



Kaart uit 1753 van Isla Más a Tierra – beter bekend als het eiland Robinson Crusoe – is het grootste eiland binnen de Juan Fernández-archipel en ligt in de Zuidelijke Grote Oceaan op 674 kilometer van Chili

# EXPEDITIE ROBINSON

## VAN ECOLOGISCHE CORRELATIE NAAR MULTI LEVEL ANALYSE

RICHARD STARMANS

### Structuren in een gedataficeerde wereld

In tijden van data science en big data is het bon ton krachtige hyperbolen niet te schuwen. Zo heeft zich volgens ingewijden een data-revolutie voltrokken, zijn we getuige van een data-explosie of – beter nog – een data-tsunami en leven we onmiskenbaar in een gedataficeerde wereld. De alomtegenwoordigheid en diversiteit van data en de snelheid waarmee zij beschikbaar komen, geven voeding aan de gedachte dat de werkelijkheid in al haar complexiteit nagenoeg geheel wordt gevat, gecodeerd en gecodificeerd in data. Deze voorstelling gaat deels terug tot 19e eeuwse opvattingen van onder meer Ernst Mach (1838-1916) en vooral Karl Pearson (1857-1937). Laatstgenoemde deed recht aan variatie en verandering in de natuur door deze niet te identificeren in errors, maar in de ver-

schijnselen zelf (gecodeerd in data), en terug te voeren tot verschillende (klassen van) kansverdelingen. Hij zag in dat vele verschijnselen niet normaal, maar scheef waren verdeeld en met behulp van vier parameters (gemiddelde, variantie, scheefheid en welving) konden worden beschreven. Kansverdelingen met bijbehorende parameters werden de objecten van de wetenschap. Als eerste gaf Pearson daarmee kansverdelingen een volwaardige plaats in de wetenschap en zag de wereld op een niveau van abstractie waarbij data, variatie in data, data genererende mechanismen en parameters van de kansverdelingen de werkelijkheid veeleer coderen en opbouwen en niet zozeer een (vermeende) fysische werkelijkheid representeren of afbeelden. Die laatste voorstelling krijgt enige plausibiliteit tegen de achtergrond van het feit dat de zeer

empiristische en anti-causalistische Pearson opmerkelijk genoeg geen materialistische, maar een idealistische denker was. De werkelijkheid geldt primair als een constructie van de menselijke geest en het doel van wetenschap is een inventaris te maken van de inhoud daarvan (Pearson, 1892). Het oude Kantiaanse idee dat het kenvermogen zelf een structuur oplegt aan de werkelijkheid maakt het idee van een volledige datafificatie minder buitenissig.

Een en ander werpt een nieuw licht op het traditionele filosofische *dualisme* van taal en werkelijkheid en de daarbij gekoesterde correspondentietheorie van de waarheid, maar ook op de methodologische queeste naar het ultieme valide meetinstrument en het wiskundige streven naar een adequate *representatiestelling*. Aan de ene kant zijn er de complexiteit en dynamiek van de werkelijkheid in al haar verschijningsvormen, processen, variatie en verandering, onderliggende structuren, inclusief de *wisselwerking van entiteiten met de omgeving* waarin zij zich bevinden c.q. de groepen waartoe zij behoren. En uiteraard de voorstellingen die wij van die werkelijkheid maken, de gepostuleerde theoretische/abstracte entiteiten en de verhouding tot de al dan niet vermeende waarneembare entiteiten, en de taal waarin we daarover spreken. Aan de andere kant zijn er de data, die aan dit alles recht moeten doen en waaraan dus hoge eisen moeten worden gesteld: betrouwbaarheid, validiteit, kennis van het data genererende mechanisme en vooral de structuur in de data. Met het oog op dit laatste zij hier opgemerkt dat statistici in de loop der tijden uitermate bedreven en volhardend zijn gebleken in het zoeken naar en creëren en analyseren van nieuwe *bronnen van varia(n)tie* door het ontwikkelen van complexere methoden en technieken. We beperken ons hier tot één aspect van dit alles, de voornoemde *wisselwerking van individuen met de omgeving*. Statistici herkennen hierin data die doorgaans als *hiërarchisch* worden opgevat; zij zijn genest, ingebed, of gelaagd, dat wil zeggen zij kennen verschillende niveaus. Een traditionele, maar geenszins exclusieve *interpretatie* (of gepostuleerde correspondentie met de werkelijkheid) hiervan, betreft de opvatting dat deze data soms eigenschappen beschrijven van *individuen*, min of meer 'ondeelbare', singuliere entiteiten, die deels invariant zijn, op zichzelf staan en een eigen persoonlijke identiteit krijgen toebedeeld. Soms beschrijven de data eigenschappen van groepen of clusters, waartoe deze entiteiten behoren, waarin zij zijn ondergebracht of georganiseerd, of waaruit zij voortkomen. Deze groepen 'bepalen' hun afkomst, soort, status,

