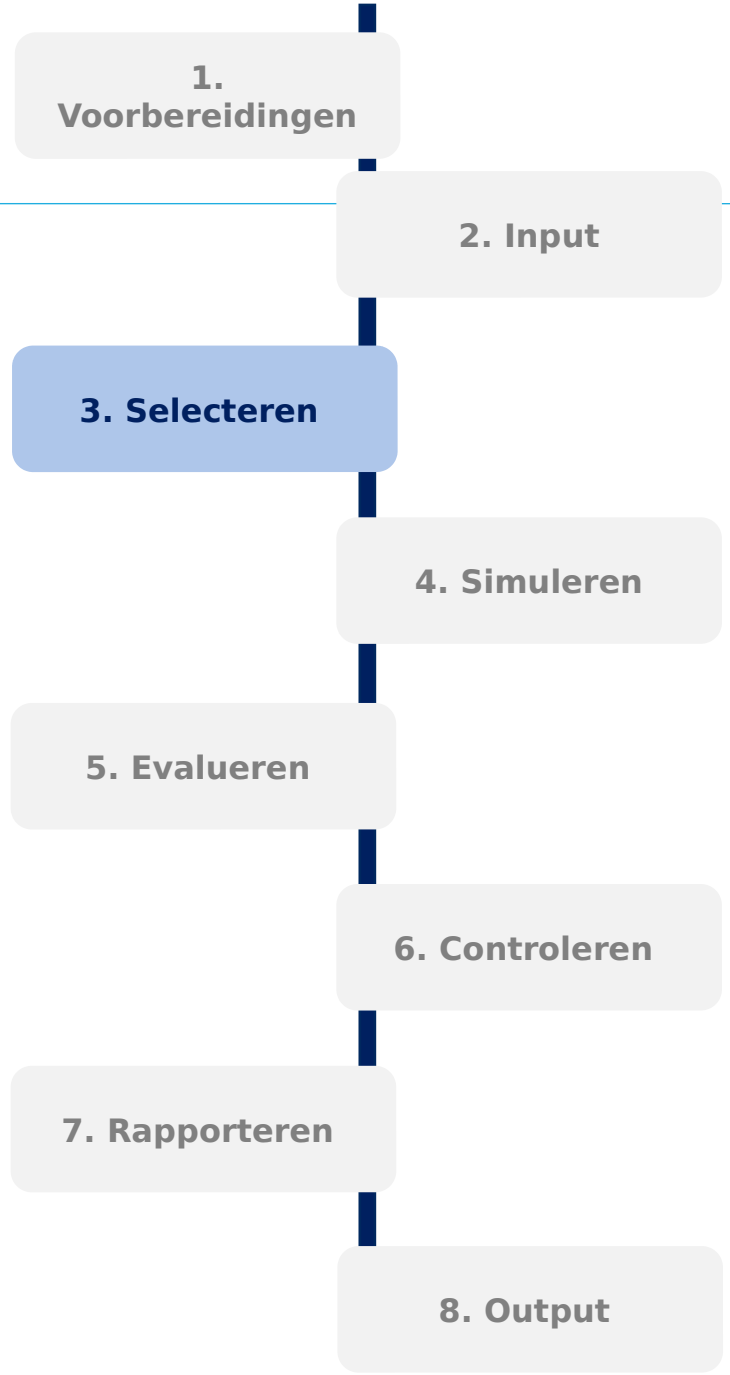
Methode

3. Selecteren

In dit stadium selecteert de gebruiker welke variabelen van de originele data zullen worden meegenomen in de creatie van de simulatiedata.

