ECONOMETRISCHE ANALYSE VAN GROTE BEDRIJFSECONOMISCHE GEGEVENSBESTANDEN

door

Philip Hans FRANSES

Econometrisch Instituut en
Rotterdams Instituut voor Bedrijfseconomische Studies,
Erasmus Universiteit Rotterdam

SAMENVATTING

In de financiering en de marketing zijn steeds meer gegevens beschikbaar op het individuele transactieniveau. De econometrische analyse van deze grote gegevensbestanden blijkt echter niet rechttoe, rechtaan. In dit artikel worden enkele kanttekeningen geplaatst bij deze analyse, waarbij de nadruk ligt op de mogelijke toepassing van complexe niet–lineaire modellen, data mining, de informatiewaarde van de gegevens, en aspecten als het schatten van parameters, modelkeuze en modelevaluatie. Een conclusie is dat voor het vakgebied econometrie, dat vooral tot bloei kwam dankzij toepassing bij macroeconomische vraagstukken, veel toekomstige ontwikkelingen te verwachten zijn, met name door haar toenemende toepassing bij de analyse van grote bedrijfseconomische gegevensbestanden.

Deze versie: 4 juni 1997. Commentaar welkom.

NOTEN

Dit artikel is geschreven voor presentatie tijdens het Achtste Symposium Statistische Software 1997 ("Verborgen Rijkdom"), 12 november 1997 (VVS en CBS), en de RECNET dag, 20 juni 1997 (ROBECO). Ik bedank Dick van Dijk voor zijn commentaar. Aan dit artikel werd begonnen in de vroege ochtend van 20 mei 1997, de dag die later de geboortedag van mijn zoon Tobias zou blijken.

Correspondentie: Econometrisch Instituut, Postbus 1738, 3000 DR Rotterdam, telefoon: 010 4081273, fax: 010 4527746, e–mail: franses@few.eur.nl

1. INLEIDING

Het vakgebied econometrie heeft zich sinds de jaren '50 enorm ontwikkeld, mede dankzij de grote belangstelling die er bestond voor de toepassing bij de analyse van macroeconomische vraagstukken. Grootschalige macro-modellen die complete nationale economiën in kaart brachten (en nog immer brengen) zijn daar de meest bekende exponenten van. Omdat het niet eenvoudig is gegevens te verzamelen die een nationale economie als geheel karakteriseren, werden en worden die grote modellen vaak gebruikt voor jaarlijkse (of hooguit kwartaal-) gegegens, waarbij geaggregeerd is over alle individuen in een economie of een sector. Gezien het feit dat nationale statistische bureaus vaak pas sinds de jaren 70 met enige precisie relevante economische grootheden als werkloosheid en nationaal produkt kunnen meten, worden veel macroeconometrische modellen beschouwd voor vaak niet meer dan 30 jaar aan waarnemingen.

De schaarste aan macroeconomische data heeft geleid tot grote aandacht in de econometrie voor de analyse van simultane modellen en voor de studie naar de kleine steekproefeigenschappen van schatters en toetsgrootheden, om maar enkele voorbeelden te noemen. Immers, als men maar beschikt over slechts 30 jaarcijfers, wat heeft men dan aan een toets op bijvoorbeeld seriecorrelatie in de residuen, als die pas bij meer dan duizend waarnemingen een groot onderscheidend vermogen heeft? De aandacht voor simultane modellen is vooral ingegeven door de mogelijkheid dat het werkelijke economische proces zich binnen een jaar afspeelt, terwijl men toch alleen jaarcijfers heeft voor de modelbouw. Immers, in dat geval lijken relaties tussen economische grootheden tegelijkertijd en niet na elkaar plaats te vinden.

Een ander toepassingsgebied van de econometrie dat heeft geleid tot veel wetenschappelijke progressie is de arbeidsmarkteconomie. Bijvoorbeeld, enkele duizenden personen worden gevolgd gedurende een aantal jaren, waarbij ze eens in de vijf jaar worden gevraagd naar hun mate van participatie op de arbeidsmarkt. Dit levert een zogenaamd panel van gegevens op, dat is, men beschikt uiteindelijk over waarnemingen voor, zeg, $T = 3$ momenten in de tijd en voor, zeg, $N = 2000$ individuen. Deze panels zijn erg kostbaar om samen te stellen, en omdat de tijdscomponent T vaak niet al te groot is, zijn ook hier de kleine steekproefeigenschappen van schatters en toetsgrootheden van belang in het geval men dynamische modellen beschouwt.

In dit artikel zal ik de aandacht richten op de mogelijke toepassing van econometrische methoden en technieken op gegevensbestanden die vele malen

groter zijn dan die bij de macro– en arbeidsmarkteconomie. Door toegenomen meet– en opslagfaciliteiten kan men in de bedrijfseconomische deelgebieden marketing en financiering vaak beschikken over bestanden met vele miljoenen waarnemingen. In veel supermarkten worden sinds een aantal jaren alle boodschappen met een zogenaamde scanner–apparaat geregistreerd, resulterend in de bekende scanning data. In de financiële economie is het de gewoonte om per minuut of soms zelfs per seconde de prijs van aandelen en de hoogte van de rente bij te houden. De bedrijfseconomische gegevens betreffen vaak de registratie van een specifieke transactie. Een produkt wordt in een winkel gekocht, en een aandeel gaat voor een zekere prijs van hand tot hand. Aangezien de transactie een elementaire economische handeling is, lijkt de weg vrij te zijn om middels de toepassing van econometrische technieken daadwerkelijk inzicht te krijgen in het economische gedrag van individuen. Echter, in dit artikel zal ik aangeven dat de bestaande technieken niet zonder meer kunnen worden aangewend voor deze doelstelling. Er ligt dan ook een grote uitdaging om de huidige methoden te verbeteren of aan te passen, en om nieuwe econometrische methoden te ontwikkelen, zodanig dat die speciaal geschikt zijn voor de analyse van grote gegevensbestanden. Het is mijn verwachting dat de toepassing op deze gegevens zal leiden tot nieuwe wetenschappelijke ontwikkelingen in het vakgebied econometrie.

