

Aggregeringsnivå, modellspecifikation och ekologisk inferens: en studie av de svenska valen 1944 och 1979

AV SVANTE ERSSON OCH INGEMAR WÖRLUND

Inledning

Med ekologidata eller aggregatdata avses inom samhällsvetenskaperna vanligen officiell statistik som samlats in för administrativa enheter på olika nivåer. Det kan vara statistik rörande fördelning av röster vid val till folkvalda församlingar eller regelbundet återkommande folkräkningar för att ge några exempel. Kännetecknande för dessa data är att de bygger på uppgifter över individer som ställs samman – aggregeras – till att gälla administrativa enheter, från vilka enheter någon återkoppling till uppgifter om enskilda individer ej är möjlig att göra.

Ekologidata kan därför användas för att beskriva variationer i politiska partiers röstetal eller andelar sysselsatta inom olika näringsgrenar inom ett land utifrån en given administrativ nivå. Med partier som analysobjekt kan det utifrån aggregatdata vara möjligt att skatta modeller som i olika avseenden kan förklara variationer i partiernas väljarstöd inom ett land (Brown, 1982). Med en sådan ansats inriktas analysen mot användningen av aggregatdata som något i sig intressant. Men det finns andra ansatser där man i uppbyggnaden av aggregatdata ser en möjlighet att utifrån dem försöka dra slutsatser om grupper av individer. Försök att dra slutsatser av denna art – ekologiska inferenser – har gjorts i olika sammanhang, men vanligtvis har de bemötts med stor skepsis alltsedan WS Robinson i sin artikel "Ecological correlations and the behavior of individuals" (1950) menade att alla försök till ekologiska inferenser obönhörligen leder till "ekologiska felslut". Rädslan att göra ekologiska felslut har därför gjort att man inom samhällsvetenskaperna ryggar tillbaka inför användningen av aggregatdata och i stor utsträckning inriktat sig på användning av direkt individbaserade data; i varje fall har aggregatdata ofta setts som underlägset i jämförelse med individdata.

Det är ändå den senare ansatsen vad gäller användning av aggregatdata som vår uppsats skall ägnas. Vår avsikt är att med svensk valdata aggregerade till olika nivåer pröva möjligheten att för två skilda tidsperioder göra ekologiska inferenser gällande väljarbeteendet hos grupper av individer med skilda näringsgrenstillhörighet. Framför allt för historiska valdata finns ett intresse att pröva ekologiska inferenser eftersom samtida surveydata rörande väljarbeteendet helt saknas. Prövningen av ekologiska inferenser har även ett intresse vad gäller användningen av nutida aggregatdata eftersom ekologiska inferenser i det sammanhanget kan ses som ett sätt att pröva hur väl specificerade aggregatdata-modeller är; en rimlig ekologisk inferens är en indikation på att modellen är väl specificerad.

Problemen kring ekologisk inferens

Robinsons artikel hade som antytt en stor genomslagskraft. Så utvecklade Alker (1969) med utgångspunkt från Robinson en typologi över olika slag av ekologiska felslut. Andra åter har diskuterat skilda förklaringar till att aggregeringsbias som leder till ekologiska felslut uppstår (Hannan, 1971; Hammond, 1973; Firebaugh, 1978; Burstein, 1978). Men såväl Robinson som Alker har mötts av invändningar och det har efterhand vuxit fram en rik litteratur kring frågan om möjligheten att göra ekologiska inferenser (Langbein & Lichman, 1978).

Även om ingen förnekar förekomsten av aggregeringsbias så har man genom simuleringar visat att aggregeringsbias inte alltid är nödvändig och kanske inte heller alltid så sannolik som man tidigare trott (Feisenmaier, 1985). Samtidigt har man utarbetat alternativa metoder för ekologisk inferens som skiljer sig från den av Robinson brukade ekologiska korrelationsmetoden. Vi kan här skilja mellan två alternativa vägar, dels en ut-

veckling av ekologisk korrelation, dels användning av ekologisk regression.

Det exempel som Robinson använde sig av gällde sambandet mellan egenskapen att vara svart och egenskapen att vara icke läskunnig på individnivå respektive delstatsnivå i USA i början av 1930-talet. Sambandet mättes på såväl individnivå som delstatsnivå med Pearson's korrelationskoefficient (r) som ju är avsett att användas för data på intervallskalenivå. Just detta förhållande att individsambandet mätts med ett olämpligt sambandsmått har tjänat som utgångspunkt för den danske statsvetaren SR Thomsens utveckling av den ekologiska korrelationsmetoden (Thomsen, 1987a; 1987b). Vad Thomsen – i all korthet – har gjort är att han utvecklat en teknik för att transformera den ekologiska korrelationen på aggregatnivå till att svara mot en individuell tetrakorisk korrelation; utifrån den skattade korrelationen på individnivå kan man så skatta olika proportioner för grupper av individer. De resultat som Thomsen redovisat är mycket lovande och visar en förvånansvärt god överensstämmelse med motsvarande skattningar gjorda baserade på surveydata (Thomsen, 1987a).

Än så länge är Thomsen tämligen ensam i sin användning av ekologisk korrelation. De flesta försök att göra ekologisk inferens har i stället byggts på ekologisk regression. Grunden för den ekologiska regressionen las av Goodman (1953; 1959) men principerna för den har behandlats av många andra (Stokes, 1969; Langbein & Lichtman, 1978; Lohmöller & Falter, 1986). Det är framför allt två egenskaper som gjort den ekologiska regressionen populär; dels kan regressionskoefficienterna – konstant och koefficient – tolkas som proportioner av en befolkningsgrupp med vissa bestämda egenskaper, dels har man visat att den ostandardiserade regressionskoefficienten inte är lika känslig för aggregeringsbias som den standardiserade korrelationskoefficienten är (Hammond, 1973; Hannan & Burstein, 1974). Förutsättningen för att arbeta med ekologisk regression är alltså att man utgår från proportioner av befolkningsgrupper – t ex proportioner av väljare som summerar till ett – så att regressionskoefficienterna också skall kunna tolkas som proportioner. Detta innebär att de proportioner som skall skattas bör hamna inom intervallet 0–1 för att anses vara giltiga; alla skattningar över 1 eller mindre än 0 är uppenbarligen ogiltiga. Problemet med ekologisk regression är emellertid att man inte alltid erhåller giltiga

skattningar i denna mening. Förklaringen till detta är att antagandet som ligger bakom den ekologiska regressionen – att de sannolikheter för ett bestämt beteende som regressionskoefficienterna uttrycker inte är desamma för alla individer i alla administrativa enheter, bortsett från en slumpmässig variation – inte alltid håller fullt ut. För att komma runt detta problem har man därför utvecklat olika alternativ inom ekologisk regressionsanalys.

