

Het gebruik van selecte registers.

5.1.2.e

28 januari 2019

Inleiding

Data wordt steeds meer ingezet als basis voor verder beleid. Algoritmen, profileringen en andere strategieën worden ontwikkeld om criminaliteit en fraude te voorspellen. Op zich zijn hier al veel methodologische kanttekeningen bij te plaatsen. Zo is het de vraag of gedrag in het verleden ook gedrag in de toekomst voorspeld: Er is vaak wel correlatie maar geen causaliteit. Voor de bronnen van de statistiek en daarmee de statistiek zelf heeft het gebruik van algoritmen echter ook gevolgen. Als bijvoorbeeld de politie meer patrouilleert in een bepaalde wijk zullen er ook meer processen-verbaal uitgeschreven worden: zoekt en gij zult vinden. Deze data komt weer terug in de statistiek en geeft de politie weer een extra bevestiging: er wordt een kunstmatige feedback-loop gecreëerd.

De elementen in de registraties zijn niet gebaseerd op een aselechte steekproef of waarneming maar op de gekozen manier van waarnemen of selectie. Het aantal elementen, het aantal processen verbaal, klopt natuurlijk wel, maar de verhoudingen tussen de wijken van crimineel gedrag is gevoelig voor de manier van waarnemen. Het is een select register. Ook zullen variabelen welke correleren met de waarneemstrategie dezelfde bias vertonen. Het principe lijkt een beetje op selecte non-respons. Maar bij selecte registers is hier lastiger voor te corrigeren.

In deze nota heb ik eerst voorbeelden proberen te vinden van selecte registers. Dit zal geen uitputtende opsomming zijn. Daarna is er gezocht naar CBS-statistieken welke gebaseerd zijn op dergelijke selecte registers. Deze zijn in de bijlage opgenomen.

Mijn conclusie is dat het CBS soms uitspraken doet welke eigenlijk het gevolg van de wijze van waarnemen en niet gebaseerd zijn op een zuiver beeld van de samenleving.

Voorbeelden van selecte registers

1. De politie gaat data gebruiken om risicovolle wijken te identificeren. Deze wijken worden dan extra geobserveerd. Dit betekent ook dat deze wijken relatief vaker in de registraties van de politie gaan voorkomen.
2. De belastingdienst gebruikt algoritmes om risicovolle respondenten te identificeren. Dergelijke respondenten krijgen relatief meer aandacht bij controle en vaststellen van aanslagen en toeslagen. Respondenten met een hoog risicoprofiel zullen dus relatief vaker in de registratie staan als fraudeur.
3. Bij het signaleren van een incident maakt de politie al dan niet onbewust keuzes: geef ik een mondelinge waarschuwing of ga ik een proces-verbaal opmaken. Het voorbeeld was de rapper in een luxe auto: een gekleurde man in een dure auto is verdacht. De politie probeert profileren op huidskleur te voorkomen maar de vraag is of dit altijd lukt. Ook hier is een potentieel gevaar voor een oververtegenwoordiging van bepaalde groepen in de registratie.
4. Onderzoek naar bijstandsfraude is waarschijnlijk ook gebaseerd op een selectieve uitvoering.
5. Big data: ook hier kan sprake zijn van een bepaalde selectie.

Zo zijn er misschien nog meer voorbeelden te verzinnen waar registraties geen aselechte afspiegeling vormen van de populatie.

In het algemeen geldt dat schatters van een populatie waarbij de doelvariabele selectief gekapt is in de registratie een bias zullen vertonen als daar niet voor gecorrigeerd wordt. In de wijk waar veel geobserveerd wordt worden meer proces-verbalen uitgeschreven. In de statistieken over deze wijken wordt niet zozeer het aantal proces-verbaal overschat, want dat is een gegeven, maar wel wordt de indruk gewekt dat deze wijken extra crimineel zijn. Er ontstaat een cirkel die zichzelf herbevestigt.

Ook bij variabelen die nauw gecorreleerd zijn met de doelvariabele en waarvoor niet via een weging gecorrigeerd kan worden, geldt dat schatters ook een bias zullen hebben die overeenkomt met de bias in het register. Een (heel) fictief voorbeeld: Als veel rijke mensen wonen in een wijk waar veel gepatrouilleerd wordt dan zullen relatief meer parkeerboetes uitgeschreven worden aan rijke mensen. Het is dan gevaarlijk daar de conclusie uit te trekken dat rijke mensen vaak fout geparkeerd staan.