Dit artikel is als volgt ingedeeld. In paragraaf 2 geef ik een indruk van de verschijningsvorm van zulke bedrijfseconomische databestanden. Daarna, in paragraaf 3, geef ik enkele kanttekeningen bij de econometrische modellering van die gegevens. Aan de orde komen de toepassing van complexe modellen, de informatiewaarde van de data, het zoeken naar structuur en het schatten van de parameters en modelkeuze. In paragraaf 4 besluit ik dit artikel met enkele laatste opmerkingen.

Tenslotte zij het vermeld dat dit artikel bedoeld is als een wat informele persoonlijke visie op sommige ontwikkelingen. Ik streef er dan ook niet naar een alomvattend overzicht te geven van het totale econometrische vakgebied. Het afsluitende literatuuroverzicht is ook verre van compleet. Als de geachtevorming over de econometrische analyse van grote bedrijfseconomische gegevensbestanden door dit artikel mede gestimuleerd wordt, is mijn doel al bereikt.

2. GROTE GEGEVENSBESTANDEN

In deze paragraaf bespreek ik een paar typische aspecten van de structuur van grote bedrijfseconomische gegevensbestanden, wederom zonder er naar te streven volledig te zijn. Ik beperk mij tot de panelstructuur, die een uitbreiding vormt op de eerder genoemde structuur bij de gegevens van aspecten van de arbeidsmarkt. Noteer y als de te verklaren variabele, en X als een $(p\times1)$ vector van exogene waargenomen variabelen, dat is $X = (x_1, x_2, .., x_p)$, waarbij x_1 een constante weergeeft. Verder, noteer ε als de storingsterm en θ als een $(p\times1)$ vector van parameters. Tenslotte, f noteert de functie die X en θ relateert aan y. Om een discussie over de structuur van een groot gegevensbestand te stroomlijnen kies ik voor de volgende modelvorm:

$$y_{i,j,k,t} = f_{i,j,k,t}(X_{i,j,k,t}, \theta_{i,j,k,t}) + \varepsilon_{i,j,k,t}, \qquad (1)$$

waarbij de indices i, j, k en t bijvoorbeeld kunnen staan voor

Individuen	$i = 1, 2, ..., I,$
Omgeving:	$j = 1, 2, ..., J,$
Type transactie	$k = 1, 2, ..., K,$
Tijd:	$t = 1, 2, ..., T.$

Als $I = J = K = T = 100$ en $p = 10$, dan heeft men $I\times J\times K\times K\times(p+1) = 1{,}100{,}000{,}000$ waarnemingen voor de verklarende en te verklaren variabelen. Hoewel model (1) natuurlijk kan worden uitgebreid met nog een extra dimensie, dat is, nog een extra index bij de variabelen en parameters, voldoet het aan zijn doeleinden voor dit moment. Hieronder geef ik voor de financiering en de marketing enkele voorbeelden van zo'n groot gegevensbestand. Overigens zij opgemerkt dat ik veronderstel dat men in staat is om al die gegevens in ieder geval tijdelijk op te slaan.

2.1 Financiering

Op veel financiële markten is het mogelijk om elke transactie vast te leggen. Van zo'n transactie kan men rapporteren of er eerst een offerte of eerst een bod is gedaan. Verder kan de transactie een zeker aandeel betreffen dat op een aandelenmarkt in een zeker land wordt verhandeld. Omdat het bekend is dat het

gedrag op financiële markten verandering ondergaat naarmate de dag vordert is het van belang niet een dwarsdoorsnede van transacties maar het gedrag door de tijd heen te analyseren. In termen van (1) levert dit op

I: transacties in aandeel i
J: beurs in land j
K: transactie van het type k
T: tijdstip van de dag t.

Zou men de wens hebben op dit gedetailleerde niveau het gedrag op financiële markten te beschrijven dan zou men al gauw $I = 25$ aandelen, $J = 5$ beurzen, met K is bijvoorbeeld 2 en $T = 8$ maal $60 = 480$ minuten kunnen bestuderen. Dit geeft in totaal 120,000 waarnemingen. Omdat het tevens bekend is dat het gedrag op een aandelenbeurs per weekdag kan verschillen, zie Andersen en Bollerslev (1995), lijkt het zinnig bijvoorbeeld een jaar aan werkdagen te bestuderen. Als de transacties over zeg 250 werkdagen per jaar worden geobserveerd, dan zou men kunnen beschikken over 30 miljoen waarnemingen per variabele. Zelfs als $K = 1$, dan zijn dit nog steeds 15 miljoen waarnemingen. Wanneer y het uiteindelijke rendement van het aandeel is, dan zou X allerlei eigenschappen van het aandeel kunnen betreffen.

Een model voor y zou kunnen trachten te achterhalen wat de achterliggende verklarende variabelen zijn die het rendement van een aandeel verklaren. Men kan ook overwegen om met y het aantal verhandelde aandelen per transactie te beschrijven. Het is bekend dat het rendement nagenoeg onvoorspelbaar lijkt, terwijl over het aantal aandelen dat van hand tot hand gaat mogelijkerwijs wel een uitspraak valt te doen. Gezien de dynamiek op de meeste aandelenmarkten lijkt het verstandig om de waarde van y in de vorige periode als verklarende variabele mee te nemen. Ook lijkt het zinvol om een maatstaf voor de onrust op de financiële markten op te nemen in het model. Omdat die maatstaf vaak weer een functie is van rendement of aantal transacties in het recente verleden, kan men zich voorstellen dat een model dat alle aspecten goed wil beschrijven al gauw nogal complex wordt.