En väg är att utnyttja regioner som inom sig är så homogena som möjligt och samtidigt skiljer sig så mycket som möjligt från andra regioner. Tanken är att inom de homogena regionerna är sannolikheterna för ett bestämt beteende – "conditional probabilities" – likartade, varför rimliga skattningar av proportioner bör vara möjliga att erhålla. En konsekvens av detta sökande efter homogena regioner är också att man eftersträvar att finna databärande enheter på lägsta tänkbara administrativa nivå (Jones, 1972; Lichtman & Langbein, 1978).

En annan väg som också prövats är den som kan kallas "constrained" eller "restricted" regression (Lewin et al, 1972; Schadee, 1987). Utgångspunkten här är att alla skattade proportioner bör hamna inom intervallet 0–1. I den mån skattade proportioner faller utanför det giltiga intervallet försöker man med program som gör iterativa skattningar erhålla proportioner som är rimliga. Tonvikten i denna ansats är alltså att genom utnyttjandet av olika skattningsförfaranden komma fram till tolkningsbara proportioner som kan ligga till grund för ekologiska inferenser.

En tredje väg – och den väg som vi skall försöka följa i denna uppsats – utgår från att i den mån som den ekologiska regressionen ger orimliga proportioner som resultat, så orsakas den av att den ekologiska regressionsmodellen har brister av olika slag (Hanushek et al., 1974; Langbein & Lichtman, 1978). Det kan vara så att aggregatmodellen är bristfälligt specificerad i förhållande till den individrelation som skall skattas (jfr Flanigan & Zingale, 1981). Det kan också föreligga mätfel hos de aggregerade variabler som ligger till grund för analysen. Gemensamt för studier inom denna tradition är att man genom att förbättra den ekologiska regressionsmodellen tror sig kunna göra rimliga ekologiska inferenser på individnivå. Ett problem med denna ansats är dock att de skattade proportionerna ibland tycks vara instabila och det är inte heller alltid säkert att de summerar till ett om skilda regressionsmodeller

utnyttjats för olika partier. Ett annat problem är att den ekologiska regressionen tycks vara bäst anpassad för skattningar vid binära val, dvs när endast två alternativ finns att välja mellan som i ett tvåpartisystem (Crewe & Payne, 1976).

Syfte

Som redan antytts är vi i denna uppsats intresserade av att pröva om ekologiska inferenser är möjliga att göra rörande väljarbeteendet hos grupper av individer med olika näringsgrenstillhörighet. Vi väljer att göra denna prövning för de två valen 1944 och 1979. Valet 1944 har valts eftersom det är det sista valtillfälle för vilket vi förfogar över ekologidata på en låg administrativ nivå med ett stort antal enheter (N=2479). Det senare valet 1979 har valts beroende på att vi för denna tidpunkt har tillgång till motsvarande surveydata på individnivå.

De ekologiska inferenserna baseras på ekologisk regression varvid vi är speciellt intresserade av att pröva om rimligheten i skattade proportioner påverkas av regressionsmodellens specificering. Vi är även intresserade att se om ekologiska inferenser påverkas av val av aggregeringsnivå. Utifrån de analyser som presenteras i uppsatsen hoppas vi kunna dra slutsatser om bland annat följande problem:

- finns det någon aggregeringsnivå som bör föredras eller undvikas?
- kan man med ekologisk regression som grund fastställa några kriterier för när ekologiska inferenser är såväl giltiga som rimliga?
- kan man utifrån jämförelser med skattningar baserade på surveydata avgöra om ekologiska skattningar är möjliga eller alls lämpliga att arbeta vidare med?

Data och metod

Endast aggregerade databärare utnyttjas för analyserna av valet 1944. Utgångsmaterial är det valdataarkiv som insamlats av projektet "Politiska partier, regional spridning, regional balans" vid Umeå universitets statsvetenskapliga institution. Antalet kommuner i detta arkiv är 2576, men för att erhålla över tiden konstanta enheter har de för denna uppsats aggregerats till ett antal av 2479. För att variera aggregeringsnivå har så dessa 2479 kommuner i sin tur aggregerats till högre administrativa nivåer = mindre antal enheter. Förutom kommunnivån om 2479 enheter finns ytterligare fyra aggregeringsnivåer: kommuner

enligt 1952 års indelning (N=991); häradsindelning (N=409); kommuner enligt 1985 års indelning (N=283); samt län enligt vår nutida indelning (N=24). Variabler som finns på samtliga fem nivåer är dels näringsgrensfördelning: proportion sysselsatta inom jordbruk, industri och övriga näringar; dels de politiska partiernas väljarstöd: proportioner för socialdemokraterna (SAP), borgerliga partier, socialistiska partier och icke röstande i förhållande till samtliga röstberättigade. Utöver dessa data ingår också variabler som mäter partiernas väljarstöd vid valet 1917 (Wörlund & Ersson, 1987a; 1987b). Vad gäller näringsgrensdata tarvar de några kommentarer. För det första avser de 1940 medan valdata hänför sig till 1944 varför det finns en tidsmässig skillnad mellan datamängderna. För det andra så refererar näringsgrensdata från 1940 till hela befolkningen och inte enbart till de yrkesverksamma. Innebörden av detta är att näringsgrenspopulationen (hela befolkningen) skiljer sig i storlek från väljarpopulationen (de röstberättigade). Det antagande vi därför tvingas göra – som inte är helt realistiskt – är att näringsgrensfördelningen inom befolkningen som helhet är ungefärligen densamma som bland de röstberättigade.

Analyserna av valet 1979 baseras på såväl aggregeratdata som surveydata. Surveydata har erhållits från väljarundersökningen 1979 och de variabler som används är de som mäter näringsgrenstillhörighet (V220) och partival 1979 (V290) (Holmberg & Nordlöf, 1982).¹ Aggregeratdata för valet finns endast för en nivå – kommunnivå (N=283). I stort sett samma typ av variabler som nyttjats för analysen av valet 1944 används vid analysen av valet 1979; det är partidata och näringsgrensdata uttryckta som proportioner plus en del andra prediktorvariabler som t ex historisk tradition, politisk konkurrens och olika omgivningsfaktorer av typ urbaniseringsgrad och inkomster.