omgeving, hetzij langs natuurlijke, hetzij langs artificiële of institutionele weg. Zo kan de inbedding plaatsvinden op basis van fysische, biologische, sociale of demografische criteria. De groepeigenschappen worden dikwijls beschreven met *geaggregeerde* data, dat wil zeggen data die ontleend zijn aan – en dus op hun beurt afhangen van – metingen verricht aan voornoemde singuliere entiteiten; gemiddelden, proporties, indexcijfers of correlaties. Omgekeerd kunnen eigenschappen van voornoemde entiteiten weer afhankelijk zijn van karakteristieken van de bewuste groepen. Zo beschouwd spelen de wisselwerking, interdependenties en interacties tussen de verschillende *levels* in vrijwel alle wetenschapsgebieden een rol, variërend van biologie, ecologie en epidemiologie tot sociologie, economie en robotica. De statisticus die recht wil doen aan de voornoemde wisselwerkingen en verschillende 'niveaus' simultaan in een analyse wil betrekken, beschikt vandaag de dag over een groot aantal technieken die hier gemakshalve als *multilevel* analyse zullen worden getypeerd. Of het nu gaat om multilevel models in strikte zin, hierarchical (linear) models, nested data, mixed models, klassieke split-plot designs, random coefficient of random-effect models, repeated measures, et cetera. Strikte definitie- en demarcatiekwesties blijven hier buiten beschouwing. Het gaat erom dat het zoeken naar en creëren van nieuwe *bronnen van varia(n)tie* leidt tot modellen met meer random effects, minder fixed effects en een meer gesofisticeerde analyse van residuele error, dit alles binnen en tussen de verschillende niveaus.

### De anatomie van de ecologische correlatie

In handboeken en overzichtswerken op het gebied van multilevel analyse wordt doorgaans benadrukt dat de problematiek een respectabele genealogie kent. Dikwijls betonen de auteurs zich daarbij op de een of andere wijze schatplichtig aan het werk van de Amerikaanse statisticus William S. Robinson (1913-1996). In 1950 verscheen van zijn hand het amper vijf bladzijden tellende artikel 'Ecological Correlations and the Behaviour of Individuals' in de *American Sociological Review*. Het behoort vandaag de dag tot de meest geciteerde papers in de methodologie van de sociale wetenschappen en is uiterst relevant in het licht van de hier geschetste problematiek. Een aantal merkwaardige aspecten van het artikel bleef lang onopgemerkt ondanks het grote aantal citaties en de nagenoeg

universele bijval die de publicatie ten deel viel. Deze betreffen zowel kleine fouten als inhoudelijke inconsequenties van het artikel. Enkele facetten ervan passeren hier kort de revue.

