

Kwalitatieve Statistiek en Econometrie
Een toepassing op regionale verschillen
in Nederland.

Prof.dr. P. Nijkamp*

Dr. P. Rietveld*

Samenvatting

Dit paper bevat een beknopt overzicht van statistische en econometrische methoden voor kwalitatieve (bijvoorbeeld ordinale) data. Speciale aandacht wordt besteed aan de overeenkomst die er bestaat tussen de structuur van Kendalls rangcorrelatiecoëfficiënt voor ordinale data en Pearsons product-moment correlatiecoëfficiënt voor cardinale data. Uitgaande van deze overeenkomst worden suggesties gedaan voor de opzet van diverse multivariate methoden voor ordinale data. Deze methoden worden toegepast op Nederlandse gegevens over interregionale verschillen betreffende sociaal-economische condities, milieu kwaliteit en infrastructuur. Daarbij wordt speciale aandacht besteed aan de vraag tot op welke hoogte de resultaten gebaseerd op ordinale data lijken op de resultaten die bereikt worden op grond van de corresponderende cardinale data.

*De auteurs zijn verbonden aan de Economische Faculteit van de Vrije Universiteit, Postbus 7161, 1007 MC Amsterdam.
De delen van dit artikel berusten op P. Nijkamp en P. Rietveld, *Soft Econometrics as a Tool for Regional Discrepancy Analysis, Papers of the Regional Science Association*, 1982 (te verschijnen).

1. Inleiding

Een van de kenmerken van empirisch (regionaal) economisch onderzoek is dat het bijna uitsluitend gebruik maakt van gegevens gemeten op cardinale schaal, zoals inkomen, bevolkingsomvang of kapitaalgoederenvoorraad. Voor een aantal onderzoeksvelden binnen de economie is dit wellicht een aanvaardbare stand van zaken, maar zeker niet voor alle. Terreinen waarbij men op problemen stuit als men zich uitsluitend baseert op cardinale gegevens zijn b.v. de analyse van woonvoorkeuren, plan/project evaluatie studies en milieuvraagstukken.

Dit soort problemen wordt gekenmerkt door aspecten die in principe of per definitie niet-cardinaal van aard zijn. Deze aspecten kunnen betrekking hebben op variabelen die op ordinale of nominale schaal gemeten worden, zoals het type woning dat men zoekt, de volgorde van belangrijkheid van evaluatie criteria of de waardering van een bepaald landschap. Wanneer men deze aspecten buiten beschouwing laat in de analyse omdat ze niet hanteerbaar zijn met cardinale methoden, is het gevaar reëel dat men tot onbetrouwbare onderzoek resultaten komt.

Er zijn diverse redenen waarom in de economie cardinale gegevens niet altijd voorhanden zijn :

- het voorkomen van meetfouten
- het bestaan van onbetrouwbare statistische informatie
- de aard van veel verschijnselen (b.v. milieuwaardering)
- de beperkte mogelijkheden van de mens om prioriteiten en percepties in cardinale termen te uiten.

Het is opvallend dat in diverse disciplines in de sociale wetenschappen (zoals de sociologie en psychologie) de laatste tientallen jaren ruime aandacht is geschonken aan de ontwikkeling van methoden om data van een dergelijk laag meetniveau te analyseren, terwijl deze ontwikkeling gedeels aan het (regionaal) economisch onderzoek is voorbijgegaan.

In par. 2 en 3 zal een indruk worden gegeven van de vorderingen die de laatste tijd gemaakt zijn op het terrein van de analyse van 'zachte' data. Daarbij zal een onderscheid worden gemaakt tussen statistische en econometrische methoden, alhoewel opgemerkt moet worden dat de overgang tussen deze twee groepen vloeiend is. In par. 4 en 5 zullen enkele statistische en econometrische methoden meer in detail worden besproken. In par. 6 zal een aantal van de gepresenteerde methoden worden toegepast op de analyse van interregionale verschillen in Nederland. Hierbij zal tevens een vergelijking worden uitgevoerd van de resultaten gebaseerd op ordinale en cardinale gegevens.

2. Een Overzicht van Kwalitatieve Statistische Methoden

Een gebruikelijke manier om verschijnselen te bestuderen die moeilijk op een cardinale schaal kunnen worden gemeten (zoals geluk, satisfactie, kwaliteit van het bestaan) is de introductie van proxy-variabelen die wel cardinaal meetbaar zijn. Adelman en Morris (1974) wijzen er evenwel terecht op dat deze operatie gemakkelijk tot ongeoorloofde simplificaties leidt.

Het is daarom de moeite waard om te bezien of er methoden zijn om niet-cardinale gegevens te analyseren.

Een reeds geruime tijd bekende benaderingswijze tot zachte data (met name ordinale informatie) is het gebruik van rangcorrelatie analyse, gebaseerd op niet-parametrische statistiek (zie Siegel, 1956). De meest bekende rangcorrelatiecoëfficiënten zijn die van Spearman en Kendall. Deze correlatiecoëfficiënten zijn enkelvoudig: ze meten de samenhang tussen twee variabelen. Het blijkt niet eenvoudig om op consistent wijze multipale rangcorrelatiecoëfficiënten af te leiden (zie b.v. Blalock, 1976 en Nijkamp en Rietveld, 1981). Sommige rangcorrelatiecoëfficiënten zijn bovendien gebaseerd op nauwelijks geoorloofde numerieke operaties met ordinale gegevens zodat men zich bewust moet zijn van de ernstige beperkingen die in sommige gevallen voor rangcorrelatie analyse gelden.

Een bijzonder bruikbare methode voor ordinale data analyse is multi-dimensionale schaling (MDS) analyse (zie voor een overzicht Nijkamp, 1979). MDS methoden kunnen gebruikt worden om ordinale gegevens in cardinale uitkomsten te transformeren op basis van twee condities: 1) een reductie van de rang van de ordinale data matrix, en 2) een "goodness-of-fit" criterium dat garandeert dat de nieuwe cardinale configuratie in maximale mate overeenstemt met de oorspronkelijke ordinale informatie. MDS analyses hebben hun bruikbaarheid in een groot aantal toepassingen bewezen: perceptie en preferentie analyse, recreatie analyse, analyse van milieukwaliteit, stadsvernieuingsbeleid en multicriteria waarderingsmethoden (zie b.v. Voogd, 1978).