2.2 Marketing

De introductie van de scanning techniek (streepjescode) heeft het mogelijk gemaakt om in supermarkten en nu ook in andere winkels elke transactie bij te

houden. Initieel was deze techniek vooral nuttig voor voorraadbeheer, maar nu levert het ook allerlei grote gegevensbestanden op die met econometrische methoden kunnen worden geanalyseerd. Doel van die analyse is vaak een studie naar de effecten van allerlei marketinginstrumenten zoals reclame, promoties, distributie en van de positionering van het schap in de winkel op de verkopen of het marktaandeel. Als men individuele consumenten thuis de gelegenheid geeft middels een scanningapparaat bij te houden welk produkt van welk merk ze waar hebben gekocht kan men al gauw een groot gegevensbestand opbouwen. In dat geval kan men transacties observeren voor

I: consument i, in
J: winkel j, die
K: produkt of merk k koopt, in
T: week t.

Het aantal consumenten I in zo'n steekproef is al gauw 1000, het aantal winkels 100, en vaak beschikt men over minstens $T = 100$ wekelijkse gegevens. Meestal bestudeert men de transacties voor een beperkt aantal merken, zeg $K = 5$. In totaal levert dit 50 miljoen waarnemingen op voor variabelen zoals verkopen of marktaandeel, distributiegraad, prijs, reclame en eventuele promoties. Naast de verkopen in het totaal, kan men ook verkiezen de variatie in het gedrag over verschillende winkeltypen of verschillende typen consumenten bestuderen. Een voorbeeld van een gelijksoortig gegevensbestand betreft alle credit card of PIN transacties gedurende een bepaalde periode.

Een andere categorie van grote gegevensbestanden met ook weer speficieke eigenschappen betreft de direct marketing. Bijvoorbeeld, veel (non–)profit bedrijven onderkennen het belang van het onderhouden van goede relaties met hun huidige klanten en bieden daarom regelmatig nieuwe produkten via een brief aan. De klant kan dan een bedrag inleggen, of besluiten niet te reageren. Een karakteristieke situatie is dat men 100,000 individuen een brief stuurt, en dat daarvan 5000 klanten responderen door een bedrag in te leggen. Omdat de eigenschappen van de klanten bekend zijn door voorgaande acties, kan men een inzicht verkrijgen in welke factoren de mate van responderen bepaalt en welke variabelen de hoogte van het ingelegde bedrag verklaren. Een overzicht van enkele nuttige econometrische modellen voor het modelleren van direct mailing respons wordt gegeven in Franses (1997).

3. ECONOMETRISCHE ANALYSE

In deze paragraaf plaats ik kanttekeningen bij de econometrische analyse van grote gegevensbestanden. In sommige opzichten vullen zij de recente opsomming in Granger (1997) aan. Evenals in dat artikel veronderstel ik dat er voorlopig nog geen bovengrens zit aan de rekencapaciteit van computers. Beter nog, ik ga er van uit dat de mogelijkheden die grote databestanden bieden juist een motivatie zijn om de rekencapaciteit op te voeren. Immers, de IBM machine "Deep Blue", die gebruikt is als schaaktegenstander van Kasparov, is er voor gemaakt om "..(te) zoeken naar verborgen relaties tussen grote hoeveelheden, op het oog onafhankelijke gegevens." (bron: NRC, Zaterdag 10 mei 1997). Verder veronderstel ik, vooral om de notatie eenvoudig te houden, dat alle variabelen met dezelfde frequentie gemeten zijn. Het onderstaande wordt ingewikkelder wanneer van sommige variabelen bijvoorbeeld metingen per seconde zijn, terwijl van andere relevante variabelen er slechts minuutcijfers bekend zijn. Hoe men zulke variabelen samenbrengt in een enkel econometrisch model is een terrein van onderzoek op zich.

Natuurlijk zijn er vele kanttekeningen te plaatsen, en ik laat het aan de lezer over om aanvullingen op de onderstaande opmerkingen te bedenken. Tevens zij gemeld dat dit artikel er vooral toe dient aan te geven waar eventuele knelpunten zitten, en niet om ook meteen de oplossingen aan te dragen. Het is vooral de intentie te laten zien dat nog heel veel theoretisch en toegepast econometrisch onderzoek nodig is om grote databestanden nuttig te analyseren.

3.1 Toepassing van complexe modellen

De meeste econometrische modellen die in de bedrijfseconomische praktijk worden toegepast zijn linear in de parameters. Dit betekent dat model (1) dan reduceert tot

$$y_{i,j,k,t} = X_{i,j,k,t}{}'\theta_{i,j,k,t} + \varepsilon_{i,j,k,t}. \qquad (2)$$

Met specifieke restricties op de θ parameters (bijvoorbeeld dat zij voor alle i gelijk zijn, zie subparagraaf 3.2), kunnen zij met gewoon kleinste kwadraten [GKK] methodes worden geschat door alle waarnemingen te stapelen in ($IJKT$×1) vectoren en de bekende GKK formules te hanteren. Hoewel een lineair model heel nuttig is om verbanden te beschrijven tussen variabelen, lijkt een lineair

model eerder geschikt voor een klein aantal waarnemingen dan voor een verzameling met miljoenen waarnemingen. Immers, de kans dat waarnemingen afwijken van de ideale lineaire regressielijn (zie wederom subparagraaf 3.2) wordt snel groter naarmate er meer waarnemingen zijn.