Robinson licht de thematiek toe aan de hand van een tweetal eenvoudige maar saillante voorbeelden, ontleend aan data van de volkstelling in de VS uit 1930. Allereerst onderzoekt hij het verband tussen etniciteit (blank versus zwart) en analfabetisme. Robinson berekent om te beginnen een zogenaamde ecologische correlatie: *'The statistical object is a group of persons. The thing described is the population of a state, and not a single individual. The variables are percentages, descriptive properties of groups, and not descriptive properties of individuals.'* (Robinson, 1950) Hij visualiseert deze met behulp van een scatterplot. De onderzoekseenheden worden gevormd door negen kiesdistricten, die elk getypeerd worden met het percentage zwarten en het percentage analfabeten aldaar woonachtig. De Pearson-correlatiecoëfficiënt tussen beide groepskenmerken blijkt zeer hoog: 0,97 (geval I). Vervangt men de kiesdistricten door de 48 verschillende staten, dan wordt de ecologische relatie 0,77. Beide ecologische correlaties suggereren ruwweg: in gebieden met veel zwarten is veel analfabetisme. Vervolgens berekent Robinson een individuele correlatie, een *'correlation in which the statistical object or thing described is indivisible'* op basis van ruim negentig miljoen onderzoekseenheden, weergegeven in een in een  $2 \times 2$  kruistabel. De individuele *fourfold-point* correlatie-coëfficiënt is weliswaar eveneens positief, maar bedraagt slechts 0,2 (geval II). Alsof dit contrast tussen I en II nog niet genoeg is, komt Robinson met een tweede voorbeeld, waarbij ook het teken en daarmee dus de richting van de correlatie veranderen. Dit voorbeeld betreft het verband tussen geboorteland (*native born* versus *foreign born*) en analfabetisme. De ecologische relaties op basis van districten (-0,62) en staten (-0,53) zijn negatief (I), de individuele correlatie blijkt positief: 0,12 (II). Daarmee sluit zijn voorbeeld nog beter aan bij de vele klassieke paradoxen en anomalieën (confounding, 'spurious correlation', Simpson-paradox) die traditiegetrouw in inleidende cursussen statistiek worden besproken.

Ogenschijnlijk is Robinsons doelstelling bescheiden. Hij beoogt niet een nieuwe analysetechniek of eigen paradigma te ontwikkelen, maar geeft in eerste instantie slechts een caveat, een profylactische waarschuwing. Naar eigen zeggen verschaft hij *'a definite answer as to whether ecological correlations can validly be used as sub-*

*stitutes for individual correlations. They cannot.'* De auteur stelt dat *'the purpose of this paper will have been accomplished, however, if it prevents the future computation of meaningless correlations and stimulates the study of similar problems with the use of meaningful correlations between the properties of individuals.'* Hoe dan ook, vandaag de dag wordt deze problematiek dikwijls getypeerd als de *ecologische drogreden*, complementair aan de *atomistische drogreden*, waarbij juist op basis van individuele data conclusies worden getrokken over (eigenschappen van) groepen.

Robinson besluit de kwestie te verhelderen door recht-toe rechtaan de 'anatomie van de ecologische correlatie' te analyseren, waarbij hij een minimum aan wiskundige voorkennis vooronderstelt. Zo laat hij onder meer zien dat de verbindende schakel tussen analyse I en II wordt gevormd door de *within-areas individual correlations*, oftewel de individuele *fourfold-point* correlaties tussen etniciteit en analfabetisme, berekend voor elk van de negen districten (III). Analyses I en II hangen beide hiervan af, maar op een verschillende manier; voor I geldt dat deze wordt bepaald *'only upon their marginal totals'*, II daarentegen *'upon the internal frequencies'*. Uiteraard bepalen in een  $2 \times 2$  kruistabel de rij- en kolomtotalen (de 'afzonderlijke' verdelingen) niet de inhoud van de vier cellen (de 'gemeenschappelijke' verdeling). Andere aspecten, waaronder de wegingen van de data die Robinson hanteert en de formele wiskundige relaties tussen I, II en III blijven hier buiten beschouwing. Zie hiervoor (Robinson, 1950) en bij voorbeeld (Kreft, 1987) en (Te Grotenhuis, 2011). Hier beperken we ons tot een tweetal opmerkelijke gezichtspunten van Robinson, die van cruciaal belang zijn, zowel binnen de context van de sociologie, als in het licht van de ontwikkelingen in de statistiek.