Tenslotte dient het gebruik van logit modellen en meer recentelijk, probit modellen te worden genoemd (zie Domencich en McFadden, 1975). Deze modellen zijn speciaal ontworpen voor de analyse van discrete gegevens. Deze analyses zijn gebaseerd op een waarschijnlijkheidsbenadering waarbij de frequentie waarin een bepaald overschijnsel zich voordoet als invoergegeven gebruikt wordt.

Logit en probit analyse spelen met name een belangrijke rol in verklarende modellen voor kwalitatieve verschijnselen zoals zal worden aangetoond in de volgende paragraaf.

Diverse ordinale data methoden - soms in combinatie met logit benaderingen - zijn voorts ontwikkeld op het gebied van de analyse van kruistabellen (zie Grizzle et al., 1969). In dit gebied is ook gebruik gemaakt van chi-kwadraat benaderingen, regressies met dummy variabelen en variantie analyse.

3. Een Overzicht van Kwalitatieve Econometrische Methoden

Kwalitatieve econometrische modellen worden gekenmerkt door het voorkomen van zachte variabelen als afhankelijke of verklarende variabelen. Gebruik makend van Wrigley (1979) kan men de volgende onderverdeling voor het meetniveau van de variabelen maken :

		afhankelijke variabele	
		cardinaal	zacht
verklarende variabele	cardinaal	I	IV
	gemengd	II	V
	zacht	III	VI

Fig. 1. Meetniveaus van verklarende en afhankelijke variabelen

Een gebruikelijke manier om met zachte informatie om te gaan (voornamelijk nominale data) is het gebruik van dummy variabelen. Dummy variabelen zijn vaak in econometrisch onderzoek gebruikt om niet-cardinaal meetbare verklarende variabelen te kunnen opnemen zoals geslacht, huwelijks staat, beroepsgroep, enz. (zie Theil, 1971). Dummy variabelen kunnen de waarden 0 of 1 aannemen om aan te geven of een waarneming al dan niet in een bepaalde nominale klasse valt. Ze zijn vnl. gebruikt om sociologische en demografische variabelen een plaats te geven in econometrisch onderzoek. Ze kunnen worden geplaatst in klasse II of III van Fig. 1. Een duidelijk nadeel van dummy variabelen is dat ze aanleiding kunnen geven tot een groot aantal verklarende variabelen wanneer diverse klassen worden onderscheiden. Voorts blijkt een 0-1 indicator voor sommige variabelen te leiden tot een inefficiënt gebruik van de beschikbare informatie (bijvoorbeeld wanneer een variabele gemeten is in termen van "goed, normaal, slecht"). Voor dit soort gevallen bestaan andere meer geschikte econometrische methoden (zie vervolg).

Een andere benadering van kwalitatieve econometrische methoden wordt gevormd door MDS methoden. Als de verzameling verklarende variabelen uit ordinale data bestaat (of een deelverzameling van ordinale data bevat), kan MDS analyse gebruikt worden om de ordinale informatie in cardinale informatie van een lagere dimensionaliteit te transformeren. Vervolgens kan deze cardinale informatie gebruikt worden als data voor traditionele schattingsmethoden. Deze benadering is verwant aan principale componenten analyse die een verzameling gegevens reduceert tot een kleinere verzameling die gebruikt kan worden in een regressiemodel. Deze MDS methoden behoren eveneens tot klasse II of III van Fig. 1.

Een andere klasse kwalitatieve econometrische modellen wordt gevormd door de eerdergenoemde logit en probit modellen. Deze stochastische modellen zijn met name geschikt wanneer de afhankelijke variabele niet-cardinaal meetbaar is. Deze modellen kunnen geplaatst worden in de klassen IV, V en VI in Figuur 1. Toepassingen van logit modellen kunnen ondermeer gevonden worden in het veld van de categorale data analyse. Categorale data zijn dichtotome of polytome data; ze zijn vaak het resultaat van enquêtes waarin de respondenten moeten aangeven of een bepaald attribuut al dan niet van toepassing is of belangrijk wordt geacht. De frequentie van bevestigende of ontkennende antwoorden kan dan gebruikt worden als input voor het logit model (zie b.v. Theil, 1971, en Bishop et al., 1975).

Een andere toepassing van logit analyse kan gevonden worden in McCullagh (1980) die een algemene klasse van regressie modellen voor ordinale gegevens heeft ontwikkeld. Deze modellen zijn gebaseerd op diverse vormen van stochastische ordening van een ordinale structuur. De auteur stelt in het bijzonder het gebruik van twee modellen voor, namelijk het "proportional odds" model (gebaseerd op een lineair of niet-lineair logit model voor de geordende categorieën van een afhankelijke variabele, gegeven de waarden van de onafhankelijke variabelen) en het "proportional hazards" model (gebaseerd op een log transformatie van een hazard functie voor een ordinale afhankelijke variabele. Deze modellen behoren tot klassen IV-VI van Figuur 1.

Een laatste manier om ordinale informatie te benutten zowel voor de afhankelijke als de verklarende variabelen, is te vinden in Nijkamp en Rietveld (1981). Deze benadering is gebaseerd op een paarsgewijze vergelijking van ordinale data gevolgd door een dominantie analyse via zogenaamde regimes (verzamelingen van dominantie en non-dominantie relaties). Vervolgens kunnen de frequenties van zulke regimes gebruikt worden in een lineair logit model zodat men waarschijnlijkheidsuitspraken kan afleiden betreffende het effect van een specifiek regime van verklarende variabelen op het voorkomen van dominantie in de afhankelijke variabelen. De laatste modellen vallen in categorie VI. Voor een meer gedetailleerde beschrijving zij verwezen naar par.5.