Het is dan ook nuttig om voor grote gegevensbestanden meer flexibele modelvormen te beschouwen. Vaak geldt ook dat zulke modellen juist beter toepasbaar zijn in het geval men veel waarnemingen heeft. Dit wordt veroorzaakt door het feit dat flexibele modellen kunnen worden gekenmerkt door veel parameters, die met een kleine verandering vergelijkbare patronen in de data kunnen beschrijven. Het is dan dus nuttig als die parameters met grote precisie kunnen worden bepaald. Een voorbeeld van een flexibel model is het artificiële neurale netwerk model [ANN], zie Kuan en White (1994) en Ripley (1994) voor recente overzichten. In het geval van één verklarende variabele, ziet een ANN met 1 verborgen laag er uit als

$$y_n = \mu + \theta x_n + \sum_{q=1}^{Q} \beta_q G(\mu_q + \theta_q x_n) + \varepsilon_n, \tag{3}$$

met $n = 1,2,..,\ N$, waarbij $N = IJKT$, en $G(.)$ is een activeringsfunctie. Vaak wordt die laatste functie gekozen als de logistische functie

$$G(a) = [1 + \exp(-a)]^{-1}. \tag{4}$$

De verborgen laag in de ANN in (3) heeft Q elementen. Kuan en White (1994) (en anderen) laten zien dat, wanneer Q maar groot genoeg wordt, model (3) kan worden gebruikt om elke functie f in $y_n = f(x_n)$ arbitrair dicht te benaderen. Door te onderzoeken of de functie $\sum_{q=1}^{Q} \beta_q G(\mu_q + \theta_q x_n)$ een bepaald herkenbaar patroon heeft voor sommige i, j, k of t zou men inzicht kunnen proberen te verkrijgen in een eventuele structuur in de gegevens. Franses en Draisma (1997a) gebruiken op deze wijze een ANN om te onderzoeken of (overigens macro-)economische variabelen in sommige seizoenen een ander gedrag vertonen. Franses en Draisma (1997b) hanteren vergelijkbare modellen om tijdvariërende elasticiteiten van prijs en distributie op het marktaandeel weer te geven.

Model (3) kan ook worden geschreven als

$$y_n = \mu_n + \theta x_n + \varepsilon_n, \tag{5}$$

ofwel, het ANN is een zogenaamd variërend parameter model waarbij de constante

term verschillende waarden aan kan nemen. Alternatieve specificaties van zulke panelmodellen met variërende constanten worden besproken in subparagraaf 3.2. Het kan zijn dat men ook de parameter θ in (5) wil laten variëren met n. In dat geval kan men (3) uitbreiden tot

$$y_n = \mu + \theta x_n + \sum_{q=1}^{Q} \beta_q G(\mu_q + \theta_q x_n) + \sum_{p=1}^{P} \alpha_p x_n G(\mu_p + \theta_p x_n) + \varepsilon_n. \qquad (6)$$

Het is duidelijk dat model (6) $2+3Q+3P$ parameters heeft terwijl er maar 1 verklarende variabele is. Om de parameters in model (6) redelijk precies te kunnen schatten zijn er derhalve veel waarnemingen nodig.

Een model dat lijkt op een ANN, en dat ook nuttig kan zijn om structuren in de data te ontwaren, is het zogenaamde gegeneralizeerde additieve model, zie bijvoorbeeld Hastie en Tibshirani (1990). Dit is een niet–parametrisch model omdat de data zelf worden gebruikt om het functionele verband tussen de verklarende en de te verklaren variabele te bepalen. Een voorbeeld in geval van twee verklarende variabelen is

$$y_n = \mu + f_1(x_{1,n}) + f_2(x_{2,n}) + \varepsilon_n, \qquad (7)$$

waar f_1 en f_2 door de data te specificeren functies zijn. Deze twee functies kunnen erg flexibel zijn, en daarmee kunnen complexe verbanden tussen x_1, x_2 en y worden beschreven. Omdat de vorm van f_1 en f_2 door de data wordt bepaald, heeft men veel waarnemingen nodig om een accurate schatting van ze te maken. Natuurlijk zijn er allerlei andere soorten niet–parametrische modellen mogelijk, zie bijvoorbeeld het overzicht in Fan en Gijbels (1996). Een nuttig voordeel van (7) is dat de partiële effecten van x_1 en x_2 op y uit elkaar kunnen worden gehouden.

Om structuur te ontdekken zullen grote gegevensbestanden met nam worden geanalyseerd middels classificatiemethoden. Voorbeelden zijn discriminantanalyse, clusteranalyse, en eventueel het veelgebruikte logitmodel. Als de grens tussen twee of meer verzamelingen van waarnemingen moet worden bepaald op basis van de data, dan is het nuttig als er heel veel waarnemingen zijn. Sequentiële toepassing van die methoden kan dan leiden tot het opdelen van de gegevensverzameling in min of meer homogene groepen, waarbinnen weer bepaalde verbanden tussen de endogene en verklarende variabelen gelden. Met een schier oneindige hoeveelheid data moet het in principe mogelijk zijn om die opdeling vrij precies te bepalen. Wanneer

segmenten kunnen worden ontdekt in een verzameling consumenten die in bepaalde winkels de boodschappen doet, of die op sommige vormen van direct mailing reageert, dan kan men doelgericht marketinginstrumenten inzetten om de verkopen trachten te bevorderen.

In het algemeen kan men concluderen dat grote hoeveelheden gegevens het mogelijk maken meer complexe, maar potentieel zeer nuttige, modellen te beschouwen. Deze modellen kunnen dan ook worden gebruikt om al zoekende structuur in de gegevens te ontdekken, zie ook subparagraaf 3.3.

3.2 Informatiewaarde van de data

Een voordeel van een kleine gegevensverzameling is dat de meeste waarnemingen van belang zijn. Bij de analyse van 30 jaarcijfers van het Bruto Nationaal Produkt [BNP], lopend van 1966 tot en met 1995 kan men begrijpen dat het weglaten van twee waarnemingen, zeg 1974 en 1975, van invloed zal zijn op de analyse. Verder kan het weglaten van 1994 en 1995 een groot effect hebben op voorspellingen omdat een mogelijk recente wending van de trend niet wordt meegenomen. Hoewel ik niet er naar streef het begrip formeel te definiëren, kan men stellen dat de 30 jaarcijfers allen veel informatiewaarde hebben. Het weglaten van enkele waarnemingen kan grote gevolgen hebben voor inferentie.