Keuze uit variabelen gerelateerd aan:

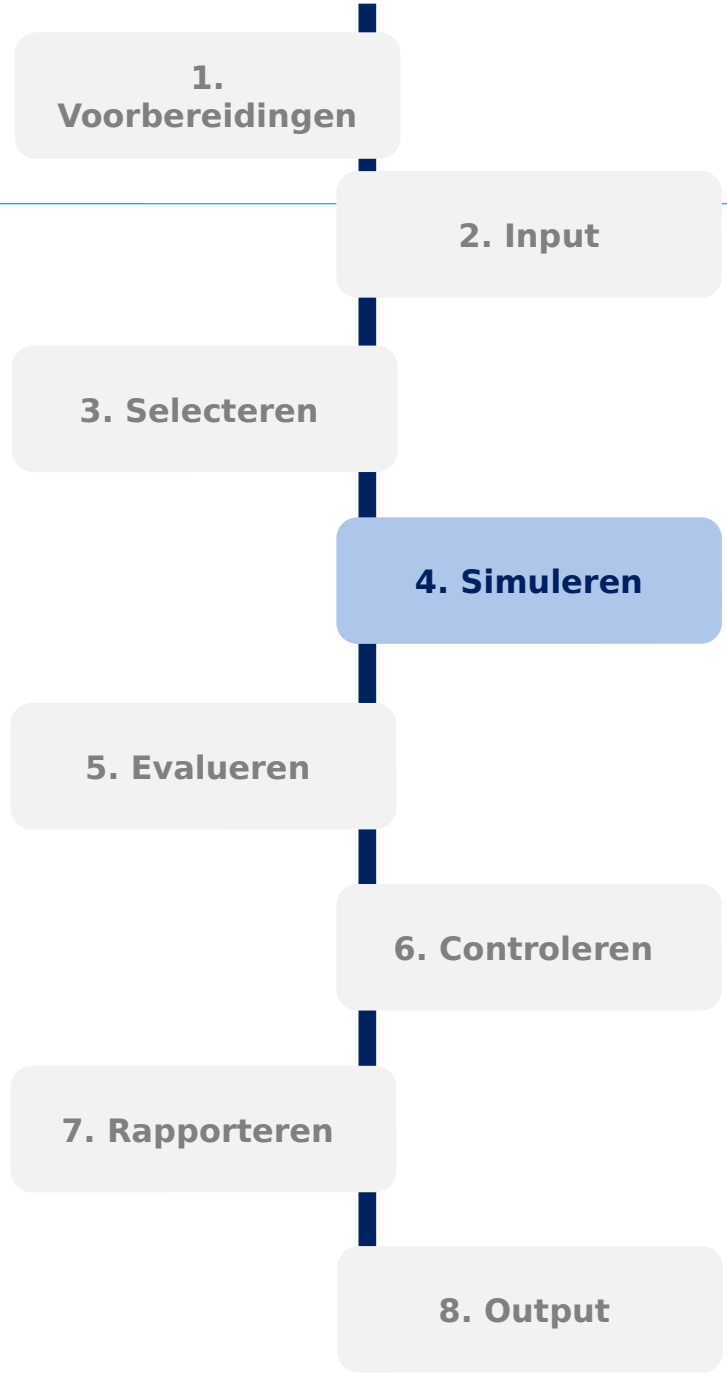
- Inschrijvingstype
- Demografie
- Aanwezigheid introductie-activiteiten
- Beoordeling van studie namens student
- Studieresultaten
- Succesvariabelen
- Resultaten uit vooropleiding



Methode

4. Simuleren

Met behulp van het synthpop package wordt synthetische c.q. gesimuleerde data gecreëerd. De onderliggende structuur en karakteristieken van de originele data blijven behouden.
Ook wordt er in dit stadium een unieke identificerende variabele gegenereerd en als een gesimuleerd studentnummer toegevoegd.



Methode

5. Evalueren (1/2)

De kwaliteit en bruikbaarheid van de gesimuleerde data wordt in twee stappen geëvalueerd, namelijk:

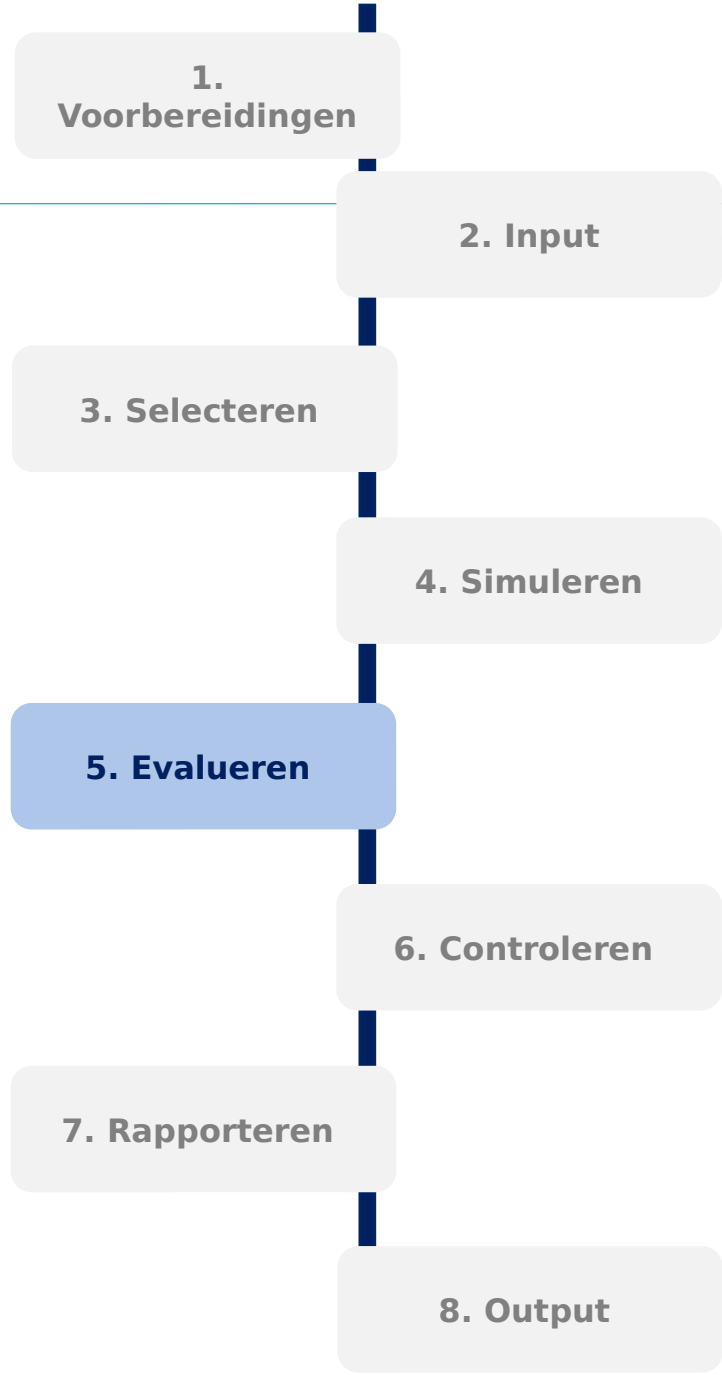
1. De onderlinge correlaties tussen variabelen in de originele data. Deze dienen behouden te blijven in de simulatiedata.
    - Numerieke variabelen worden getoetst met:
        - Gemiddelde, variantie, min, max
        - Onderlinge correlatiecoëfficiënt
    - Categoriale variabelen worden getoetst met:
        - Frequentieverdeling van observaties
        - Onderlinge frequentieverdeling (χ^2 test)

Methode

5. Evalueren (2/2)

2. De afwijking van de simulatiedata ten opzichte van de originele data. Deze dient minimaal te zijn en wordt getoetst aan de hand van:

   - Relatieve frequentieverdeling via synthpop:compare functie.
   - Test voor rijen die identiek zijn in simulatie-data en originele data.
   - Utility:gen & Utility:var scores om variabelen ten opzichte van elkaar te evalueren.
   - Utility:tab om originele en gesimuleerde data in een kruistabel te vergelijken via Voas- Williamson statistiek.