4. Kwalitatieve Statistische Methoden : een meer gedetailleerde beschrijving van enkele methoden

In deze paragraaf zullen we een meer gedetailleerde beschrijving geven van enkele statistische methoden voor ordinale gegevens die in de laatste paragraaf zullen worden toegepast. Het uitgangspunt van de methoden is een verzameling gegevens over attributen van bepaalde eenheden. In dit stuk nemen we hierbij aan dat het gaat om indicatoren van milieu, infrastructuur en de sociaal economische ontwikkeling van een aantal regio's. Indien we R regio's en J indicatoren onderscheiden kan de volgende profielmatrix geconstrueerd worden (zie ook Nijkamp en Rietveld, 1982).

$$X = \begin{bmatrix} x_{11} & \dots & x_{1R} \\ \vdots & & \vdots \\ x_{J1} & \dots & x_{JR} \end{bmatrix} \quad (1)$$

Wanneer slechts ordinale gegevens voorhanden zijn bevatten de rijen van X rangnummers van 1 tot R, bijvoorbeeld (2, 1, 5, 3, 4). Een score 4 voor de laatste regio betekent eenvoudig dat deze regio voor de desbetreffende indicator de op een na gunstige plaats inneemt. Het is duidelijk dat aan de uitkomst 4 niet de gebruikelijke kwantitatieve betekenis kan worden gehecht zodat elementaire operaties zoals optellen of vermenigvuldigen niet op een ordinale datamatrix X mogen worden toegepast. Achtereenvolgens zullen nu de volgende statistische methoden worden behandeld: correlatie analyse, principale componenten analyse, interdependentie analyse en cluster analyse.

4.1. Correlatie Analyse

Ordinale gegevens bevatten slechts informatie over het teken van het verschil tussen twee waarnemingen en niet over de omvang van dat verschil. Daarom ging Kendall bij de ontwikkeling van een rangcorrelatiecoëfficiënt als volgt te werk. Beschouw alle paren regio's voor twee indicatoren. Een tweetal regio's vormen een concordant paar als voor beide indicatoren geldt dat de rangorde van de twee regio's overeenkomt. Wanneer de rangorde tegengesteld is, spreken we van een discordant paar.

Het totaal aantal concordante en discordante paren wordt weergegeven door respectievelijk S^+ en S^- . Voorts worden T_x en T_y gedefinieerd als de aantallen paren regio's waarvoor een gelijke uitkomst wordt gevonden bij indicator x, respectievelijk y. Dan kan Kendalls rangcorrelatiecoëfficiënt beschreven worden als (zie Kendall, 1970):

$$\tau_b = \frac{S^+ - S^-}{\sqrt{(S^+ + S^- + T_x)(S^+ + S^- + T_y)}} \quad (2)$$

Deze coëfficiënt kan waarden aannemen tussen -1 en +1. We geven nog een andere manier om τ_b af te leiden aangezien deze ook van belang blijkt bij de presentatie van andere methoden in dit stuk.

Gegeven een ordinaal gemeten variabele x met waarnemingen ($x_1, x_2, \dots, x_R$) definiëren we een nieuwe variabele $\tilde{x}$ met behulp van de volgende transformatie voor alle $R(R-1)$ paren regio's

$$\begin{cases} \tilde{x}_{rs} = 1 & \text{als } x_r > x_s \\ \tilde{x}_{rs} = 0 & \text{als } x_r = x_s \\ \tilde{x}_{rs} = -1 & \text{als } x_r < x_s \end{cases} \quad (3)$$

Op dezelfde wijze kan de variabele $\tilde{y}$ worden gedefinieerd. Dan kan bewezen worden (zie Nijkamp en Rietveld, 1981), dat :

$$\tau_b = \frac{\sum \tilde{x}_{rs} \tilde{y}_{rs}}{\sqrt{\sum \tilde{x}_{rs}^2 \cdot \sum \tilde{y}_{rs}^2}} \quad (4)$$

Het opmerkelijke van (4) is, dat deze uitdrukking voor een rangcorrelatiecoëfficiënt identiek is aan de gebruikelijke correlatiecoëfficiënt voor cardinale data. Deze conclusie is belangrijk aangezien ze suggereert dat ordinale data na transformatie m.b.v. (3) kunnen worden behandeld alsof het cardinale data zijn. Alhoewel aan deze suggestie een aantal haken en ogen zitten waaraan we hier voorbijgaan, is deze interessant genoeg om ook bij andere methoden toe passen.

Een probleem dat hierbij vermeld moet worden is dat de data na de betreffende transformatie betrekking hebben op paren waarnemingen i.p.v. op individuele waarnemingen. Als gevolg hiervan zijn de 'waarnemingen' in $\tilde{x}$ en $\tilde{y}$ niet onafhankelijk, wat repercussies heeft voor de mogelijkheid om statistische toetsen toe te passen.

4.2 Principale Componenten Analyse

Het doel van principale componenten analyse is het weergeven van J variabelen dooreen kleiner aantalte constateren variabelen (componenten genaamd) met een hoge mate van nauwkeurigheid. Wanneer men opzoek is naar slechts één component om J ordinale indicatoren weer te geven kan men als volgt te werk gaan. Bepaal een zodanige rangorde van regio's dat de som van de J rangcorrelatiecoëfficiënten tussen deze rangorde en de J oorspronkelijke variabelen zo groot mogelijk is.

Een andere methode om principale componenten af te leiden op grond van een ordinale profielmatrix X is het toepassen van transformatie (3) waardoor een nieuwe matrix $\tilde{X}$ ontstaat met $R(R-1)$ rijen en J kolommen. Vervolgens kan men de gebruikelijke stappen ondernemen om de principale componenten van $\tilde{X}$ te bepalen.