Bij een kleine dataset hebben individuele waarnemingen dan ook een groot effect op de resultaten van een econometrische analyse. Dit betekent dat een afwijkende waarneming een flink verstorende werking kan hebben, niet alleen op het regressiemodel zelf, maar ook op de eigenschappen van de geschatte residuen. Dit heeft geleid tot veel onderzoek naar econometrische methoden om afwijkende waarnemingen op te sporen. Deze data worden dan op een of andere manier gewogen, en in een volgende ronde van parameters schatten worden de gewogen waarnemingen beschouwd. Daarna wordt het model weer onderzocht op de verstorende werking van enkele data, enzovoort totdat een model resulteert dat in zekere zin optimaal is, zie Belsley, Kuh en Welsh (1980) voor enkele relevante methoden voor cross–sectiegegevens en Tsay (1988) en Chen en Liu (1993) voor tijdreeksgegevens.

Bij een grote gegevensverzameling met miljoenen gegevens lijkt het onwaarschijnlijk dat enkele waarnemingen dermate afwijkend zijn dat ze het uiteindelijke model kunnen verpesten. Als afwijkende waarnemingen erg invloedrijk zijn, dan moet het om grote hoeveelheden gaan. Bijvoorbeeld, een beurs in een land j vertoont ander gedrag dan de beurzen in andere landen (is

bijvoorbeeld erg onrustig), of winkel j ligt zodanig ver van de ander $J-1$ winkels dat men kan stellen dat de consumenten die winkel niet echt als één van de mogelijk winkels zien om de boodschappen te doen. In deze gevallen kan men overwegen een deelverzameling van de waarnemingen niet mee te nemen bij de analyse.

De huidige kleine steekproef–gebaseerde methoden voor het ontdekken van afwijkende waarnemingen moeten worden aangepast om voor grote hoeveelheden gegevens toepasbaar te worden. Het iteratief zoeken naar en aanpassen van slechts 1 of enkele waarnemingen per keer zal veel te tijdrovend en onhandig zijn. Immers, men wil ook voorkomen dat de aangewezen afwijkende observaties kriskras door de data te vinden zijn. Het lijkt eerder nuttig om robuuste schattingsmethoden te ontwerpen die alle gegevens tegelijk beschouwen en die automatisch waarnemingen een kleiner gewicht geven wanneer ze een verstorende werking hebben. In Lucas (1995) worden zulke methoden voorgesteld voor kleine verzamelingen van tijdreeksgegevens, en deze methoden lijken in principe uit te breiden voor grote hoeveelheden data.

Voordat men robuuste technieken kan toepassen is het natuurlijk eerst van belang om te definiëren wat eigenlijk een afwijkende waarneming in een hele grote gegevensverzameling is. Immers, bij zo'n mega–panel (zoals Granger (1997) het noemt) is de individuele informatiewaarde van een enkele waarneming veel minder groot dan in het 30 jaar voorbeeld hierboven. Sterker nog, hebben al die data wel evenveel informatiewaarde? Hoe schadelijk is het om een aantal van die waarnemingen weg te laten? Een concreet voorbeeld betreft de direct marketing respons. Als 5000 personen reageren en 95000 niet, is het de vraag of men al die nonrespondenten moet meenemen in een model. Fragiele verbanden tussen de te verklaren en verklarende variabelen voor de 5000 respondenten kunnen danig worden verstoord door de mate van ruis verzoorzaakt door de toegevoegde groep van nonrespondenten. Men zou kunnen overwegen slechts een deel van die 95000 te beschouwen, eventueel met herhaalde steekproeftrekking.

Als het duidelijk is dat niet alle waarnemingen even relevant zijn, kan men tot aggregatie overgaan. Aggregatie kan plaatsvinden over alle dimensies, maar ook over deelverzamelingen. Bijvoorbeeld, men kan de Aziatische beurzen samen nemen, of alle mannelijke consumenten. Dit geeft dan de gelegenheid in te zoomen op voor de onderzoeker meer relevante aspecten. Over de effecten van aggregatie is al veel nagedacht, maar er lijkt niet een algemeen geldende regel vast te stellen. In sommige gevallen werkt het in het voordeel, in weer andere gevallen vertroebelt het onderliggende structuren. Bij een mega–panel

lijkt de kans op vertroebeling wat kleiner als men niet al te rigoreus de gegevens samenvat, en bijvoorbeeld de aggregatie baseert op de resultaten van de toepassing van classificatiemethoden zoals eerder beschreven.

Het is ook mogelijk om niet de waarnemingen zelf, maar als het ware de parameters te aggregeren. Dit wordt vaak "pooling" genoemd, zie Blattberg en George (1991) en Maddala en anderen (1997) voor een discussie van de huidige stand van zaken met betrekking tot pooling. Een voorbeeld is dat men voor het model

$$y_{i,j,k,t} = X_{i,j,k,t}'\theta_{i,j,k,t} + \varepsilon_{i,j,k,t}, \tag{8}$$

veronderstelt dat

$$\theta_{i,j,k,t} = \theta + \mu_i, \tag{9}$$

met

$$\mu_i \sim N(0,\Sigma_\mu), \tag{10}$$

waarbij Σ_μ een $(p\times p)$ matrix is als de vector X geen constante bevat. Het aantal parameters wordt op deze manier flink teruggebracht, terwijl de parameters toch met dezelfde hoeveelheid waarnemingen worden geschat. Uit de vergelijkingen (9)–(10) valt al af te lezen dat er erg veel varianten mogelijk zijn, hetgeen betekent dat de selectie van de meest optimale mate van pooling onderwerp van veel toekomstig onderzoek zal zijn. Men kan overwegen in dat geval de aanpak in Hausman (1978) aan te passen voor mega–panels.

Grote gegevensbestanden impliceren (mogelijk nieuwe) definities van de informatiewaarde van individuele of verzamelingen van waarnemingen, maar ook van het concept afwijkende waarnemingen. De huidige methoden zijn vooral geschikt voor gegevens op redelijk kleine schaal. Veel observaties leiden ook weer tot meer en nieuwe beslissingsregels, en het is zinvol te onderzoeken hoe de bestaande technieken kunnen worden aangepast.