Det är med ekologisk regression som metod som vi gör ekologiska inferenser för 1944 och 1979 års val baserade på aggregeratdata. De inferenser vi är intresserade av innebär försök att skatta hur stor andel av de sysselsatta inom respektive jordbruk, industri och övriga näringar som röstar på olika partier. De skattade proportionerna kommer att erhållas utifrån tre grundmodeller för ekologisk regression: regression utan intercept [1]; bivariat regression [2]; multipel regression [3]. Låt oss här bara kortfattat antyda hur man kan erhålla de skattade

proportionerna utifrån de tre grundmodellerna. Vi kan som exempel ta en skattning av proportionen sysselsatta inom industrin som 1979 röstar på socialdemokraterna (SAP). Följande modeller kan då specificeras:

$$\text{SAP} = b_1\text{IND} + b_2\text{JBR} + b_3\text{ÖVR} + e \quad [1]$$

$$\text{SAP} = a + b_1\text{IND} + e \quad [2]$$

$$\text{SAP} = a + b_1\text{IND} + b_2\text{HIST} + b_3\text{OMG} + e \quad [3]$$

där SAP står för proportion röstande på socialdemokraterna; IND, JBR och ÖVR står för näringsgrenstillhörighet; HIST och OMG står för en historisk traditionsfaktor respektive en omgivningsfaktor. I modell [1] kan koefficienten b_1 tolkas som proportionen industrissysselsatta som röstar på SAP; i modell [2] erhålls motsvarande proportion genom att konstanten a adderas till koefficienten b_1 ; i modell [3] skattas samma proportion genom att konstanten a adderas till koefficienten b_1 plus värdet för summan b_2 multiplicerat med medelvärdet för variabeln HIST plus värdet för summan b_3 multiplicerad med medelvärdet för variabeln OMG.

Vad gäller ekologisk regression som metod för att göra ekologiska inferenser är det främst tre problem vi är intresserade av att beakta. Det första berör frågan om man kan påvisa någon form av aggregeringsbias när man går från en lägre nivå med ett stort antal analysenheter till en högre nivå med ett mindre antal enheter. Om de skattade proportionerna förändras från en nivå till en annan kan det vara en följd av aggregeringsbias. För att mer exakt pröva eventuell förekomst av aggregeringsbias kommer vi dessutom att närmare jämföra två aggregeringsnivåer genom att testa den s.k. "X-rule" som föreslagits av Firebaugh (1978). Detta problem kommer att beaktas vid analysen av valet 1944.

Ett andra problem gäller huruvida en modell som prövas är riktigt specificerad eller ej. I praktiken känner man aldrig den "sanna" modellen och därför kan heller aldrig det "sanna" värdet för en regressionskoefficient skattas, utan man kan endast komma mer eller mindre nära det "sanna" värdet. Bortsett från rena mätfel är en felaktig modellspecifikation den största svårigheten vid regressionsanalys (Deegan, 1976). I en modell där en viktig prediktorvariabel utelämnats tenderar parametrarna i den ingående modellen att överskattas. Därför är det viktigt att i en reg-

ressionsmodell försöka ta in variabler som utifrån teoretiska och empiriska erfarenheter har effekt på den beroende variabeln. Tar man å andra sidan in alltför många variabler riskerar man att ställas inför ett annat problem: multikolliniaritet. Det problem som multikolliniaritet är uttryck för är att två eller flera prediktorvariabler starkt samvarierar, varför det kan vara svårt att skilja effekten av en prediktorvariabel från effekten av en annan prediktorvariabel på den beroende variabeln. Ändå bör man nog inte överbetona problemet med multikolliniaritet eftersom det i olika sammanhang visats att regressionskoefficienternas storlek endast marginellt påverkas av kolliniariteten; vad som i stället påverkas är varianserna för de skattade regressionskoefficienterna. Därför kan man nog hävda att problemet med felaktig modellspecifikation är allvarigare än problemet med multikolliniaritet (Hanushek & Jackson, 1977). Av denna anledning kommer vi att pröva ett flertal olika modellspecifikationer för att se om någon modell ger bättre skattningar än andra modeller. Något entydigt kriterium för att avgöra om en modell är väl specificerad eller ej finns inte, däremot kan man vid jämförelser mellan olika modeller med en och samma beroende variabel använda det justerade R^2 (R^2A) som mått på modellens anpassning till data (King, 1986:677). Dessutom rapporterar vi även ett mått för multikolliniaritet - R^2D - som uttrycker R^2 när en oberoende variabel regressas mot övriga oberoende variabler i modellen.

Ett tredje problem är huruvida man vid aggregatdataanalyser bör använda viktade eller oviktade data. Erfarenheter från regressionsanalyser med aggregerade data visar att heteroskedasticitet kan inverka menligt på skattningar som görs. Vanligt har då varit att de databärande enheternas varierande storlek medverkat till den uppkomna heteroskedasticiteten (Lemieux, 1976). För att beakta detta problem kommer vi att redovisa såväl viktade som oviktade skattningar; viktade skattningar baseras på att de databärande enheterna i fråga viktas efter befolkningsstorlek.

Till detta skall också läggas att när vi gör ekologiska inferenser rörande politiska partiers väljarstöd görs skillnad mellan inferenser för enskilda partier och inferenser för politiska block. Utifrån tidigare redovisade erfarenheter bör man förvänta sig säkrare skattningar som rör politiska block (nära binärt val) än skattningar för enskilda partier. Genomgående gäller att proportioner som redovisas för enskilda partier eller partiblock

avser proportioner i förhållande till samtliga röstberättigade. Därav följer att förutom för de traditionella partierna redovisas skattningar för gruppen icke röstande.

Ett av syftena med uppsatsen är att försöka avgöra vilka skattningar som är mest rimliga eller vilka skattningar som skiljer sig mest från varandra. Vad gäller valet 1979 finns survey-skattningar som en referenspunkt medan vi för analysen av valet 1944 är hänvisade till andra skattningar av icke helt jämförbara grupper. För båda analyserna gäller det att avgöra vilka skattningar som närmast sammanfaller. För att mäta grad av överensstämmelse används därför två mått: ett första mått (T) mäter hur nära två värden ligger en given diagonal i ett koordinatsystem (Klevmarken, 1970:262); ett andra mått (NS) mäter hur nära i förhållande till varandra en serie motsvarande cellvärden i två tabeller ligger (Thomsen, 1987a:28). Med dessa två mått, (T,NS) kan vi avläsa vilka ekologisk-skattningar som visar den största överensstämmelsen med survey-skattningar för 1979.

Den följande analysen delas in i två avsnitt: ett första avsnitt ägnas valet 1944 med skattningar för olika aggregeringsnivåer som avser dels SAP enbart, dels socialister, borgerliga och icke röstande; det andra analysavsnittet ägnas valet 1979 där jämförelsen mellan aggregat-dataskattningar och surveydata-skattningar står i förgrunden. För att komma fram till en bästa möjlig skattning utifrån aggregatdata krävs ett stort antal modellprovningar.