Een nadeel van principale componenten analyse (zowel voor cardinale als voor ordinale data) is het feit dat nieuwe variabelen geconstrueerd worden die weliswaar op basis van factor ladingen geïnterpreteerd kunnen worden, maar die op zich geen duidelijke betekenis behoeven te hebben. Een methode die aan dit bezwaar tegemoet komt is interdependentie analyse, het onderwerp van de volgende paragraaf.

4.3 Interdependentie Analyse

Interdependentie analyse is een techniek om uit een verzameling van J variabelen een aantal variabelen te kiezen die tezamen een zo goed mogelijke representatie vormen van de J variabelen. Met "een zo goed mogelijke representatie" wordt daarbij bedoeld dat de minimum multiple correlatiecoëfficiënt tussen aan de ene kant ieder van de niet geselecteerde variabelen en aan de andere kant de groep wel geselecteerde variabelen zo groot mogelijk moet zijn. (zie Boyce et al. 1974).

Interdependentie analyse kan voor ordinale data uitgevoerd worden na toepassing van transformatie (3) op X .

4.4 Cluster Analyse

Het doel van cluster analyse is de afleiding van op elkaar gelijkende verzamelingen variabelen (of regio's). Er bestaan veel soorten cluster methoden (zie b.v. Hartigan, 1975); zij onderscheiden zich bijvoorbeeld door de manier waarop de overeenkomst tussen twee variabelen wordt gedefinieerd en de wijze waarop clusters worden gevormd (hiërarchisch versus niet-hiërarchisch).

Het voornaamste probleem met ordinale gegevens bij cluster analyse is dat een bruikbaar criterium moet worden gevonden om de overeenkomst tussen variabelen te meten. De in par. 4.1 besproken rangcorrelatiecoëfficiënt is daarvoor een geschikte kandidaat.

5. Kwalitatieve Econometrische Methoden : een meer gedetailleerde beschrijving van de regime methode

Stel dat men het verband wil schatten tussen een afhankelijke variabele y en verklarende variabelen $x_1, \dots, x_j$ en dat gegevens beschikbaar zijn over R regio's. Dan kan een lineair verband als volgt worden geformuleerd:

$$y_r = \alpha_0 + \alpha_1 x_{1r} + \dots + \alpha_j x_{jr} \quad r=1, \dots, R \quad (5)$$

waarbij de α 's de te schatten coëfficiënten zijn.

We stellen ons nu de vraag of een dergelijke vergelijking ook geschat kan worden als y en x_j slechts ordinaal meetbaar zijn. We zullen aantonen dat als gebruik gemaakt wordt van een paarsgewijze vergelijking van regio's analoog aan de benadering zoals uiteengezet in par. 4.1, inderdaad een schatting kan worden uitgevoerd van het verband tussen de x_j en y . Daartoe beschrijven we (5) eerst zodat deze in termen van verschillen tussen regio's gaat luiden:

$$y_r - y_s = \alpha_1 (x_{1r} - x_{1s}) + \dots + \alpha_j (x_{jr} - x_{js}) \quad \begin{matrix} r,s=1,\dots,R \\ r \neq s \end{matrix} \quad (6)$$

Vervolgens passen we transformatie (3) toe op de uitkomsten voor alle paren regio's. Voor ieder paar kan dan de vector met verklarende variabelen geschreven worden als een reeks van J getallen die de waarden 1, 0 en -1 kunnen aannemen. Een bepaalde combinatie van de waarden noemt men een regime. In principe zijn er 3^J verschillende regime's.

Beschouw een bepaald regime m . We definiëren de structuur van regime m als $(z_{m1}, \dots, z_{mJ})$, waarbij de z_{mj} de waarden 1, 0 en -1 kunnen aannemen. F_m is het aantal paren waarnemingen (regio's) dat aan regime m voldoet. Bij deze paren waarnemingen doen zich drie mogelijkheden voor: y_r kan zijn: groter dan, gelijk aan, of kleiner dan y_s . De aantallen van deze paren worden weergegeven door respectievelijk F_{m1} , F_{m0} en F_{m-1} . Uiteraard geldt dat $F_{m1} + F_{m0} + F_{m-1} = F_m$. De relatieve aandelen van deze drie gevallen worden weergegeven door: f_{m1} , f_{m0} en f_{m-1} .

Dus : $f_{m1} = F_{m1}/F_m$, etc.

Het verband tussen de verklarende variabelen en de afhankelijke variabele kan nu als volgt beschreven worden:

$$\ln\left(\frac{f_{m1}}{f_{m-1}}\right) = \beta_1 z_{m1} + \beta_2 z_{m2} + \dots + \beta_J z_{mJ} \quad (7)$$

Voor de achtergronden van deze specificatie verwijzen we naar Theil, 1981. Vergelijking (7) kan worden geschat m.b.v. gewogen kleinste kwadraten.

We merken op, dat de ordinale en cardinale analyses in het geval van de regime methode in het algemene niet ongeveer gelijke coëfficiënten en R^2 zullen leveren. Hierin onderscheidt de regime method zich van de methoden beschreven in par. 4.

6. Toepassing

De concepten en methoden die in de voorgaande paragrafen werden gepresenteerd zullen worden toegepast op Nederlandse regionale gegevens¹⁾ uit de periode 1976-1978. De gegevens hebben betrekking op 40 regio's (COROP gebieden) en 13 indicatoren ($R = 40$, $J = 13$). De regio's zijn afgebeeld in Fig.2.



Figuur 2. De locatie van de COROP gebieden

¹⁾ De auteurs danken Wouter van Veenendaal en Jan Broersma voor hun hulp bij de data verzameling. De basis gegevens voor dit artikel zijn te vinden in W.M. van Veenendaal, 1932.

Per regio zijn de volgende kenmerken in beschouwing genomen¹⁾:

Sociaal-economisch profiel

- s_1 : belastbaar inkomen per hoofd
- s_2 : werkloosheidspercentage
- s_3 : vermogen per hoofd
- s_4 : index voor de kosten van levensonderhoud (in het bijzonder huur).