### De erfenis van Durkheim

Het eerste opmerkelijke gezichtspunt van Robinson wordt duidelijk wanneer hij onomwonden stelt dat gebruik van ecologische correlaties niet geschiedt omdat de auteurs geïnteresseerd zouden zijn in (correlaties tussen de) eigenschappen van gebieden, groepen et cetera. Bij alle ecologische studies is *'the obvious purpose'* iets te leren over het gedrag van individuen. *'Ecological correlations are used simply because correlations between the properties of individuals are not available. In each instance however the substitution is made tacitly rather than explicitly'*. Cu-

rius is dat de uitspraak wordt gedaan in een tijdschrift voor sociologie, omdat de wisselwerking tussen individu en groep juist daar op bijzondere, eigen manier gestalte krijgt. Centraal staan hier zowel de vraag naar de ontologische status van groepen (instituties, collectieve representaties) als gepostuleerde 'echte' bouwstenen van de samenleving, als de vraag naar de rol van de statistiek bij dit alles. Een en ander werd reeds actueel sedert de opkomst van de sociologie in de 19e eeuw toen pioniers als Karl Marx, Adolphe Quetelet, Auguste Comte en later vooral Max Weber (1864-1920) en Emile Durkheim (1858-1917) trachtten het vak een eigen wetenschappelijke status te geven door het af te splitsen van de wijsbegeerte en de daarmee verbonden speculatieve traditie. Daarmee werd de jonge discipline direct in een krachtenveld van enkele klassieke positiebepalingen of tegenstellingen getrokken: empirisme versus rationalisme, *Erklären* versus *Verstehen* (Dilthey), het nomothetische versus het idiografische (Windelband) en vooral een deterministisch versus een probabilistisch wereldbeeld.

Tegen deze achtergrond moesten object en methode worden bepaald. Zo lag het in de rede aansluiting te zoeken bij het dominante empirisme, dat onder meer culmineerde in Comte's positivism. Dat empirisme vroeg om een sterke empirische basis, hetgeen in 'de met cijfers bedekte 19e eeuw' (Stamhuis, 1992), waarin een *'avalanche of numbers'* (Hacking, 1989) waarneembaar was, niet problematisch leek. Bovendien was er naast deze -door de net opgerichte statistische bureaus in Europa vergaarde- data nog een tweede bron van gegevens, de etnografische gegevens van uitheemse culturen, verzameld in overzeese gebieden. Volgens (Oberschall, 1989) vormen beide databronnen de *'empirical roots of social theory'*. Hoe dan ook, om de wetenschappelijke status te verhogen lag het volgens velen meer voor de hand aansluiting te zoeken bij de op *Erklären* gerichte nomothetische natuurwetenschappen, dan eigen spelregels te formuleren of zich te richten op *Verstehen* en een idiografische benadering (waarvoor de etnografische data natuurlijk meer geschikt leken). En als men dan voor de natuurwetenschappelijke basis koos (Comte, Quetelet), moest dan een klassieke deterministische benadering (die in de 19e eeuw op zijn retour was) het uitgangspunt zijn of moest juist gebruik worden gemaakt van de opkomende 'nieuwe' statistiek in een eeuw waarin variatie en verandering in geschiedenis en natuur als inherent werden beschouwd en een probabilistisch wereldbeeld zich begon af te tekenen?

Eén probleem bij dit alles was dat het empirisme naast een empirische basis ook een sterk anti-metafysische, anti-speculatieve strekking vertoonde. Wetenschappers mochten niet teruggrijpen op abstracties waarbij onwaarneembare entiteiten werden gepostuleerd. Gedenkwaardig was het verzet van fysici als Mach en Ostwald tegen een realistische interpretatie van de zeer succesvolle atoomtheorie. Pas na het werk van Einstein en vooral Perrins studie *Les Atomes* uit 1913 kantelde de discussie (Starmans, 2013). Dat alles pleit op het eerste gezicht natuurlijk ook in de sociologie niet voor het postuleren van instituties en collectieve representaties als *causaal werkzame grootheden* of *krachten*, die de samenleving - al dan niet beschouwd als mechanisme of organisme- opbouwen. Ook leek het eenvoudiger de nieuwe statistiek toe de passen op individuen en hun eigenschappen, dan op voornoemde abstracte entiteiten. Mede door dit alles worstelden velen met de statistiek en de toepassing ervan in de 19e eeuwse sociologie was allerm minst een klare zege-tocht. Veeleer vertoonde deze toepassing kenmerken van de Processie van Echternach, die sedert de Middeleeuwen jaarlijks in de gelijknamige Luxemburgse stad wordt gehouden en waarbij door de gelovigen altemeerend drie stappen vooruit en twee stappen achteruit worden gezet. Vele tijdgenoten aarzelden, anderen bleken spijtoptanten en keerden op hun schreden terug van het statistische avontuur (Oberschall, 1989; Porter, 1986).