Ekologiska inferenser för valet 1944

I detta avsnitt skall vi redovisa skattningar av hur man inom grupper med varierande näringsgruppstillhörighet röstat på olika partier vid valet 1944. Vi är intresserade av att veta om och i så fall hur dessa skattningar påverkas av olika aggregeringsnivåer likaväl som skilda modellspecifikationer. Låt oss emellertid först se om aggregeringsnivå inverkar på variabelvärdenas fördelningar i olika avseenden; relevanta data redovisas i tabell 1 för fem aggregeringsnivåer såväl viktade som oviktade.

Det är framförallt två förhållanden som förtjänar att påverkas. Dels är det tydligt att variationen i variablerna minskar med ökande aggregering, dels framgår klart att skillnaden mellan viktade och oviktade värden är betydande på låga aggregeringsnivåer men mycket måttliga på den högsta

aggregeringsnivån, länet. Genomgående gäller också att skillnaderna i fördelningar – bortsett från medelvärden – är störst mellan län och övriga aggregeringsnivåer.

När man laborerar med olika aggregeringsnivåer är det också av intresse att skapa sig en uppfattning om hur mycket av variationen i en variabel man förlorar när man går från en lägre nivå till en annan, högre nivå. Hur mycket av den totala variationen kan tillskrivas variationen inom den högre administrativa enheter gentemot den variation som finns mellan de högre administrativa enheterna? Utan att ställa upp några definitiva gränser för vad som är accetabelt kan man tycka det vara rimligt att variationen mellan de nya högre enheterna bör vara större än inom-gruppsvariationen i samma enheter. Låt oss här med en variansanalys belysa hur aggregeringar påverkar inom- och mellan-gruppsvariationen. E^2 -värdet som redovisas i tabell 2 nedan skall tolkas så att ju högre värde, desto mer av variansen i variabeln i fråga kan föras tillbaka på gruppstillhörigheten; om E^2 är .5 kan således den ena hälften av variabelns varians tillskrivas inom-gruppsvariation, medan den andra hälften kan tillskrivas mellan-gruppsvariationen. I tabellen redovisas samma variabeluppsättning som i föregående tabell med både viktade och oviktade värden.

Även här framträder en skillnad mellan viktade och oviktade värden, liksom skillnader mellan enskilda variabler. Det mönster som framträder är annars att län som aggregeringsnivå skiljer sig från övriga aggregeringsnivåer. Gentemot kommun₂₄₇₉ visar länet för alla variabler – viktade som oviktade – hela tiden E^2 -värden mindre än .5. Nära denna kritiska gräns .5 – ibland över, ibland under – ligger kommun₂₈₃, men jämförd med häradnivån är skillnaderna i värden inte speciellt markant. Den försiktiga slutsats vi på detta stadium möjligen kan dra är att användningen av databärare på en så hög aggregeringsnivå som län ter sig problematisk och att den högsta aggregeringsnivå som det kan vara försvarligt att arbeta med tycks vara den nuvarande kommunindelningen, dvs kommun₂₈₃.

Ett test har utvecklats av Firebaugh (1978:560) för att pröva om det föreligger aggregeringsbias när man går från en lägre aggregeringsnivå till en högre aggregeringsnivå. För att genomföra testet krävs att man förfogar över data dels för den lägre aggregeringsnivån (X_1), dels för den högre aggregeringsnivån som kan överföras till de databärande enheterna på den lägre nivån (MX_1); MX_1

Tabell 1. Fördelningar för olika variabler på olika aggregeringsnivåer: viktade och oviktade värden

Variabel	Värde	Viktning	Aggregeringsnivå				
			Kom ₂₄₇₉	Kom ₉₉₁	Har ₄₀₉	Kom ₂₈₃	Lan ₂₄
SAP44	Medel	Vikt	33.6	33.6	33.6	33.6	33.6
	Medel	Ovikt	28.5	31.7	32.5	33.1	33.1
	SD	Vikt	10.4	9.6	8.1	7.4	5.3
	SD	Ovikt	13.3	11.0	9.4	8.7	5.1
	CV	Vikt	31.0	28.4	24.2	22.1	15.6
	CV	Ovikt	46.8	34.8	29.1	26.4	15.5
	SKEW	Vikt	.019	.150	.133	.353	.756
	SKEW	Ovikt	.135	.223	-.134	.260	.652
	SOC44	Medel	Vikt	41.1	41.1	41.1	41.0
Medel		Ovikt	31.6	36.5	38.0	39.3	39.6
SD		Vikt	12.9	12.0	10.3	9.3	5.9
SD		Ovikt	15.2	13.4	11.8	11.2	6.6
CV		Vikt	31.4	29.3	25.0	22.5	14.4
CV		Ovikt	48.2	36.8	31.0	28.6	16.6
SKEW		Vikt	-.368	-.269	-.381	-.217	-.530
SKEW		Ovikt	.184	.214	-.174	.120	-.201
BORG44		Medel	Vikt	30.9	30.9	30.9	30.9
	Medel	Ovikt	40.0	34.4	33.0	31.6	31.7
	SD	Vikt	11.1	10.3	9.0	8.1	5.4
	SD	Ovikt	14.6	11.9	10.5	10.1	6.6
	CV	Vikt	35.9	33.3	29.2	26.2	17.5
	CV	Ovikt	36.4	34.5	32.0	32.0	20.7
	SKEW	Vikt	.892	.655	.676	.457	.468
	SKEW	Ovikt	.354	.185	.105	.150	.401
	ICKE44	Medel	Vikt	28.1	28.1	28.1	28.1
Medel		Ovikt	28.4	29.0	29.1	29.1	28.7
SD		Vikt	7.2	6.4	5.5	5.3	3.5
SD		Ovikt	9.1	7.8	6.9	6.4	3.3
CV		Vikt	25.7	23.0	19.7	18.7	12.3
CV		Ovikt	31.9	26.7	23.7	21.9	11.4
SKEW		Vikt	.253	.704	.780	.721	-.670
SKEW		Ovikt	.321	.787	1.512	.693	-.799
IND40		Medel	Vikt	36.5	36.4	36.4	36.1
	Medel	Ovikt	23.0	29.1	32.3	33.0	34.3
	SD	Vikt	16.9	15.9	13.5	11.7	6.3
	SD	Ovikt	15.4	16.0	15.0	13.9	7.3
	CV	Vikt	46.2	43.7	37.1	32.2	17.4
	CV	Ovikt	66.9	55.1	46.5	42.0	21.4
	SKEW	Vikt	.062	.153	.265	.128	-.963
	SKEW	Ovikt	1.127	.796	.548	.389	-.442

Tabell 2. Aggregering av variabler från kommun (N=2479) till högre nivåer; effekt av variationen mellan grupper uttryckt som E^2 .