Milieuprofiel

- s_5 : bevolkingsdichtheid
- s_6 : natuurlijk milieu als percentage van provinciale oppervlakte
- s_7 : industrialisatie-index
- s_8 : emissie-index voor vervuiling

Infrastructuur-profiel

- s_9 : dichtheid van vervoersnetwerk
- s_{10} : index voor culturele voorzieningen per hoofd
- s_{11} : index voor onderwijsvoorzieningen per hoofd
- s_{12} : afstand t.o.v. het centrum van Nederland
- s_{13} : index voor medische voorzieningen per hoofd.

Om de bruikbaarheid van analyse methoden voor ordinale data te testen is als volgt te werk gegaan. Eerst wordt voor iedere methode de analyse uitgevoerd op de gebruikelijke wijze met behulp van de beschikbare cardinale gegevens. Vervolgens wordt aangenomen dat slechts de rangorde van de cardinale gegevens bekend is, m.a.w. dat de variabelen slechts op ordinaal niveau gemeten zijn. Op deze ordinale gegevens worden dan de voor dit doel ontwikkelde methoden toegepast. Tenslotte worden de uitkomsten van beide benaderingen met elkaar vergeleken om te zien in hoeverre de op ordinaal niveau gemeten variabelen aanleiding geven tot gelijklopende conclusies als de op cardinaal niveau gemeten variabelen.

1) Deze kenmerken zijn ook gehanteerd in Blommenstein, Nijkamp en Rietveld, 1981 voor provinciale data. In de analyses zijn de variabelen 2, 4, 5, 7, 8 en 12 voorvermenigvuldigd met een factor -1 zodat voor alle variabelen geldt dat een hogere waarde geprefereerd worden boven een lagere.

6.1.a Correlatie Analyse: Cardinaal

De correlatiecoëfficiënten tussen de 13 variabelen zijn weergegeven in Tabel 1. De correlaties binnen de drie onderscheiden blokken blijken

1	2	3	4	5	6	7	8	9	10	11	12	13
1.00	.57	.60	-.17	-.84	.30	-.63	-.83	.77	.05	.15	.62	.55
.57	1.00	.49	-.13	-.40	.31	-.36	-.37	.25	-.15	.17	.72	.24
.60	.49	1.00	.06	-.39	.29	-.11	-.36	.36	.12	.15	.28	.46
-.17	-.13	.06	1.00	.04	-.24	-.09	.01	-.01	.17	-.06	-.28	.04
-.84	-.40	-.39	.04	1.00	-.19	.75	.97	-.94	-.10	-.01	-.52	-.47
.30	.31	.29	-.24	-.19	1.00	-.17	-.09	.15	-.04	.04	.25	.19
-.63	-.36	-.11	-.09	.75	-.17	1.00	.69	-.68	-.03	.03	-.33	-.40
-.83	-.37	-.36	.01	.97	-.09	.69	1.00	-.91	-.11	-.01	-.52	-.47
.77	.25	.36	-.01	-.94	.15	-.68	-.91	1.00	.04	-.02	.46	.38
.05	-.15	.12	.17	-.10	-.04	-.03	-.11	.04	1.00	.01	-.17	.09
.15	.17	.15	-.06	-.01	.04	.03	-.01	-.02	.01	1.00	.12	.27
.62	.72	.28	-.28	-.52	.25	-.33	-.52	.46	-.17	.12	1.00	.18
.55	.24	.46	.04	-.47	.19	-.40	-.47	.38	.09	.27	.18	1.00

Tabel 1. Correlatie matrix voor cardinale gegevens.

overwegend positief te zijn. Relatief hoge correlaties worden gevonden voor het sociaal economische profiel (variabelen 1 - 3) en het milieu-profiel (variabelen 5, 7 en 8).

Op grond van de correlatiematrix kunnen twee duidelijk verschillende deelverzamelingen van variabelen onderkend worden.

De eerste deelverzameling S_1 bestaat uit de variabelen 5, 7 en 8. De tweede verzameling bestaat uit de variabelen 1, 2, 3, 6, 9, 12 en 13. Binnen beide deelverzamelingen zijn alle correlaties positief, terwijl ze tussen alle paren variabelen uit verschillende deelverzamelingen negatief zijn. De variabelen in S_1 zijn duidelijk gerelateerd aan de negatieve aspecten van geurbaniseerde regio's, terwijl de variabelen uit S_2 juist voornamelijk betrekking hebben op de positieve aspecten van deze regio's. Dit resultaat suggereert dat in Nederland een aanzienlijk deel van de interregionale welvaartsverschillen gerelateerd is aan de urbanisatie graad. Er zijn slechts drie variabelen (4, 10 en 11) die niet in één van beide deelverzamelingen kunnen worden opgenomen.

6.1.b Correlatie-Analyse: Ordinaal

De rangcorrelatiecoëfficiënt tussen de variabelen zijn te vinden in Tabel 2.

1	2	3	4	5	6	7	8	9	10	11	12	13
1.00	.44	.25	-.19	-.55	.13	-.48	-.57	.50	-.07	-.08	.55	.40
.44	1.00	.37	-.06	-.30	.11	-.24	-.34	.25	-.16	-.12	.59	.21
.25	.37	1.00	-.04	-.04	.02	-.01	-.10	.02	.03	.13	.19	.28
-.19	-.06	-.04	1.00	.16	-.14	.06	.13	-.14	.10	-.01	-.19	-.13
-.55	-.30	-.04	.16	1.00	-.11	.74	.82	-.74	.11	.04	-.53	-.43
.13	.11	.02	-.14	-.11	1.00	-.06	-.05	.10	.07	.10	.09	.24
-.48	-.24	-.01	.06	.74	-.06	1.00	.70	-.67	.03	-.01	-.36	-.45
-.57	-.34	-.10	.13	.82	-.05	.70	1.00	-.72	.13	.06	-.57	-.40
.50	.25	.02	-.14	-.74	.10	-.67	-.72	1.00	-.13	-.07	.50	.33
-.07	-.16	.03	.10	.11	.07	.03	.13	-.13	1.00	.37	-.23	.07
-.08	-.12	.13	-.01	.04	.10	-.01	.06	-.07	.37	1.00	-.21	.20
.55	.59	.19	-.19	-.53	.09	-.36	-.57	.50	-.23	-.21	1.00	.21
.40	.21	.28	-.13	-.43	.24	-.45	-.40	.33	.07	.20	.21	1.00

Tabel 2. Correlatie matrix voor ordinale gegevens.