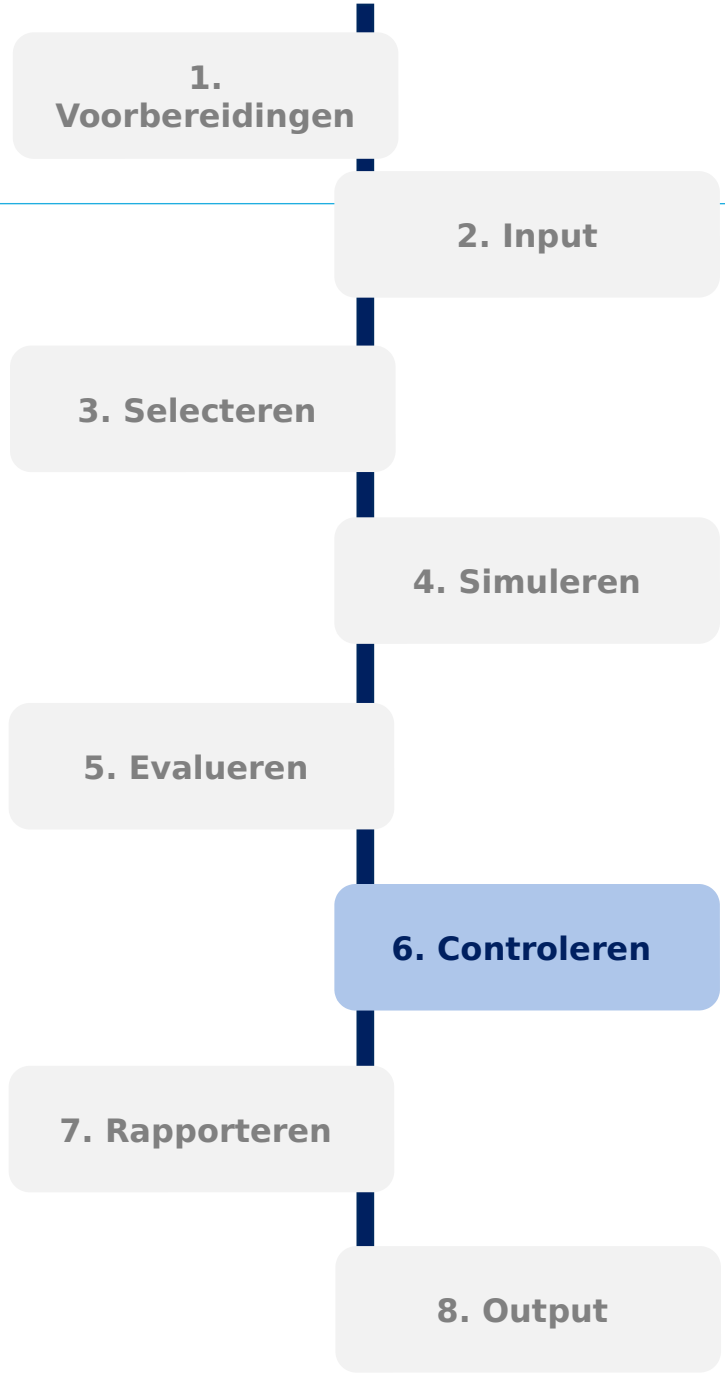


Methode

6. Controleren: logica (1/2)

Voorafgaand aan de simulatie is voor elke variabele uit de originele data vastgesteld welke logica-regels van toepassing zijn voor het genereren van gesimuleerde waarden. In dit stadium worden deze regels als functie toegepast op de simulatiedata. De volgende controles worden uitgevoerd:

- Of de variabele geslacht consistent is per studentnummer;
- Of het eerst voorkomende studiejaar correct is;
- Of studiejaren per student opeenvolgend zijn;
- Of de aanmelddatum en inschrijfjaar zich logisch tot elkaar verhouden.