3.3 Zoeken naar structuur

Een bekend probleem bij de toepassing van de econometrische technieken op werkelijke gegevens is dat zij worden gebruikt zowel om het model te bepalen

als om het te evalueren. Aangezien die constructie vaak tot doel heeft een model te vinden dat aan bepaalde eisen voldoet, kan men vraagtekens zetten bij de bruikbaarheid van de methoden voor evaluatie. In een klassiek artikel van Lovell (1983) wordt dit probleem zeer illustratief duidelijk gemaakt, en sinds dit artikel staat het zoeken en evalueren van een model op basis van dezelfde gegevens bekend als "data mining". Een voorbeeld uit Lovell (1983) betreft een simulatie waarin wordt getoond dat als men sequentieel t-toetsen gebruikt, men met een kans bijna gelijk aan 1 op een (overigens lineair) model uitkomt dat niet het model is waarmee de gegevens zijn gegenereerd.

Data mining is een belangrijk probleem omdat elk model slechts een poging tot een benadering is, en verder niemand vooraf kennis heeft over welk model het beste is. Economische theorie leidt soms tot enigzins behulpzame aanwijzingen, zoals tot de suggestie welke variabelen in een vergelijking ter verklaring van werkloosheid horen, maar in welke vorm, in welk functioneel verband en met hoeveel vertraging, daarover doet de theorie vaak geen uitspraken. Kortom, elk model komt tot stand na een lange of korte zoekprocedure. Er zijn mogelijkheden om enkele beslissingen tijdens de zoekprocedure expliciet te verwerken in de econometrische methode, zie bijvoorbeeld Judge en Bock (1978), maar vaak levert dit zeer gecompliceerde technieken op, die zelden bruikbaar zijn in de praktijk. Ook blijkt het niet mogelijk om al te veel beslissingen mee te nemen, omdat dan al helemaal geen expliciete resultaten meer te behalen zijn.

Ruwweg gesteld kent het data mining probleem een eenvoudige oplossing. Wanneer men een model maakt, en men heeft N waarnemingen, dan kan men N–M van die data nemen om het model te construeren, en M observaties om het (samen met eventueel andere modellen) te evalueren. Men kan het model opnieuw schatten voor die M waarnemingen en de resultaten met die voor N–M vergelijken, maar men kan ook M waarnemingen voorspellen en slechts de voorspellingen voor de modellen vergelijken. De vraag is dan natuurlijk hoe groot M moet zijn, en op welke wijze je de voorspellingen moet evalueren, maar dat is vaak een praktische kwestie. Men kan het belangrijk vinden dat juist de waarnemingen in jaar t of in winkel j het best worden voorspeld. Immers, het kan bijvoorbeeld niet zo interessant zijn om een rustig economisch jaar als 1986 te voorspellen, terwijl de waarnemingen rondom Augustus 1990 (de inval van Irak in Koeweit) wel weer van belang kunnen zijn voor bijvoorbeeld financiële markten en industriën die olie als grondstof behoeven. Als een model tegen grote schokken kan, dan heeft men er meer vertrouwen in.

De reden dat het fenomeen data mining zoveel aandacht heeft gekregen in de econometrische literatuur (zie bijvoorbeeld de literatuur over "bootstrap", samengevat in Hall (1994), en "pretesting" in Judge en Bock (1980)) komt vooral door de beschikbaarheid van vaak maar weinig macroeconomische gegevens. Het wordt dan niet eenvoudig M te bepalen, maar ook kan N–M te klein worden om nog een zinnig model te maken. Echter, in het geval dat men over miljoenen gegevens beschikt, zoals in marketing en de financiering, dan wordt het probleem van data mining veel minder belangrijk. Sterker nog, de term "data mining" duikt steeds vaker op in juist positieve zin, dat wil zeggen in de betekenis van het bewust zoeken naar structuur in een grote hoeveelheid data. Een economische theorie zal wel iets algemeens kunnen adviseren, maar het lijkt onwaarschijnlijk dat men daarmee kan volstaan in de confrontatie met een groot gegevensbestand. Dat is op zich geen probleem. Immers, bij wetenschappelijk onderzoek op het terrein van de marketing heeft men de neiging "de data te laten vertellen wat de structuur is". Die structuur wordt dan meegenomen in de daaropvolgende theorievorming, zie bijvoorbeeld het themanummer van *Marketing Science* in 1995 met als onderwerp "Empirical generalizations in marketing".

Aangezien bij grote bedrijfseconomische gegevensbestanden vaak nog geen structuur voorhanden is, zullen toekomstige econometrische methoden vooral aandacht besteden aan de verschillende stappen in de zoekprocedure. Daarbij zal dan niet zozeer de precieze verdeling van elke toetsgrootheid van belang zijn, alswel de wijze waarop achtereenvolgende stappen moeten worden genomen. Het achterhouden van een grote hoeveelheid waarnemingen kan dan een eerlijke selectie van een goed model garanderen. Men kan tevens het modelbouwen als een leerproces vormgeven, bijvoorbeeld door voor N data een model te maken, dit te evalueren voor M_1 waarnemingen, eventuele gebreken opsporen en het model aanpassen en schatten voor M_2 waarnemingen, enzovoorts. Als men beschikt over M_1 tot en met M_{1000} steekproeven dan mag men wellicht verwachten dat het laatste model over zekere optimale eigenschappen beschikt.

3.4 Parameters schatten, diagnostiek en modelkeuze

De laatste kanttekening die ik wil plaatsen bij de econometrische analyse van grote gegevensbestanden betreft de meer statistische analyse van het model. In deze paragraaf maak ik slechts een paar opmerkingen, en ik verwijs de lezer naar het artikel van Granger (1997) waarin vergelijkbare en aanvullende

opmerkingen worden gemaakt.