Wanneer we de uitkomsten van Tabel 2 met Tabel 1 vergelijken vinden we een sterk gelijkend patroon, zij het dat de orde van grootte van de coëfficiënten in Tabel 2 wat kleiner is. Ook in Tabel 2 kunnen twee deelverzamelingen van variabelen onderscheiden worden met hoge interne homogeniteit en een sterke externe heterogeniteit. We vinden:

$S_1 = \{4, 5, 7, 8\}$ en $S_2 = \{1, 2, 3, 6, 9, 12, 13\}$ Deze twee deelverzamelingen vertonen een sterke gelijkenis met de verzamelingen uit par. 5.1.a, zodat we mogen concluderen dat voor de beschouwde gegevens correlatie analyse vergelijkbare uitkomsten oplevert voor cardinale en ordinale data.

6.2.a Principale Componenten Analyse: Cardinaal

We beperken ons hier tot de eerste principale component. Deze blijkt 42% van de totale variatie in X te verklaren. De factorladingen zijn te vinden in de eerste rij van Tabel 3. De ladingen zijn in overeenstemming

variabele	1	2	3	4	5	6	7	8	9	10	11	12	13
cardinale data	.94	.61	.57	-.10	-.93	.32	-.73	-.90	.86	.04	.13	.67	.69
ordinale data	.76	.54	.23	-.23	-.87	.17	-.77	-.87	.80	.17	-.09	.72	.54

Tabel 3. Ladingen van de eerste principale component

met het patroon dat gevonden werd bij de correlatie analyse de elementen van S_1 hebben negatieve ladingen terwijl die bij S_2 juist positief zijn. De ladingen van de overige variabelen wijken niet sterk af van nul. Het is dus redelijk de eerste principale component te interpreteren als een indicator van de urbanisatie die positief is gerelateerd aan de voornaamste sociaal economische variabelen en negatief aan de voornaamste milieu variabelen.

6.2.b Principale Componenten Analyse: Ordinaal

De eerste principale component is berekend door transformatie (3) toe te passen op de ordinale data matrix en vervolgens de standaard principale componenten analyse uit te voeren voor de getransformeerde matrix $\tilde{X}$. De eerste component blijkt 36% te verklaren van de totale variatie in $\tilde{X}$. Zijn ladingen zijn te vinden in de tweede rij van Tabel 3. De overeenkomst tussen de resultaat voor ordinale en cardinale data is opvallend.

6.3.a Interdependentie Analyse: Cardinaal

In Tabel 4 zijn de resultaten weergegeven van een cardinale interdependentie analyse. In de laatste kolom is de minimale multiple correlatiecoëfficiënt te vinden die resulteert uit de regressie van iedere variabele met de geselecteerde variabelen die daarbij als verklarende variabelen fungeren. Slechts variabele 12 blijkt in drie opeenvolgende selecties voor te komen.

aantal geselecteerde variabelen	geselecteerde variabelen	minimale multiple correlatie coëfficiënt
1	2	.02
2	12, 13	.04
3	8, 12, 13	.08
4	6, 10, 11, 12	.13

Tabel 4. Interdependentie analyse (cardinaal)

Overigens is de verzameling geselecteerde variabelen niet erg stabiel. Dit is een weinig bevredigend resultaat, want het impliceert dat de samenstelling van selectie in sterke mate afhangt van de omvang van de selectie.

6.3.b Interdependentie Analyse: Ordinaal

De resultaten van de ordinale interdependentie analyse zijn te vinden in Tabel 5. Ook hier blijkt zich een tamelijk instabiel patroon van

aantal geselecteerde variabelen	geselecteerde variabelen	minimale multiple correlatie coëfficiënt
1	12	.01
2	12, 13	.04
3	4, 10, 13	.07
4	1, 4, 6, 11	.09

Tabel 5. Interdependentie analyse (ordinaal).

geselecteerde variabelen voor te doen. De selecties bij ordinale en cardinale data vertonen weliswaar gelijkenis maar zijn in de meeste gevallen zeker niet identiek.

6.4.a Cluster Analyse: Cardinaal

In deze paragraaf worden de resultaten van een clustering van variabelen gepresenteerd. Daarbij zijn correlatie coëfficiënten gebruikt als maatstaf voor de gelijkenis van variabelen. De clusters zijn op zodanige wijze gevormd dat de interne homogeniteit - gemeten als de minimale correlatie tussen ieder paar variabelen in een cluster - in iedere cluster zo groot mogelijk is. De resultaten zijn te vinden in Tabel 6. Uit deze tabel

aantal clusters	samenstelling van de clusters	minimale correlatie tussen cluster elementen
4	$C_1 = \{1, 2, 3, 6, 9, 12, 13\}$ $C_2 = \{5, 7, 8, \}$ $C_3 = \{4, 10\}$ $C_4 = \{11\}$	$r_{6,9} = .154$
3	$C_1 = \{1, 2, 3, 6, 9, 12, 13\}$ $C_2 = \{5, 7, 8, 11\}$ $C_3 = \{4, 10\}$	$r_{5,11} = -.015$
2	$C_1 = \{1, 2, 3, 6, 9, 12, 13\}$ $C_2 = \{4, 5, 7, 8, 10, 11\}$	$r_{8,10} = -.109$
1	$C_1 = \{1, 2, \dots, 13\}$	$r_{5,9} = -.941$

Tabel 6. Cluster analyse voor 13 variabelen (cardinale data).

blijkt dat de overgang van 2 naar 1 cluster gepaard gaat met een bijzonder sterke terugval van de homogeniteit. Dit is in overeenstemming met de resultaten van par. 6.1.