Variabel	Viktning	Aggregeringsnivå			
		Kommun ₉₉₁	Härad ₄₀₉	Kommun ₂₈₃	Län ₂₄
SAP44	Vikt	.85	.61	.51	.24
SAP44	Ovikt	.70	.51	.46	.28
SOC44	Vikt	.88	.63	.51	.20
SOC44	Ovikt	.73	.53	.47	.27
BORG44	Vikt	.87	.66	.53	.23
BORG44	Ovikt	.73	.57	.51	.31
ICKE44	Vikt	.82	.59	.53	.22
ICKE44	Ovikt	.65	.44	.42	.20
IND40	Vikt	.90	.64	.48	.14
IND40	Ovikt	.69	.40	.30	.07

är medelvärdet för variabel X_1 för den aggregerade enhet som respektive enhet på den lägre nivån tillhör. Firebaugh förklarar sin regel "Mx-rule" på detta sätt:

"Cross-level bias is absent when, and only when, $B_2 = 0$ in the structural equation $Y = a + B_1X_1 + B_2MX_1 + e \dots$ " (Firebaugh 1978:560; avvikelser från originalet vad gäller symbol-uppsättning).

En prövning av detta test har gjorts för en enkel modell applicerad på kommun₂₄₇₉ i förhållande till kommun₂₈₃ för att utröna om det föreligger någon aggregeringsbias på den högre nivån. Prövningen görs med följande modell och avser både viktade och oviktade värden:

$$SAP44 = a + B_1IND44_{2489} + b_2MIND44_{283} + e[4]$$

där $IND44_{2479}$ står för proportion sysselsatta inom industri på kommunal nivå med $N=2479$, medan $MIND44_{283}$ refererar till proportion sysselsatta inom industri på kommunal nivå med $N=283$. Skattningen av modellen ger följande parametervärden:

$$SAP44 = .263 + .567IND44_{2489} + .002MIND44_{283} \text{ (oviktad) respektive} \\ SAP44 = .330 + .495IND44_{2489} - .131MIND44_{283} \text{ (viktad).}$$

Av denna prövning att döma är inslaget av aggregeringsbias mindre besvärande när oviktade värden används än när viktade värden utnyttjas. Utfallet kan tyckas något överraskande med hänsyn tagen till tidigare redovisade resultat från variansanalysen, men det ger anledning till viss efter-

tanke vid tolkning av skattningar som baseras på viktade data.

Nästa steg i analysen av valet 1944 avser redovisning av skattade proportioner för sysselsatta inom industrin som röstat på socialdemokratin. Två modeller har prövats för samtliga fem aggregeringsnivåer med både viktade och oviktade data. I den ena modellen [5] är sysselsatta inom industri (IND40) enda prediktor medan i den andra modellen [6] dessutom ingår en historisk traditionsfaktor (SAP17) och en politisk konkurrensfaktor (SKP44) som prediktorvariabler. Det finns anledning anta att den senare modellen är bättre specificerad eftersom såväl historiska traditioner som politisk konkurrens inverkar på variationen i ett politiskt partis väljarstöd (jfr Selle, 1983). Resultatet av skattningarna av de två modellerna redovisas i tabell 3; modellerna specificeras på detta sätt:

$$SAP44 = a + b_1IND40 + e \quad [5]$$

$$SAP44 = a + b_1IND40 + b_2SAP17 + b_3SKP44 + e \quad [6]$$

I tabellen redovisas förutom den skattade proportionen sysselsatta inom industri som röstat på SAP också den förklarade variansen (R^2A) för motsvarande modell. Därvid kan först konstateras att variansen i den beroende variabeln har konsekvenser för modellernas förklarade varians: ju högre aggregeringsnivå, desto lägre R^2A . Markerade skillnader finns också för de skattade proportionerna beroende på val av aggregeringsnivå, val av modell eller viktning av data. En ökande aggregeringsnivå driver ner skattningarna, men det är endast skattningarna baserade på län

Tabell 3. Skattningar av proportionen industrisysselsatta som röstat på SAP 1944: modell [5] och [6]

	Modell [5]				Modell [6]			
	Viktad		Oviktad		Viktad		Oviktad	
Aggregeringsnivå	Prop	R ² A	Prop	R ² A	Prop	R ² A	Prop	R ² A
Kommun ₂₄₇₉	.574	.37	.637	.28	.533	.49	.495	.53
Kommun ₉₉₁	.567	.37	.604	.35	.529	.49	.517	.50
Härad ₄₀₉	.553	.32	.577	.35	.507	.45	.464	.50
Kommun ₂₈₃	.549	.32	.566	.31	.505	.42	.452	.45
Län ₂₄	.504	.06	.525	.14	.362	.23	.336	.27

som drastiskt avviker från övriga skattningar. Likaså leder den mer utvecklade modellen [6] till att proportionerna minskas jämfört med skattningar i modell [5]. Viktning av data påverkar också skattningarna på olika sätt; i modell [5] är de oviktade proportionerna högre än de viktade, medan det omvända gäller i modell [6].

Något självklart kriterium mot vilka våra skattningar kan avstämmas finns inte eftersom såväl Ingulfson och Hagmans (1950:169) som Lewins et al (1972:147) skattningar avser proportioner inom arbetarklassen och ej näringsgrenar som vi arbetar med. Låt oss ändå anta att de skattningar som dessa redovisat avseende arbetarklassen i stort också torde gälla sysselsatta inom industrin. Om så vore fallet skulle som referenspunkt andelen sysselsatta inom industrin som röstat på SAP vara mellan 70 och 71 % bland deltagande i val och mellan 50 och 51 % bland de röstberättigade. Om vi antar att denna proportion ligger mellan .50 och .51 så vore de bästa skattningarna de som erhålls med modell [6] för aggregeringsnivåer under länet, medan däremot den bästa skattningen för modell [5] skulle erhållas med län som databärande enhet. Nu vet vi inte de verkliga värdena men det verkar rimligt anta att de bästa skattningarna för SAP erhålls av den mer utvecklade modellen [6] när den appliceras på databärande enheter med ett N=283 eller mer.