Stel men beschikt over N waarnemingen voor drie variabelen y, x_1 en x_2. De vaak gehanteerde procedure om tot een adequaat en mogelijk nuttig model te komen start met een, zeg, basismodel (ik laat even in het midden of het een lineair of niet–lineair model is) waarvan de parameters worden geschat met een zekere schattingsmethode. Die methode levert standaardfouten, en t–waarden. Deze t–waarden zijn een functie van N, van de standaardfout van het hele model en van de geschatte parameter. Als N in de miljoenen loopt, dan zal deze factor de t–waarde gaan overheersen. Losjes gezegd betekent dit dat op het eerste gezicht "alle parameters significant zijn". Een t–waarde van 2 lijkt dan niet een goed criterium om de bijdrage van een variabele te evalueren. In termen van Bayesiaanse analyse, waarbij een voorverdeling voor een parameter wordt vergeleken met de naverdeling gegeven de data, betekent dit dat de informatie in de gegevens al gauw die in de voorverdeling zal overheersen. In andere woorden, met moet wel een heel extreme voorverdeling veronderstellen om iets anders uit de data te halen dan hetgeen de data zelf al aangeven.

Voor het schatten van parameters wil men graag schatters hanteren die consistent en efficiënt zijn. Efficiënt betekent dat naarmate N groter wordt de schatter voor, zeg, θ steeds dichter in de buurt van de werkelijke θ komt. Soms beschikt men over meerdere schatters voor dezelfde parameter, met elk een zekere graad van efficiëntie. Bij kleine steekproeven heeft men allicht een voorkeur voor de meest efficiënte schatter. Bij grote steekproeven maakt dit echter niet meer zoveel uit. In dat geval kan men verkiezen de meest eenvoudige aanpak te hanteren.

Nu is het eigenlijk nooit verstandig om t–waarden te evalueren in het geval men nog niet heeft vastgesteld of het model wel aan bepaalde eisen voldoet, zie Hendry (1994). Het is daarom goed gebruik om een model te evalueren aan de hand van diagnostische toetsen. Veel handige toetsen zijn gebaseerd op het "Lagrange multiplier" [LM] principe, zie Engle (1984) voor een overzicht van dit en andere principes. Meestal vormt men dan een hulpregressie waarin aan het bestaande model (of residuen daarvan) een aantal verklarende variabelen (vaak functies van residuen) worden toegevoegd, en men onderzoekt of die toevoeging een verklarende waarde heeft. In veel gevallen kan men de relevante toets uitrekenen als N maal de R^2 van die hulpregressie. Wanneer N bijzonder groot is, kan men verwachten (alweer losjes gezegd) dat elke nulhypothese zal worden verworpen. Ofwel, er blijken aanwijzingen te zijn dat het model op alle terreinen verbeterd moet worden. Een toets op de

normaliteit van de residuen is een functie van N en de geschatte derde en vierde momenten van de verdeling van de residuen. Ook deze toets zal altijd lijken te verwerpen, en dus de impressie geven dat er veel afwijkende waarnemingen zijn.

In het geval men meerdere modellen naast elkaar wil zetten, om uit hun midden het beste model te kiezen, kan men gebruik maken van de modelkeuzecriteria die de volgende vorm hebben:

$$\text{Criterium} = N.f_1(\text{fit}) + f_2(\text{aantal parameters}), \tag{11}$$

met $\partial f_1(a)/\partial a < 0$ en $\partial f_2(a)/\partial a > 0$. De keuze voor f_1 en f_2 geeft aan dat er een afweging plaatsvindt tussen de fit (vaak een schatter van de residuele variantie) en het aantal parameters. Immers, als het aantal parameters toeneemt, verbetert automatisch de fit. Het model met de laagste waarde voor het criterium in (11) wordt verkozen. Voorbeelden van zulke criteria zijn de Akaike en Schwarz criteria, genoemd naar hun bedenkers. De uitdrukking in (11) geeft al aan dat de term N in het eerste deel de waarde van de criterium gaat overheersen wanneer het heel groot wordt. Met andere woorden, het aantal parameters doet er niet meer zoveel toe, wanneer men verschillende modellen met elkaar vergelijkt.

Al met al lijkt het er op dat de econometrische analyse van veel data andere dan de huidige methoden behoeft om parameterschattingen te evalueren, de veronderstellingen van het model te toetsen en om alternatieve modellen met elkaar te vergelijken. Misschien is het werkbaar om de gegevensverzameling in veel kleine steekproeven op te delen, om daarna toch weer de gewoonlijke aanpak te hanteren. Nadeel is dan natuurlijk dat men weer werkt met relatief kleine steekproeven. Het kan ook zijn dat de evaluatie van het model voor de waarnemingen in een achter de hand gehouden steekproef afdoende is om tot een nuttig uiteindelijk model te komen.

4. ENKELE LAATSTE OPMERKINGEN

In dit artikel heb ik geprobeerd aan te geven welke de uitdagingen zijn die gepaard gaan met de mogelijkheid om grote bedrijfeconomische gegevensbestanden op econometrische wijze te analyseren. Het lijkt mij evident dat men die vele uitdagingen aan moet gaan, al is het maar vanwege het feit dat die gegevens veel inzicht kunnen verschaffen in het economisch handelen van individuen. In

tegenstelling tot de macroeconomie, waarvan vaak maar een handjevol over alle individuen geaggregeerde jaarcijfers ter beschikking staan, kunnen grote bedrijfseconomische gegevensbestanden inzicht geven in het gedrag van individuen op transactieniveau. Aangezien dit het meest elementaire niveau is van economisch handelen, kunnen deze gegevens, wanneer ze adequaat zijn samengevat in een econometrisch model, nuttig zijn voor het ontwikkelen van economische theorie. Zo'n theorie hoeft dan mogelijkerwijs niet uit te gaan van een zogenaamd representatieve economische agent.

De econometrische methoden, die vooral van belang zijn voor de analyse van kleine hoeveelheden gegevens, zullen minder relevant zijn voor de analyse van de grote bedrijfseconomische gegevensbestanden. Voorbeelden hiervan zijn methoden zoals pretesting en de bootstrap. Het lijkt ook mogelijk om eenvoudig toe te passen efficiënte schattingsmethoden te verkiezen boven ingewikkelde methoden met iets meer efficiëntie. Het onderzoek naar de eigenschappen van schatters in kleine steekproeven is ook minder van belang. Verder zijn alle methoden die naar exacte resultaten streven eveneens minder bruikbaar voor mega–panels, omdat deze exacte methoden vaak direct afhankelijk zijn van het aantal waarnemingen. Tenslotte is het de vraag op welke wijze de Bayesiaanse technieken relevant blijven.