6.4.b Cluster Analyse: Ordinaal

In Tabel 7 zijn de resultaten te vinden van een cluster analyse toegepast op de rangcorrelatiecoëfficiënten van Tabel 2. De resultaten van Tabel 6 en 7 hebben duidelijk een sterke overeenkomst.

aantal clusters	samenstelling van de clusters	minimale correlatie tussen cluster elementen
4	$C_1 = \{1, 2, 6, 9, 12, 13\}$ $C_2 = \{4, 5, 7, 8\}$ $C_3 = \{3\}$ $C_4 = \{10, 11\}$	$r_{4,7} = .059$
3	$C_1 = \{1, 2, 6, 9, 12, 13\}$ $C_2 = \{4, 5, 7, 8\}$ $C_3 = \{3, 10, 11\}$	$r_{3,10} = .028$
2	$C_1 = \{1, 2, 3, 6, 9, 12, 13\}$ $C_2 = \{4, 5, 7, 8, 10, 11\}$	$r_{4,11} = -.007$
1	$C_1 = \{1, 2, \dots, 13\}$	$r_{5,9} = -.737$

Tabel 7. Cluster analyse voor 13 variabelen (ordinaire data).

6.5.a Interregionale Migratie en Interregionale Welvaartsverschillen: Cardinale Analyse

In deze paragraaf zal de informatie over de regionale welvaartsindicatoren gebruikt worden om interregionale migratiestromen te verklaren. We gaan daarbij uit van de volgende specificatie:

$$\frac{M^{sr} - M^{rs}}{p^r \cdot p^s} = (d^{rs})^{-\alpha_0} \{ \alpha_1 (v_1^r - v_1^s) + \dots + \alpha_4 (v_4^r - v_4^s) \} \quad (8)$$

waarbij M^{rs} : aantal migranten per jaar van r naar s

p^r : bevolkingsomvang van regio r

d^{rs} : afstand tussen r en s

v_k^r : kenmerk van regio r ($k = 1, \dots, 4$)

In deze vergelijking wordt het migratiesaldo tussen 2 regio's (na standaardisatie) verklaard door de afstand tussen beide regio's en het verschil tussen beide regio's voor een viertal welvaartsindicatoren. De eerste drie welvaartsindicatoren zijn verkregen als ongewogen gemiddelden van de genormaliseerde variabelen binnen de drie sub-profielen: het sociaal economisch-profiel, het milieu-profiel en het infrastructuur-profiel. Als vierde variabele is een indicator toegevoegd voor de beschikbaarheid van woningen. Deze variabele is gedefiniëerd als het verschil tussen de groeivoet van de

woningvoorraad en de groeivoet van de bevolking (afgezien van migratie). Voor een nadere bespreking van vergelijking (8) verwijzen we naar Suyker, 1982.

Bij de schatting van (8) hebben we de beschikking over 780 (1.40.39) waarnemingen. We schatten ($\alpha_1, \dots, \alpha_4$) gegeven diverse gepredermineerde waarden van α_0 . De beste resultaten worden bereikt voor $\alpha_0 = 1.3$ (zie Tabel 8).

α_0	α_1	α_2	α_3	α_4	R^2
1.3	-1.25	3.42	-.68	3.69	.296
	(-1.47)	(2.66)	(-.76)	(15.08)	

Tabel 8. Schattingsresultaten van de migratie vergelijking
(t-waarden staan tussen haakjes)

De geschatte waarden van α_2 en α_4 (milieu en woningsituatie) hebben het juiste teken en zijn significant verschillend van nul. De geschatte waarden van α_1 en α_3 (sociaal-economische condities en infrastructuur) zijn negatief hoewel niet in significante mate. Dit betekent dat interregionale migratie in Nederland vooral verklaard wordt door interregionale verschillen in de milieu kwaliteit en de situatie op de woningmarkt. De negatieve waarden voor de coëfficiënten α_1 en α_3 zijn onbevredigend, hoewel niet geheel onverwacht. Het migratiepatroon in Nederland wordt gekarakteriseerd door een sterke uitstroom uit de geurbaniseerde regio's naar de rest van het land. De verstedelijkte regio's hebben in het algemeen een gunstige sociaal-economische en infrastructurele positie. Negatieve tekens voor α_1 en α_3 kunnen dan ook aanduiden dat v_1 en v_3 als proxy variabelen dienen van de negatieve aspecten van geurbaniseerde regio's. Dit zou betekenen dat v_2 en v_4 deze negatieve aspecten niet volledig representeren. Speciaal voor v_4 is dit niet verrassend, aangezien deze variabele (t.g.v. data beperkingen) op een tamelijk primitieve wijze is gedefiniëerd.

6.5.b Interregionale Migratie en Interregionale Welvaartsverschillen:

Ordinale Data

Het verband tussen interregionale welvaartsverschillen en interregionale migratie bij ordinale data is als volgt geschat. Allereerst is voor ieder van de drie subprofielen een ordinale indicator v_k geconstrueerd ($k = 1, 2, 3$) zoals aangegeven in par. 4.2. Vervolgens is transformatie (3) toegepast op de 4 indicatoren ($v_1 \dots, v_4$), waarna de frequenties $F_1, F_2, \dots$ van de diverse regimes kunnen worden vastgesteld. Voor ieder paar regio's dat valt onder regime m is daarop nagegaan of de afhankelijke variabele - het migratiesaldo - groter dan, gelijk aan, of kleiner dan nul is. Langs deze weg zijn voor ieder regime m de frequenties F_m, F_{m0} en F_{m-1} vastgesteld.

