Onthullingsrisico Corona gedragsonderzoek

5.1.2e, 4 februari 2020

Bij het bepalen van het onthullingsrisico maak ik gebruik van een methode die afkomstig is van het CBS. Voorafgaand aan de feitelijke analyse (m.b.v. de software Mu-Argus) worden de variabelen van een dataset ingedeeld in categorieën op basis van de mate waarin elke variabele zou kunnen bijdragen tot onthulling. Daarbij wordt gewerkt met de categorieën Hoog, Gemiddeld en Laag.

Variabelen waarvan je op voorhand denkt dat ze in sterke mate kunnen bijdragen tot herleiden van personen worden gerekend tot categorie Hoog. Daarnaast gebruiken we dus nog de categorieën Gemiddeld en Laag.

Bij deze eerste aanzet tot bepalen onthullingsrisico beperk ik me tot de data van ronde1. Daarbij heb ik de volgende variabelen gecategoriseerd:

Risico categorie	Variabele
Hoog	A01 (Geslacht)
Hoog	A02 (Leeftijd)
Hoog	A03 (Gemeente)
Midden	GGD_TOT (GGD regio)
Midden	A05 (Geboorteland)
Laag	A04 (Opleiding)
Laag	A06_1 t/m A06_9 (Woonsituatie)
Laag	A07_1 t/m A07_4 (Buitenplek)
Laag	A09_1 t/m A09_10 (Werksituatie)
Laag	A12 (Cruciaal beroep)

De variabelen **A04_9_text** (Opleiding – anders, namelijk: ....) en **A05_9_text** (Geboorteland – anders, namelijk: ...) vind ik sowieso te risicovol omdat hier zeer specifieke uitzonderlijke situaties kunnen worden ingevuld. Deze 2 variabelen wil ik daarom sowieso verwijderen uit eventueel beschikbaar te stellen data

Variabelen waarvan je vooraf kunt bedenken dat ze in het geheel niet kunnen bijdragen tot herleiden van personen vallen buiten deze categorie-indeling. Zij worden ook niet meegenomen bij de analyse onthullingsrisico. Dit geldt voor alle vragen uit de blokken B, C en D. Daarbij wordt gevraagd naar gedrag, welzijn, gevoelens en meningen. Op basis van de antwoorden op deze vragen is het vrijwel onmogelijk om iemand te identificeren. Deze variabelen kunnen volgens mij dus zonder probleem gepubliceerd worden.

Zie bestand CoronaGedrag_R1_variabelen.xlsx voor een volledige lijst met variabelen.