Daarentegen kan men verwachten dat de grote gegevensbestanden ook weer aanleiding zijn tot het ontwikkelen van nieuwe econometrische methoden of het aanpassen van de huidige methoden aan de nieuwe situatie. Complexe modellen met ingenieuze structuren kunnen worden ontwikkeld en toegepast. De aspecten van aggregatie en pooling binnen een mega–panel worden zeer actueel. Gezien de mogelijkheid dat er veel afwijkende waarnemingen tussen zitten lijkt het gewenst om robuuste schattingsmethoden toe te passen, bij voorkeur voor complexe niet–lineaire modellen zoals neurale netwerken. Clusteringmethoden zullen eveneens relevant blijken. Niet zozeer de eigenschappen van de schatters, alswel de stappen die men moet zetten (en dat kunnen er heel veel zijn) om tot een optimale indeling van de data te komen zullen nader moeten worden bestudeerd. Tenslotte zullen nieuwe methoden moeten worden bedacht die de onderzoeker in staat stellen een model te evalueren, aangezien de gebruikelijke t–waarden, veel diagnostische LM–toetsen en de bekende modelkeuzecriteria niet zo nuttig lijken in het geval men de beschikking heeft over miljoenen waarnemingen.

Op basis van al het bovenstaande lijkt het niet onzinnig om te stellen dat de grote bedrijfseconomische gegevensbestanden de econometrische

wetenschap de gelegenheid geven progressie te maken door nieuwe methoden van analyse te onwikkelen. Het is wellicht prematuur om stellig te concluderen dat het verleden van de econometrie werd gekenmerkt door een samenspel met de macroeconomie en dat de toekomst van econometrie vooral zal worden gekenmerkt door bedrijfseconomische toepassingen, maar het is toch wel mijn verwachting dat de grote gegevensbestanden in juist de bedrijfseconomie de econometrie als wetenschap een flinke impuls zullen geven.

LITERATUURVERWIJZINGEN

Andersen, T.G. en T. Bollerslev (1995), Intraday Seasonality and Volatility Persistence in Financial Markets, in *Proceedings of the First International Conference on High Frequency Data in Finance*, Zurich.

Belsley, D.A., E. Kuh en R.E. Welsch (1980), *Regression Diagnostics: Identification of Influential Data and Sources of Collinearity*, New York; Wiley.

Blattberg, R.C. en E.I. George (1991), Shrinkage Estimation of Price and Promotional Elasticities: Seemingly Unrelated Equations, *Journal of* the *American Statistical Association*,

Chen, C. and L. Liu (1993), Joint Estimation of Model Parameters and Outlier Effects in Time Series, *Journal of the American Statistical Association*, 88, 284–297.

Engle, R.F. (1984), Wald, Likelihood Ratio, and Lagrange Multiplier Tests in Econometrics, in *Handbook of Econometrics*, *Volume II* (Z. Griliches en M.D. Intrilligator, red.), Amsterdam: Elsevier.

Fan, J. en I. Gijbels (1996), *Local Polynomial Modelling and Its Applications*, London: Chapman and Hall.

Franses, P.H. (1997), On the Econometrics of Modeling Direct Marketing Response, Ongepubliceerd manuscript, Erasmus Universiteit Rotterdam and IRIS Kwantitatief Onderzoek, ROBECO Group Rotterdam.

Franses, P.H. (1998, red.), *Statistical Analysis of Large Data Sets in Business Economics*, Thema editie van *Statistica Neerlandica.*

Franses, P.H. en G. Draisma (1997a), Recognizing Changing Seasonal Patterns Using Artificial Neural Networks, *Journal of Econometrics*, te verschijnen (najaar 1997).

Franses, P.H. en G. Draisma (1997b), Modeling Varying Elasticity Using Neural Networks for Market Share, Ongepubliceerd manuscript, Erasmus Universiteit Rotterdam.

Granger, C.W.J. (1997), Extracting Information from Mega–Panels and High–Frequency Data, Ongepubliceerd en incompleet manuscript, Department of Economics, University of California, San Diego, te verschijnen in Franses (1998, red.).

Hall, P. (1994), Methodology and Theory for the Bootstrap, in *Handbook of Econometrics*, *Volume IV* (R.F. Engle en D.L. McFadden, red.), Amsterdam: Elsevier.

Hastie, T.J. en R.J Tibshirani (1990), *Generalized Additive Models*, London:

Chapman Hall.

Hausman, J.A. (1978), Specification Tests in Econometrics, *Econometrica*, 46, 1251–1272.

Hendry, D.F. (1995), *Dynamic Econometrics*, Oxford: Oxford University Press.

Judge, G.G. en M.E. Bock (1978), *The Statistical Implications of Pre-Testing and Stein-Rule Estimators in Econmerics*, Amsterdam: North-Holland.

Kuan, C.–M. en H. White (1994), Artificial Neural Networks: An Econometric Perspective (met discussie), *Econometric Reviews*, 13, 1–91

Lovell, M. (1983), Data Mining, *Review of Economics and Statistics*, 65, 1–11.

Lucas, A. (1996), *Outlier Robust Unit Root Analysis*, Amsterdam: Thesis.

Maddala, G.S., R.P Trost, H. Li en F. Joutz (1997), Estimation of Short–Run and Long–Run Elasticities of Energy Demand From Panel Data Using Shrinkage Estimators, *Journal of Business and Economic Statistics*, 15, 90–100.

Ripley, B.D. (1994), Neural Networks and Related Methods for Classification, *Journal of the Royal Statistical Society* B, 56, 409–456.

Tsay, R.S. (1988), Outliers, Level Shifts, and Variance Changes in Time Series, *Journal of Forecasting*, 7, 1–20.