In Tabel 9 zijn de verschillende regimes m , de bij behorende frequenties F_m, F_{m1}, F_{m-1} , alsmede de waarden van f_{m1} en $\ln(f_{m1}/f_{m-1})$ weergegeven.

regime	regime structuur	F	F_1	F_{-1}	F_1/F	$\ln \frac{F_1}{F_{-1}}$
1	-1 -1 -1 -1	46	16	29	.35	-.59
2	1 -1 -1 -1	53	7	45	.13	-1.86
3	-1 1 -1 -1	140	106	33	.76	1.17
4	1 1 -1 -1	87	45	42	.52	.07
5	-1 -1 1 -1	90	30	59	.33	-.68
6	1 -1 1 -1	258	38	217	.15	-1.74
7	-1 1 1 -1	56	33	20	.59	.50
8	1 1 1 -1	49	22	27	.45	-.20
9	-1 -1 -1 1	49	27	22	.55	.20
10	1 -1 -1 1	56	20	33	.36	-.50
11	-1 1 -1 1	258	217	38	.84	1.74
12	1 1 -1 1	90	59	30	.66	.68
13	-1 -1 1 1	87	42	45	.48	-.07
14	1 -1 1 1	140	33	106	.24	-1.17
15	-1 1 1 1	53	45	7	.85	1.86
16	1 1 1 1	46	29	16	.63	.59

Tabel 9. Gegevens voor de regime-methode voor ordinale data.

Merk op dat Tabel 9 geen regimes bevat met de waarde nul voor één van de z_{mj} . In werkelijkheid is er één paar regio's waarvoor zich een dergelijk geval voor-
doet. Tabel 9 heeft dan ook betrekking op alle paren regio's (780) op één na.

Voorts noteren wij dat in de meeste gevallen F_{m0} klein is. Dit wijst erop dat het migratiesaldo tussen 2 regio's slechts in uitzonderingsgevallen exact gelijk aan nul is, wat overigens geen verbazing wekt.

Deze toepassing van de regime method wijkt enigszins af van de beschrijving in par. 5. Immers, in par. 5 wordt uitgegaan van R waarnemingen van een ordinaal gemeten afhankelijke variabele die middels (3) worden getransformeerd in $\frac{1}{2} R(R-1)$ waarnemingsparen die de waarden 1, 0 of -1 aannemen. In het geval van de migratiesaldi wordt onmiddellijk uitgegaan van $\frac{1}{2} R(R-1)$ min of meer onafhankelijke interregionele saldi die de waarden 1, 0 of -1 aannemen.

Een gewogen kleinste kwadraten schatting van (7) gebaseerd op Tabel 8 leidt tot de volgende resultaten (zie Tabel 10).

β_1	β_2	β_3	β_4	R^2
-0.51	.78	-.14	.36	.986

Tabel 10. Schattingsresultaten voor de migratievergelijking met ordinale data.

De tekens van de coëfficiënten komen overeen met die uit Tabel 8. De coëfficiënt voor de infrastructuur variabele is in beide gevallen duidelijk kleiner (in absolute zin) dan die van de andere variabelen.

7. Conclusie

De overeenkomst tussen de uitkomsten met ordinale en cardinale data is in een aantal gevallen treffend. De overeenkomst is duidelijk het minst bij de interdependentie analyse. In geen enkel geval is er sprake van duidelijke contradicties tussen een cardinale en ordinale analyse. We concluderen dan ook dat op grond van ordinale data wel dergelijk zinvolle statistische en econometrische analyses kunnen worden uitgevoerd.

Referenties

- Adelman, I., en C.T. Morris, The Deviation of Cardinal Scales from Ordinal Data, Economic Development and Planning, (W. Sellekaerts, Ed.), MacMillan, London, 1974, pp. 1-39.
- Bishop, Y.M.M., S.E. Fienberg en P.W. Holland, Discrete Multivariate Analysis, MIT Press, Cambridge, Mass., 1975.
- Blalock, H.M., Can We Find a Genuine Ordinal Scope Analogue?, Sociological Methodology (D.R. Heise, Ed.), Jossey-Bass Publ., San Francisco, 1976, pp. 195-226.
- Blommestein, H.J., Nijkamp, P. and P. Rietveld, A Multivariate Analysis of Spatial Inequalities, in (W. Buhr en P. Friedrich, eds.), Regional Development Under Stagnation, Nomos, Baden-Baden, 1981, pp. 293-318.
- Boyce, D.E., A. Fahri en R. Weischedel, Optimal Subset Selection, Springer, Berlin, 1974.
- McCullagh, P., Regression Models for Ordinal Data, Journal of the Royal Statistical Society, vol. 42, no. 2, 1980, pp. 109-142.
- Domencich, F.A., en D. McFadden, Urban Travel Demand, a Behavioral Analysis, North-Holland Publ. Co., Amsterdam, 1975.
- Grizzle, J.E., C.F. Starmer en G.G. Koch, Analysis of Categorical Data by Linear Models, Biometrics, vol. 25, 1969, pp. 489-504.
- Hartigan, J.A., Clustering Algorithms, Wiley, New York, 1975.
- Kendall, M.G., Rank Correlation Methods, Griffin, London, 1970.
- Nijkamp, P. en P. Rietveld, Ordinal Multivariate Analysis, Professional Paper PP-81-2, IIASA, Laxenburg, 1981.
- Nijkamp, P., Multidimensional Spatial Data and Decision Analysis, Wiley, Chichester/New York, 1979.
- Siegel, S., Nonparametric Statistics for the Behavioral Sciences, McGraw-Hill, New York, 1956.
- Suyker, W.B.C., An Analysis of Interprovincial Migration in the Netherlands, Papers of the Regional Science Association, 1982 (te verschijnen).
- Theil, H., Principles of Econometrics, Wiley, New York, 1971.
- Veenendaal, W.M. van, Regionale Welvaart in Nederland, Economische Faculteit, Vrije Universiteit, Amsterdam, 1981.
- Voogd, J.H., On the Principles of Ordinal Geometric Scaling, Research Paper no. 10, Planologisch Studiecentrum, PSC-TNO, Delft, 1978.
- Wrigley, N., Development in the Statistical Analysis of Categorical Data, Progress in Human Geography, vol. 3, 1979, pp. 315-355.