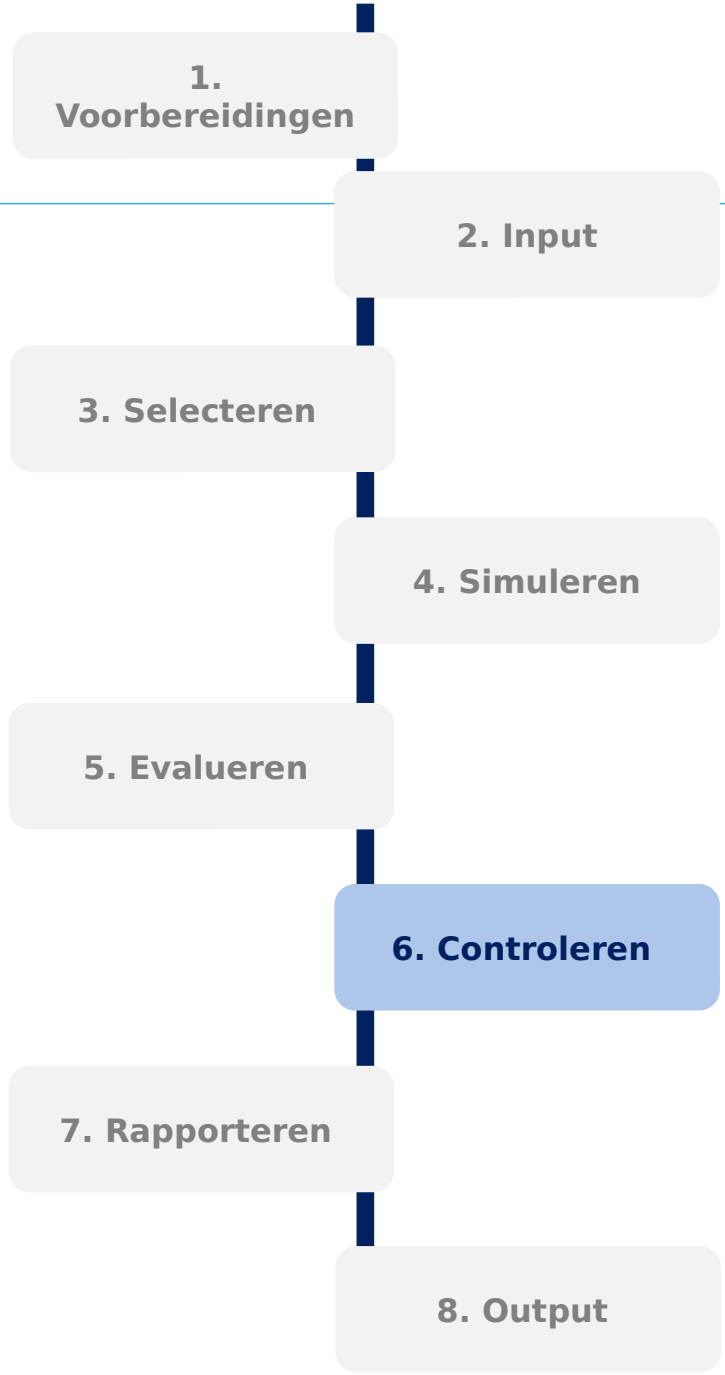


Methode

6. Controleren: privacy (2/2)

Ook wordt de privacy van studenten uit de originele data gewaarborgd door de volgende controle:

- Of er toevalligerwijs geen studenten bij toeval zijn gesimuleerd die echt bestaan, door te controleren of er identieke rijen in de simulatiedata en de brondata voorkomen.



Methode

7. Rapporteren

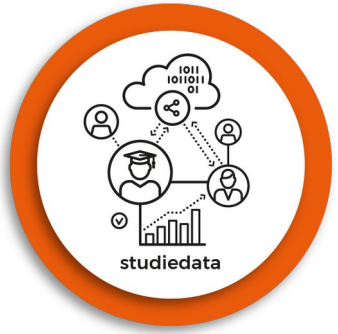
In dit stadium worden voorgaande hoofdstukken samengevat en gerapporteerd aan de gebruiker in Markdown. Dit document bevat de volgende onderdelen:

- Samenvatting en basisgegevens van gekozen variabelen uit brondata die zijn meegenomen in de simulatie;
- Beschrijving van de volgorde waarin de gekozen variabelen zijn gesimuleerd;
- Statistieken ter kwaliteitscontrole en evaluatie van de gesimuleerde data;
- Een handleiding voor interpretatie van deze evaluatiecriteria.