De centrale figuur in dit alles rond de eeuwwisseling was Emile Durkheim, die trachtte de verschillende stromingen met elkaar te verzoenen. Van een reductionisme van instituties naar individuen was bij hem geen sprake. Voorop stonden 'sociale feiten', die niet van individuele preferenties, overtuigingen of intenties afhangen, maar alleen van andere sociale feiten en feiten die gezamenlijk de structuren in de wereld constitueren. Beroemd werden zijn studies over religie als bindend element in moderne samenlevingen. In het klassieke *Le Suicide* (Durkheim, 1897) gebruikte Durkheim vele statistische gegevens, maar gaandeweg evolueerde van hij van aanhanger van statistiek naar scepticus en liet hij etnografische data prevaleren boven de Westerse numerieke data. Steeds meer richtte hij zich op het functionalisme dat in de biologie opgang maakte en paste hij evolutionaire principes toe op groepen en instituties. Anderen, zoals Max Weber, kozen wel de individuele benadering die mede zou leiden tot het methodologische individualisme in de sociale wetenschappen. Een en ander moet in dit korte essay

verder buiten beschouwing blijven, maar de problematiek van de status van theoretische concepten binnen een *niet-reductionistische* sociologie, de micro-macroloof en de vraag hoe statistiek op dit geaggregeerde of abstracte niveau kon worden toegepast blijft actueel. De ecologische correlatie is daarvan hoe dan ook een manifestatie en kan dus geenszins als overbodig of gratuit worden aangemerkt.

### Fishers handboek

Een ander opmerkelijk gezichtspunt van Robinson sluit hierbij aan en is om twee redenen historisch-filosofisch relevant. Allereerst constateert hij dat *'ecological correlations are used in an impressive number of quantitative sociological studies, some of which by now have attained the status of classics'*. Hij beseft dat zijn conclusie *'...has serious consequences, and that its effect appears wholly negative because it throws serious doubt upon the validity of a number of important studies made in recent years'*. Vervolgens noemt Robinson een groot aantal artikelen met naam en toenaam. De daarin gehanteerde technieken betreffen onder meer factoranalyse, canonische correlatie-analyse en andere multivariate technieken waaronder de door Sewall Wright ontwikkelde pad-analyse die later tot structurele vergelijkingsmodellen zou uitgroeien. In al deze technieken wordt getracht op basis van gemeten variabelen op individueel niveau, latente variabelen te creëren die corresponderen met theoretische termen, gepostuleerde entiteiten, et cetera. De vraag of deze ook een causale interpretatie toelaten is tot op heden een kernvraag binnen zowel methodologie als epistemologie. (Denis, 2006; Subramanian, 2007) Het teruggrijpen op en omarmen van deze methoden door sociologen kan uiteraard worden beschouwd als pogingen de (multivariate) statistiek te combineren met Durkheimiaanse inzichten.