För att erhålla ett bredare underlag för att avgöra vilka skattningar som är närmast överensstämmande redovisas andra skattade proportioner baserade på valet 1944. I stället för att titta på ett enskilt parti prövar vi här modeller som avser röstande på borgerliga partier (BORG44 = HÖ+BO+FP+ÖVR), socialistiska partier (SOC44 = SAP+SKP) samt icke röstande (ICKE44). Tre grundmodeller prövas varav vi i den multipla regressionsmodellen där så är lämpligt som prediktorer använder oss av historiska traditionsvariabler (BORG17, SOC17), valdelta-gande 1944 (VD44) och proportion yrkesverk-

samma 1940 (YV40). De tre grundmodeller som analyseras med SOC44, BORG44 och ICKE44 som beroende variabler kan därför specificeras som följer:

$$\text{BLOCK44} = b_1\text{JBR40} + b_2\text{IND40} + b_3\text{ÖVR40} + e \quad [7]$$

$$\text{BLOCK44} = a + b_1\text{NÄRING40} (\text{JBR, IND, ÖVR}) + e \quad [8]$$

$$\text{BLOCK44} = a + b_1\text{NÄRING40} + b_2\text{HIST17} + b_3\text{VD44} + b_4\text{YV40} + e \quad [9]$$

Med modellerna [7] till [9] som grund anpassas ett stort antal regressionsekvationer; 3 beroende variabler för varje grundmodell; fem skilda aggregeringsnivåer; och 2 skilda viktningssystem. I tabell 4 nedan presenteras endast de proportioner som skattats för socialistiska väljare.

Av tabellen framgår att skillnader finns mellan de skattade proportionerna. Endast en mindre del av dessa skillnader är emellertid beroende av viktningssystemet; så leder viktningen till att proportionen sysselsatta inom jordbruk som röstar på socialistiska partier ökas något. Val av aggregeringsnivå har en tydligare effekt på variationen i de skattade proportionerna, men bortser man från den högsta nivån – länet – är skillnaderna i skattningar förhållandevis små. Tydligt är i alla fall att skattningar med län som databärande enhet avviker från skattningar baserade på databärare på lägre aggregeringsnivåer. De största skillnaderna i skattningarna är i stället beroende av val av grundmodell där modell [7] avviker mest från de två övriga [8] och [9]. Så skattas proportionen industrisysselsatta betydligt högre i modellerna [7] och [8] än i modell [9], medan proportionen sysselsatta i övriga näringar skattas lägre i modell [7] än i modellerna [8] och [9]. De skillnader i skattningar som konstaterats för röstande på socialistpartier gäller också i stort för

Tabell 4. Skattningar av proportioner sysselsatta inom olika näringsgrenar röstande på socialistpartierna 1944: grundmodell [7] till [9]

Modell	Nivå (N)	Proportion jordbruk		Proportion industri		Proportion övriga	
		Viktad	Oviktad	Viktad	Oviktad	Viktad	Oviktad
[7]	2479	.204	.165	.742	.735	.235	.268
[7]	991	.204	.201	.739	.753	.239	.216
[7]	409	.213	.216	.722	.733	.249	.205
[7]	283	.207	.228	.709	.728	.273	.228
[7]	24	.252	.223	.585	.594	.370	.382
[8]	2479	.223	.150	.759	.777	.527	.662
[8]	991	.226	.186	.758	.758	.525	.547
[8]	409	.253	.230	.744	.727	.518	.504
[8]	283	.240	.198	.740	.727	.503	.466
[8]	24	.288	.229	.656	.665	.518	.575
[9]	2479	.280	.229	.587	.547	.479	.489
[9]	991	.277	.266	.582	.565	.481	.471
[9]	409	.285	.289	.568	.554	.492	.452
[9]	283	.289	.300	.552	.530	.436	.429
[9]	24	.212	.219	.458	.482	.623	.640

skattningar av röstande på borgerliga partier respektive icke röstande.

Låt oss även här relatera våra skattningar gentemot de skattningar som redovisats av Ingulfson och Hagman (1950:169) och Lewin et al (1972:147). Samma reservationer gäller som tidigare angående jämförbarhet mellan grupper avseende arbetarklass eller sysselsatta inom industri. Om vi ändå tillåter oss det antagandet att de proportioner som dessa skattat ligger rätt nära det sanna värdet så har vi ännu en referenspunkt att utgå från. Vi begränsar oss då till skattningen avseende sysselsatta inom industrin som röstat på socialistpartier vid valet 1944. Vi känner redan proportionen som avser SAP (.71 eller .70 respektive .51 eller .50) varför vi här måste komplettera med proportionen avseende SKP. Medan proportionen för SAP nära överensstämmer finns avvikelser i fråga om SKP; Lewins et al skattning för 1944 är .17 respektive .12 medan Ingulfson och Hagman för 1946 redovisar en skattad proportion om .12 respektive .09. Till detta kan följande kommentarer fogas: å ena sidan är det ett faktum att SKP 1946 i röstetal ökade till 11.2 % från 10.3 % 1944 varför proportionen sysselsatta inom industri borde vara högre 1946 än 1944; å andra sidan kan värdet för 1944 vara något över-skattat och värdet för 1946 vara något under-skattat. Om dessa skattningar nu ligger nära det verkliga värdet – låt oss anta det – skulle andelen sysselsatta inom industrin som röstat på SAP el-

ler SKP i förhållande till samtliga röstberättigade vara någonstans kring .60. Om vi använder .60 som riktvärde, vilka av våra ekologiskattningar hamnar närmast? Tydligt är att modell [9] ger de bästa skattningarna om man bortser från skattningar med län som databärare; en svag under-skattning av proportionen kan det vara fråga om, men denna avvikelse är mindre än den överskattning som modellerna [7] och [8] leder till. Återigen kan vi dock konstatera att med län som databärande enhet erhålls rimliga skattningar med modeller – [7] och [8] – som utifrån teoretiska utgångspunkter borde vara underspecificerade.

Eftersom vi inte känner de "sanna" proportionerna får vi här nöja oss med ett konstaterande att skattningar baserade på länsnivån skiljer ut sig från övriga skattningar; vidare har vi noterat att regressioner utan intercept – modell [7] – avviker något mer från övriga modellbaserade skattningar. Om tidigare i litteraturen redovisade skattningar av proportioner inom arbetarklassen som stöder SAP eller SKP kan ses som ett närmevärde för proportioner sysselsatta inom industrin som har samma politiska preferenser så antyder de dels att mer utvecklade modeller – bättre specificerade – ger rimligare skattningar för de fyra lägsta aggregeringsnivåerna, dels att för den högsta aggregeringsnivån ger den enklast specificerade modellen den rimligaste skattningen. För att hitta en bättre referenspunkt gentemot vilken de olika ekologiska inferenserna kan avstämmas

måste vi emellertid gå vidare till en analys av valet 1979.

Surveybaserade och ekologiska inferenser för valet 1979

Vi disponerar detta avsnitt så att först redovisas de skattningar som bygger på surveydata. Därefter presenterar vi de skattningar som är resultaten av en stor mängd prövade ekologdata-modeller vilka har följt i stort samma uppläggning som vid analysen av 1944 års val. Till sist försöker vi i detta avsnitt avgöra vilka ekologdata-skattningar som närmast överensstämmer med de skattningar som bygger på surveydata. Gemensamt för dessa redovisningar är att såväl skattningar för enskilda partier – från moderater till övriga samt icke-röstande – som skattningar för partiblock presenteras. De skattningar som avses är hela tiden proportionen sysselsatta inom en bestämd näringsgren – jordbruk, industri eller övriga näringar – som röstar på ett visst politiskt parti eller ett visst politiskt block.