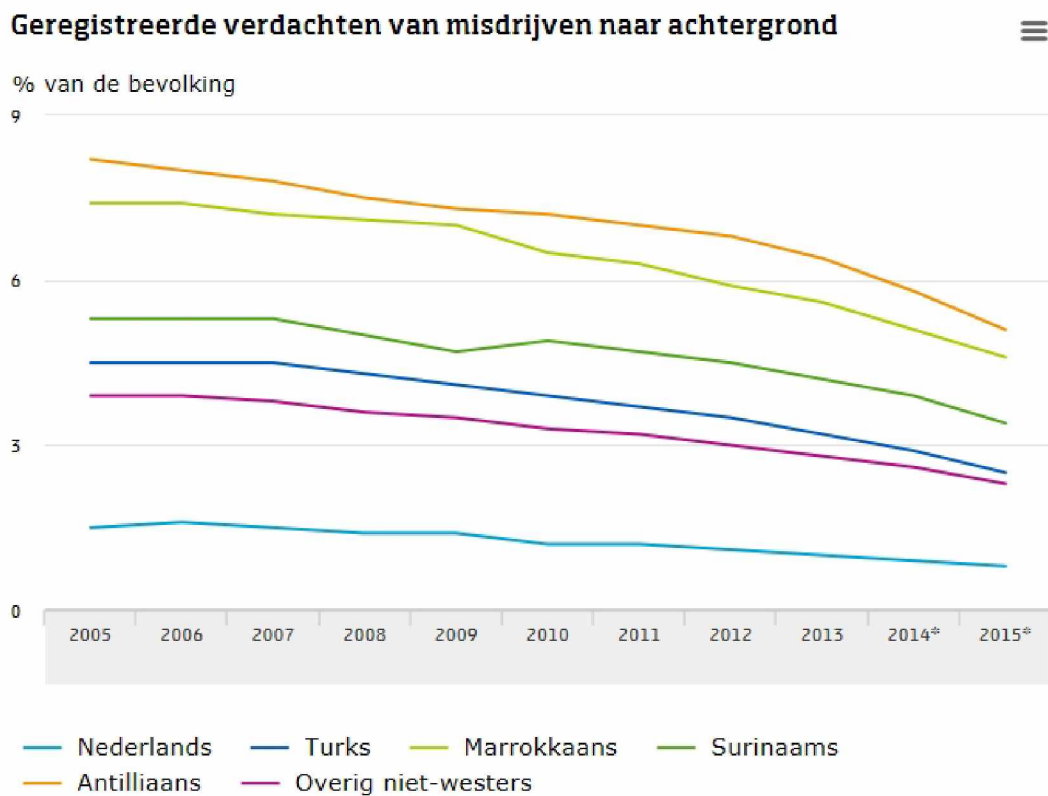
Bij de directe waarneming door het CBS op basis van een aselechte steekproef bij respondenten zien we een dergelijke bias optreden bij de non respons: het type non respondent is niet gelijkmatig verdeeld over de populatie. Hiervoor bestaan correctiemechanismen op basis van bekende eigenschappen van én de populatie én de steekproefelementen.

Bij aselechte registraties is het veel lastiger om dergelijke correctiemechanismen te bepalen. De politie zou bijvoorbeeld kunnen aangeven hoeveel extra aandacht in termen van tijd(?) een wijk gegeven is. De vraag is dus of we dergelijke registraties waarbij de regio als indeling gebruikt is voor extra observatie, moeten gebruiken om regionale statistieken te maken zolang het CBS niet kan corrigeren voor de bias in de registratie.

Bij het derde voorbeeld is het vinden van een correctiemechanisme nog veel lastiger.

Bijlage met voorbeelden

Sinds 2005 is het percentage door de politie geregistreerde verdachten van misdrijven onder alle herkomstgroeperingen bijna gehalveerd. Dat geldt zowel voor personen met een Nederlandse achtergrond als voor personen met Turkse, Marokkaanse, Antilliaanse of Surinaamse achtergrond.



<https://www.cbs.nl/nl-nl/achtergrond/2016/47/criminaliteit>

Het relatieve verschil tussen de gekleurde lijnen is gevoelig voor de manier van waarnemen.

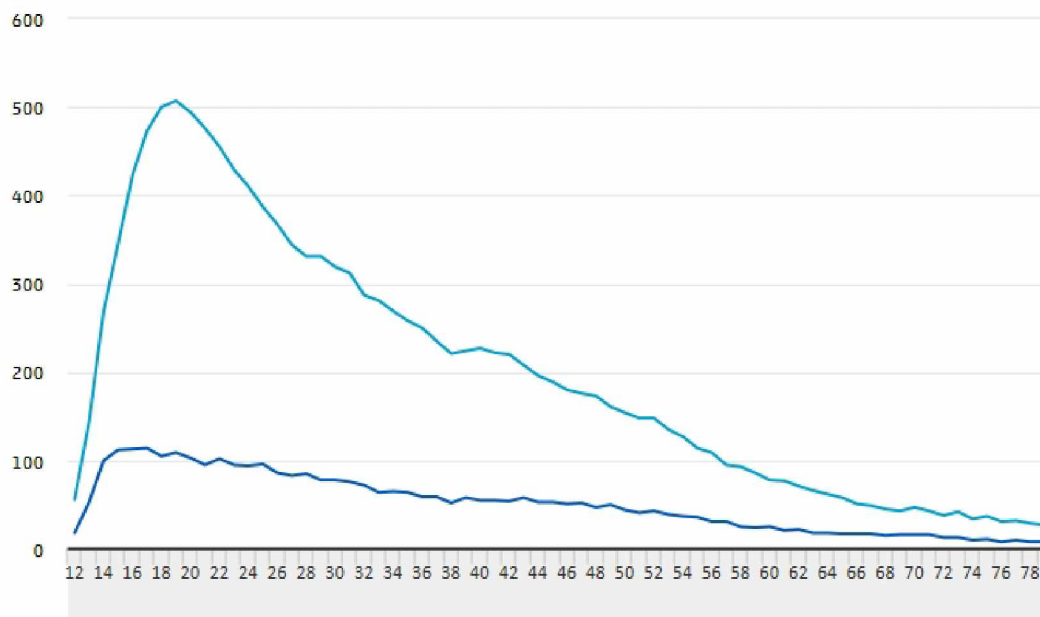
<https://www.cbs.nl/nl-nl/nieuws/2015/51/vrouwen-vier-keer-minder-verdacht-dan-mannen>