Maar er is nog een ander, wezenlijk verschillend aspect. Het artikel van Robinson verschijnt precies 25 jaar na de poging van Ronald Fisher (1890-1962) met zijn *Statistical methods for research workers* (1925) statistische leken een handboek te verschaffen dat beoogde een state-of-the-art overzicht te geven op het gebied van *inferentiële statistiek*; onproblematische, 'veilige' technieken en inzichten, die de kloof tussen steekproef en populatie kunnen overbruggen. In weerwil van talloze herdrukken,

was dit streven deels illusoir. Dat bleek onder meer uit het feit dat onmiddellijk na het verschijnen van het boek de oude vete tussen Fisher enerzijds en Jerzy Neyman (1894-1981) en de familie Pearson anderzijds weer opblaide (Pearson, 1926). Veel technieken werden weliswaar grif toegepast in biologie en psychologie, maar waren nog allerminst uitgekristalliseerd. Gert Gigerenzer laat in zijn essay 'The inference experts' (Gigerenzer, 1989) zien hoe zelfs over de meest elementaire begrippen geen overeenstemming bestond: p-waarden, parameterschatting, betrouwbaarheidsintervallen, hypothese toetsing, et cetera. Hybride methoden en onversneden eclecticisme vierden hoogtij. Waar op methodologisch gebied convergentie, standaardisering en consensus soms mogelijk bleken (Blalock, A.D. de Groot) bleek op statistisch gebied anno 1950 zo'n handboek – waaraan nu juist binnen de kwantitatieve sociologie behoefte bestond – een brug te ver. De vraag of dit ook heden ten dage de status quo is, blijft hier buiten beschouwing (Starmans, 2018).

### Epiloog

Tot slot enige kanttekeningen.

Al is de relevantie van Robinsons paper voor multilevel analyse evident, het is de vraag of hij als pionier en wegbereider ervan moet worden beschouwd. Hij kon latere ontwikkelingen op het gebied vanzelfsprekend niet voorzien en het was evenmin zijn oogmerk de weg te plaveien voor een nieuwe methode. De studie van contingentietabellen, categoriale data-analyse en multilevel methoden zouden pas in de decennia daarna een hoge vlucht nemen. Omdat daarmee ook complexere hiërarchische structuren mogelijk worden, is de traditionele *interpretatie* van hiërarchische data als 'indivisible' individu versus groep, die wij hanteerden om aan te sluiten bij Robinsons definitie, uiteraard rigide. Het onderscheid wordt relatief als het individu zelf 'groep' wordt: herhaalde metingen genest binnen een proefpersoon, studies genest in een meta-analyse, geïnterviewden genest in een interviewer, et cetera. Dat geldt a fortiori in ketens van inbeddingen: genen, cellen, organen, individuen, families, wijken, steden, et cetera, waarin het niveau van beschouwing al helemaal geen a priori gegeven kan zijn. Hoe dan ook, in het licht van de beschikbare technieken en de status quo binnen de kwantitatieve sociologie rond

1950 was Robinsons caveat zeker terecht.

Het artikel kan dan ook zonder overdrijving een klassieker worden genoemd, temeer omdat het vrijwel onmiddellijk leidde tot reacties van statistici om het probleem te repareren, (Duncan, 1953; Goodman, 1853, 1959) en Robinsons werk een vernietigende kritiek gaf, die in de sociologie nog steeds nagalmt. In 2011 wijdde de *Journal of Epidemiology* een themanummer aan Robinsons paper en gaf het opnieuw uit. De socioloog Te Grotenhuis ontdekte verschillende nog niet eerder opgemerkte (kleine) fouten en onnauwkeurigheden in het artikel, voerde een replicatiestudie en een multilevel analyse uit op de oorspronkelijke data (Te Grotenhuis, 2011). Deze fouten van Robinson en zijn wellicht enigszins geringe empathisch vermogen betreffende de oogmerken en noden van de sociologie doen aan dit alles niets af.

Ecologische analyses vonden reed plaats in vroege 17e-eeuwse statistische studies van John Graunt en William Petty en blijven omwille van ethiek en privacy dikwijls ook onmisbaar in het tijdperk van data science. In weerwil van de volop beschikbare geavanceerde multilevel software maakt de hier beknopt geschetste problematiek duidelijk dat Fishers goedwillende 'research workers', die zich een weg moeten banen in de jungle van vele data-analytische technieken en statistische methoden, nog steeds voor een ongewis en hachelijk avontuur kunnen komen te staan. Zo beschouwd is de expliciete verwijzing naar de populaire *reality soap*, waar goed bedoelende avonturiers in onherbergzame gebieden trachten te overleven ten aanschouwen van een miljoenen publiek, wellicht niet geheel misplaatst, zeker tegen de achtergrond van de eerder gesignaleerde hyperbolen in tijden van data science en big data.