Methode

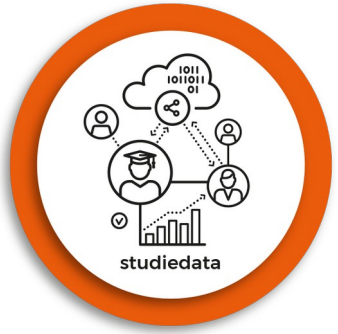
8. Output

Het script wordt na afronding gesorteerd en opgeruimd en de gebruiker kan de gesimuleerde data openen voor gebruik.



Resultaten

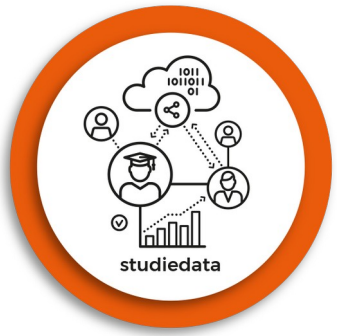
1. .CSV bestand met simulatiedata
2. Markdown file met statistieken m.b.t. kwaliteitscontrole
3. R script om simulatiedata te reproduceren
4. R package om zelf simulatiedata te genereren



Revisie

Kwaliteitsborging

1. De ontwikkeling is geëvalueerd en getoetst door data-analisten van SURF.
2. De simulatiedataset is online beschikbaar gesteld evenals de broncode in R.



Casus

Verdere ontwikkeling

Er zijn momenteel twee ambities opgenomen in de toekomstvisie van dit project:

1. Het script voor de generatie van simulatiedata voor universiteiten passend maken voor gebruik door hogescholen. Een verkennend traject hiervoor is afgerond. Komend jaar zal de bouw van de dataset voor hogescholen volgen.
2. Toevoeging van minder gestructureerde data uit leermanagement systemen om simulatie uit te breiden. Dit deelproject voert de zone uit in samenwerking met SURF.

Bijlage 1. Versiegeschiedenis

V0.4 – Verwerking feedback van Theo
V0.5 – aanpassing naar huisstijl Versnellingsplan

Voorwaarden voor gebruik van deze publicatie (1/2)

Dit is de vereenvoudigde (human-readable) versie van de volledige licentie en geen vervanging van de volledige licentie. Vrijwaring.

Je bent vrij om:
- **het werk te delen** — te kopiëren, te verspreiden en door te geven via elk medium of bestandsformaat
- **het werk te bewerken** — te remixen, te veranderen en afgeleide werken te maken
- De licentiegever kan deze toestemming niet intrekken zolang aan de licentievoorwaarden voldaan wordt.

Onder de volgende voorwaarden:
- **Naamsvermelding** — De gebruiker dient de maker van het werk te vermelden, een link naar de licentie te plaatsen en aan te geven of het werk veranderd is. Je mag dat op redelijke wijze doen, maar niet zodanig dat de indruk gewekt wordt dat de licentiegever instemt met je werk of je gebruik van het werk.
- **NietCommercieel** — Je mag het werk niet gebruiken voor commerciële doeleinden.
- **GelijkDelen** — Als je het werk hebt geremixt, veranderd, of op het werk hebt voortgebouwd, moet je het veranderde materiaal verspreiden onder dezelfde licentie als het originele werk.
- **Geen aanvullende restricties** — Je mag geen juridische voorwaarden of technologische voorzieningen toepassen die anderen er juridisch in beperken om iets te doen wat de licentie toestaat.

Voorwaarden voor gebruik van deze publicatie (2/2)

(Vervolg)

Let op:
Voor elementen van het materiaal die zich in het publieke domein bevinden, en voor vormen van gebruik die worden toegestaan via een uitzondering of beperking in de Auteurswet, hoef je je niet aan de voorwaarden van de licentie te houden.

Er worden geen garanties afgegeven. Het is mogelijk dat de licentie je niet alle gebruiksvrijheden geeft die nodig zijn voor het beoogde gebruik. Bijvoorbeeld, andere rechten zoals publiciteits-, privacy- en morele rechten kunnen het gebruik van een werk beperken.

De volledige versie van de licentie op deze publicatie is van toepassing.
Zie https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode.nl