Ook hier is het relatieve verschil tussen mannen en vrouwen gevoelig.

Geregistreeerde verdachten van misdrijven naar geslacht en leeftijd, 2014



per 10 000



— Mannen — Vrouwen

<https://www.cbs.nl/-/media/ excel/2017/50/tabel%20piot%20jeugdcriminaliteit.xlsx>

22

23

24

25

26

27

28

29

30

31

32

33

34

35

36

37

38

39

40

41

42

43

44

45

46

47

48

49

50

51

52


53

54

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V		
2) Gemeente-indeling per 1-1-2017. Onderzoeksresultaten voor de gemeenten Bergeijk, Bladel, Eersel en Reusel - de Mierden. De gemeente Oirschot doet vanuit de Kemengemeenten niet mee aan dit onderzoek.																							
3) Het totaal van de gemeenten Bergeijk, Bladel, Eersel en Reusel - de Mierden.																							
4) De Kemengemeenten willen graag de onderzoeksresultaten afzetten tegen en vergelijken met de resultaten van de provincie Noord-Brabant.																							
Tabel 3																							
Geregistreerde verdachten van misdrijven en jeugdige geregistreerde verdachten van misdrijven per 10 000 inwoners ^{a)} , uitgesplitst naar geslacht, herkomst, naar leeftijd, naar gemeente ^{b)} , naar totaal vier Kemengemeenten ^{c)} , naar provincie Noord-Brabant ^{d)} , voor de verslagjaren 2010 t/m 2016																							
		2010						2011						2012									
		Totaal geregisteerde verdachten			waaronder Totaal jeugdige geregisteerde verdachten			Totaal geregisteerde verdachten			waaronder Totaal jeugdige geregisteerde verdachten			Totaal geregisteerde verdachten			waaronder Totaal jeugdige geregisteerde verdachten						
					Geslacht		Leeftijd				Geslacht		Leeftijd										
					man vrouw		12 t/m 14 jaar 15 t/m 17 jaar				man vrouw		12 t/m 14 jaar 15 t/m 17 jaar										
														</									

<https://imopendata.cbs.nl/#/JM/nl/dataset/20209NED/table?ts=1548245712110>

De Jeugdmonitor: de verschillen tussen het relatief aantal verdachten naar achtergrondkenmerken is gevoelig.

 StatLine		Jeugdmonitor		
Verdachten van 12 tot 25 jaar; delictgroep, persoonskenmerken				
Gewijzigd op: 30 maart 2018				
Variabelen kunnen gesleept worden naar de kop, rijen of kolommen van de tabel. In de kop is maar één item van een variabele te selecteren.				
Leeftijd 12 tot 25 jaar				
		Herkomstgroepering Geslacht Perioden		
Onderwerp		Totaal	Nederlandse achtergrond	Met migratieachtergrond
		Totaal mannen en vrouwen	Totaal mannen en vrouwen	Totaal mannen en vrouwen
		2017 ^a	2017 ^a	2017 ^a
Geregistreerde verdachten				
Totaal verdachten				
Totaal verdachten van misdrijven	aantal	55 290	27 450	27 790
Verdachten van geweldsmisdrijven	aantal	12 800	6 520	6 270
Verdachten van vermogensmisdrijven	aantal	23 300	9 430	13 850
Verdachten vernieling en openbare orde	aantal	9 770	5 410	4 350
Verdachten van verkeersmisdrijven	aantal	8 130	4 960	3 170
Verdachten van drugsmisdrijven	aantal	4 870	2 310	2 550
Verdachten van vuurwapenmisdrijven	aantal	2 040	1 010	1 050
Verdachten (relatief)				
Totaal verdachten van misdrijven	aantal per 10 000 inwoners	181	135	307
Verdachten van geweldsmisdrijven	aantal per 10 000 inwoners	44	32	77
Verdachten van vermogensmisdrijven	aantal per 10 000 inwoners	72	46	146
Verdachten vernieling en openbare orde	aantal per 10 000 inwoners	33	27	51
Verdachten van verkeersmisdrijven	aantal per 10 000 inwoners	28	24	38
Verdachten van drugsmisdrijven	aantal per 10 000 inwoners	15	11	26
Verdachten van vuurwapenmisdrijven	aantal per 10 000 inwoners	7	5	12
Aangehouden verdachten				