### LITERATUUR

- Denis, D.J., & Legerski, J. (2006). Causal Modelling and the origins of Path Analysis. *Theory and Science*, 7(1).
- Duncan, O.D., & Davis, B. (1953). An alternative to ecological correlation. *American Sociological Review*, 18, 665–66.
- Durkheim, E. (1897). *Le suicide*. Paris: F. Alcan; English translation by J A Spalding, 1951. Toronto: Free Press, Collier-MacMillan, Toronto.
- Gigerenzer, G. (1989). *The Empire of Chance; how probability changed science and everyday life*. Cambridge: Cambridge University Press.
- Goodman, L. (1953). Ecological regression and the behavior of individuals. *American Sociological Review*, 18, 663–64.
- Goodman, L. (1959). Some alternatives to ecological correla-

tion. *American Journal of Sociology*, 64, 610–625.

- Grotenhuis, M. Te, Eisinga, R., & Subramanian, S.V. (2011). Robinson's Ecological Correlations and the Behavior of Individuals: methodological corrections. *The International Journal of Epidemiology*, 40(4).
- Hacking, I. (1989). *The Taming of Chance*. Oxford: Oxford Publishers.
- Oberschall, A. (1989). The two empirical roots of Social Theory and the Probabilistic Revolution. In: Lorenz Kruger (ed.) *The Scientific Revolution, Part II*. Cambridge: MIT Press, Cambridge.
- Selvin, H.C., (1958). Durkheim's suicide and problems of empirical research. *American Journal of Sociology*, 63, 607–619.
- King, G. (1997). *A Solution to the Ecological Inference Problem: Reconstructing Individual Behavior from Aggregate Data*. Princeton: Princeton University Press.
- Kreft, I., & De Leeuw, E.D. (1987). The see-saw-effect: a multi-level problem? A reanalysis of some findings of Hox and De Leeuw. *Quality and Quantity*, 22, 127–137.
- Pearson, K. (1892) *The Grammar of Science*. London.
- Pearson, E., (1926). Review (1926) of Statistical Methods for Research Workers (R. A. Fisher). *Science Progress*, 20, 733–734.
- Porter, T.M. (1986). *The Rise of Statistical Thinking, 1829–1900*. Princeton, New Jersey: Princeton University Press.
- Robinson, W.S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, 15, 351–357.
- Stamhuis, J.H., & Knecht-van Eekelen, A. de (1992). *De met cijfers bedekte negentiende eeuw: Toepassing van statistiek en waarschijnlijkheidsrekening in Nederland en Vlaanderen 1840–1920*. Rotterdam: Erasmus Publishing.
- Starmans, R.J.C.M. (2013). Perrin and the Partisans of the Void (in Dutch). *Nederlands Tijdschrift voor Natuurkunde*, 79(12).
- Starmans, R.J.C.M. (2018). The Predicament of Truth: on Statistics, Causality, Physics and the Philosophy of Science. In: Mark J. Van der Laan & Sherri Rose (Eds.), *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*. Springer Series in Statistics. New York: Springer.
- Subramanian, S.V. (2004) The relevance of multilevel statistical methods for identifying causal neighborhood effects. *Social Science & Medicine*, 58, 1961–1967.
- Subramanian, S.V., Glymour, M.M., & Kawachi, I. (2007) Identifying causal ecologic effects on health: a methodologic assessment. In: Galea S, editor. *Macrosocial Determinants of Population Health*. New York: Springer Media. pp. 301–331.

RICHARD STARMANS is verbonden aan de Faculteit Bètawetenschappen (Department of Information and Computing Sciences) van de Universiteit Utrecht. Hij doet onderzoek op het snijvlak van filosofie, statistiek en informatica. E-mail: starmans@cs.uu.nl