Surveydata-skattningar

De survey-skattningar vi använder för 1979 års val bygger på data vi erhållit genom 1979 års väljarundersökning. Med vårt syfte är vi endast intresserade av data rörande partival 1979 (V290) och näringsgrenstillhörighet (V220). Låt oss först se på bortfallet. Det bortfall som avser partivalet är dels de som varken uppgett uppgifter om röstning eller valdeltagande (N=5), dels de som röstat men ej angett partival (N=784). Motsvarande bortfall för uppgifter angående näringstillhörighet är de som ej intervjuats (N=853), saknar uppgift (N=164) eller för vilka frågan ej är tillämplig (N=170). Sammantaget innebär detta internbortfall att det samlade antalet intervjupersoner som svarat på båda dessa frågor är nära 2 500 (N=2 484).

Klassificeringen av partivalet är självklar, medan man vad gäller näringsgrensfördelningen kan tänka sig olika varianter. Vi har valt att följa valundersökningens näringsgrensfördelning (VNI) och som sysselsatta inom jordbruk räknas VNI=1-3; till sysselsatta inom industri räknar vi VNI=11-42; följaktligen är sysselsatta inom övriga näringar VNI=51-99 (Holmberg & Nordlöf, 1982).

Beroende på det interna bortfallet visar sig andelen icke röstande i urvalet (5.2%) vara underskattad i jämförelse med det verkliga ickeröst-

det vid valet 1979 (9.3%). Av den anledningen har vi valt att redovisa dels *oviktade* skattningar utan korrigeringar för internbortfallet, dels *viktade* skattningar där urvalet (N=2484) viktats så att de svarar mot de andelar respektive partier eller partiblock hade bland väljarna vid valet 1979. I fråga om blockvalet har den lilla gruppen övriga hänförs till det socialistiska blocket eftersom en relativt stor andel av dessa kan antas ha stött APK.

Nedan redovisas i tabellerna 5 till 8 skattningar av andelar sysselsatta inom de tre nämnda näringsgrenarna dels med avseende på parti- eller blockval, dels med avseende på skilda viktningsskattningar. Det kan därvid konstateras att bortsett från gruppen icke-röstande är skillnaderna mellan viktade och icke-viktade skattningar ytterst små; skattningarna rörande blockval är desamma som gäller för partival, men de redovisas ändå separat för att underlätta senare jämförelser med motsvarande ekologdata-skattningar. De

Tabell 5. Survey-skattningar (viktade) av partival efter näringsgren 1979

Parti	Andel sysselsatta inom näringsgren		
	Jbr (7.2)	Ind (23.9)	Övr (68.9)
Icke (9.3)	8.9	10.2	9.0
Mod (18.4)	16.8	13.2	20.4
CP (16.4)	52.2	12.4	14.0
FP (9.6)	6.7	8.1	10.5
KDS (1.3)	0	1.1	1.5
SAP (39.2)	14.9	49.7	38.1
VPK (5.1)	0.6	5.1	5.6
Övr (0.7)	0	0.3	0.9

Tabell 6. Survey-skattningar (icke-viktade) av partival efter näringsgren 1979

Parti	Andel sysselsatta inom näringsgren		
	Jbr (6.9)	Ind (24.0)	Övr (69.0)
Icke (5.2)	5.2	5.7	5.1
Mod (19.0)	18.0	13.6	21.0
CP (15.6)	51.7	11.7	13.4
FP (10.5)	7.6	8.7	11.4
KDS (1.4)	0	1.2	1.6
SAP (42.8)	16.9	53.9	41.5
VPK (5.1)	0.6	5.0	5.6
Övr (0.4)	0	0.2	0.5

skattningar som redovisas svarar rätt väl mot vad man kan förvänta: Centerpartiets relativt starka stöd bland sysselsatta inom jordbruk framträder klart och SAP:s stöd bland industrisysselsatta kan likaså tydligt avläsas.

Tabell 7. Survey-skattningar (viktade) av blockval efter näringsgren 1979

Block	Andel sysselsatta inom näringsgren		
	Jbr (7.2)	Ind (23.9)	Övr (69.9)
Icke (9.3)	8.9	10.2	9.0
Borg (45.7)	75.7	34.8	46.4
Soc (45.0)	15.5	55.1	44.6

Tabell 8. Survey-skattningar (icke-viktade) av blockval efter näringsgren 1979

Block	Andel sysselsatta inom näringsgren		
	Jbr (6.9)	Ind (24.0)	Övr (69.0)
Icke (5.2)	5.2	5.7	5.1
Borg (46.5)	77.3	35.2	47.4
Soc (48.3)	17.5	59.1	47.6

Ekologdata-skattningar

De ekologiska inferenser som görs för valet 1979 baseras endast på en aggregeringsnivå, den nuvarande kommunindelningen (kommun₂₈₃). Analysen indelas i två moment, så att först görs skattningar utifrån partival för att därefter bygga på blockval. Förutom att pröva viktade respektive oviktade skattningar kommer i likhet med analysen av valet 1944 tre grundmodeller för ekologisk regression att utnyttjas: regression utan intercept, bivariat regression och multipel regression. Det nya med analysen av valet 1979 är att den multipla regressionen kommer att prövas mer systematiskt. Fyra olika modeller kommer att ligga till grund för de ekologiska skattningarna; en första med historisk tradition som prediktor; en andra där också politisk konkurrens ingår som en prediktorvariabel; den tredje modellen tar in olika omgivningsfaktorer som är uttryck för religiös orientering, tätortsgrad, inkomster med mera för respektive kommuner. För att erhålla rimliga

mått som uttryck för dessa omgivningsfaktorer har de underkastats en faktoranalys som genererat tre faktorer som kan benämnas på följande sätt: höga inkomster, urbanisering och arbetslöshet (jfr Kelley & McAllister, 1983). I den tredje modellen läggs dessa tre faktorer – mätta med faktor-scores – till som prediktorvariabler. I den fjärde och sista av de multipla regressionsmodellerna modifieras de olika modellerna så att endast variabler vars parametrar är signifikanta i den tredje modellen beaktas vid skattningen. Med detta som bakgrund kan vi specificera de olika modeller som skall ligga till grund för våra ekologiska skattningar av proportioner sysselsatta inom olika näringsgrenar som röstar på olika partier:

$$\text{PARTI79} = b_1\text{JBR80} + b_2\text{IND80} + b_3\text{ÖVR80} + e \quad [10]$$

$$\text{PARTI79} = a + b_1\text{NÄRING80 (JBR, IND, ÖVR)} + e \quad [11]$$

$$\text{PARTI79} = a + b_1\text{NÄRING80} + b_2\text{HIST} + e \quad [12]$$

$$\text{PARTI79} = a + b_1\text{NÄRING80} + b_2\text{HIST} + b_3\text{POL} + e \quad [13]$$

$$\text{PARTI79} = a + b_1\text{NÄRING80} + b_2\text{HIST} + b_3\text{POL} + b_4\text{FAKTOR1} + b_5\text{FAKTOR2} + b_6\text{FAKTOR3} + e \quad [14]$$

Den fjärde varianten av multipel regression som talas om ovan är alltså en modifikation av modell [14] som vi väljer att kalla modell [15]. Det dataunderlag som utnyttjas finns närmare beskrivet i Wörlund & Ersson (1987b), men det kan ändå påpekas att indelningen i näringsgrenar gjorts så att den skall så nära som möjligt svara mot den som utnyttjats vid survey-skattningarna. Tilläggas bör också att val av historisk traditionsfaktor och politisk konkurrensvariabel gjorts med beaktande av vilket parti som skattningen avser. Vid redovisningen av de skattningar som gäller modellerna [12] till [15] lämnas även information rörande andel förklarad varians (R^2A) respektive ett mått på multikollinearitet för den variabel vars proportion skattas (R^2D). Skattningar som gjorts avseende röstande på SAP presenteras nedan i tabell 9.

Tabell 9. Skattningar av proportioner sysselsatta inom olika näringsgrenar röstande på SAP 1979: grundmodell [10] till [15]

Modell	Proportion jordbruk		Proportion industri		Proportion övriga	
	Viktad	Oviktad	Viktad	Oviktad	Viktad	Oviktad
[10]	.067	-.013	.675	.619	.312	.358
[11]	.229	-.024	.647	.618	.328	.348
[12]	.496	.362	.574	.538	.332	.331
[13]	.159	.136	.444	.474	.395	.390
[14]	.226	.251	.452	.456	.371	.369
[15]	.151	.261	.453	.450	.371	.375

Som framgår av tabellen finns skillnader mellan olika skattningar, men vanligen är skillnaderna förhållandevis små mellan viktade och oviktade värden, medan framför allt i skattningen av proportionen sysselsatta inom jordbruk skillnaderna mellan de olika modellerna är betydande. Ser man till de viktade skattningarna är samtliga giltiga, men uppenbart är att modell [12] leder till en överskattning av denna proportion; genom att beakta olika omgivningsfaktorer blir den skattade proportionen rimligare samtidigt som man i modell [14] kan finna ett högt R²D för jordbruksvariabeln. Det är också vid skattningen av jordbruksvariabeln som man genomgående stöter på problemet med multikolliniaritet i vissa modeller ([14] och [15]). En hög multikolliniaritet behöver dock inte leda till sämre skattningar; snarare tycks det vara så att en hög förklarad varians bäst indikerar den rimligaste skattningen. I de flesta fall är det så att den multipla regression som visar den högsta förklarade variansen är den modell som modifierats (modell [15]).

För flertalet partier är det skattningen av proportionen jordbrukssysselsatta som ställer till med problem och det är egentligen bara för CP och ICKE som skillnaderna i skattning är måttlig. Däremot är skattningarna av de två övriga proportionerna mindre varierande och det är endast för VPK och ÖVRIGA som skattningen av de industrisysselsatta medför problem.

Analysen av blockval ansluter mycket nära till vad som gällt vid analysen av partival. I stället för enskilda partier som beroende variabler provas nu de politiska blocken – socialistisk, borgerlig samt icke-röstande – som beroende variabler. Specificeringen av de multipla regressionerna måste också ändras såtillvida att en modell med politisk konkurrens som prediktor inte är meningsfull vid blockval. Med dessa skillnader i minnet kan de modeller som är aktuella för analysen specificeras enligt följande:

$$\text{BLOCK79} = b_1\text{JBR80} + b_2\text{IND80} + b_3\text{ÖVR80} + e \quad [16]$$

$$\text{BLOCK79} = a + b_1\text{NÄRING80} (\text{JBR, IND, ÖVR}) + e \quad [17]$$

$$\text{BLOCK79} = a + b_1\text{NÄRING80} + b_2\text{HIST} + e \quad [18]$$

$$\text{BLOCK79} = a + b_1\text{NÄRING80} + b_2\text{HIST} + b_3\text{FAKTOR1} + b_4\text{FAKTOR2} + b_5\text{FAKTOR3} + e \quad [19]$$

Med en modifiering av modell [19] erhålls den tredje varianten av en multipel regression av blockval, en variant som vi benämner modell [20]. I tabell 10 nedan presenteras de skattningar som avser de som vid valet 1979 röstar på socialistpartierna.

Tabell 10. Skattningar av proportioner sysselsatta inom olika näringsgrenar röstande på socialistpartierna 1979: grundmodell [16] och [20]

Modell	Proportion jordbruk		Proportion industri		Proportion övriga	
	Viktad	Oviktad	Viktad	Oviktad	Viktad	Oviktad
[16]	-.073	-.081	.687	.639	.406	.439
[17]	.053	-.091	.632	.639	.426	.427
[18]	.626	.501	.572	.533	.502	.392
[19]	.414	.322	.479	.459	.437	.441
[20]	.373	.312	.473	.457	.447	.442

Även här kan konstateras att skillnader i viktningförloppet endast marginellt påverkar skattningarna medan variationer i modellutformning har en större betydelse. Dels finns skillnader mellan de tre grundmodellerna, dels finns skillnader mellan multipla regressionsmodeller med olika modellutformning. Speciellt markerat är detta vad gäller skattningen av proportionen sysselsatta inom jordbruk som röstar på socialistpartier. Regressionen utan intercept [16] leder till icke giltiga proportioner (dvs underskattning) medan den multipla regressionen med historisk tradition och näringsstruktur som prediktorvariabler [18] starkt överskattar samma proportion. Skattningarna av proportionerna sysselsatta inom industri och övriga näringar är däremot stabilare oberoende av om det gäller röstande på socialistpartier, borgerliga partier eller icke röstande.

Jämförelser mellan surveydata- och ekologdata-skattningar

Efter att ha redovisat skattningar av hur grupper av individer med olika näringsgrenstillhörighet röstat på olika partier eller partiblock återstår att försöka utvärdera vilka ekologdata-skattningar som närmast överensstämmer med skattningarna baserade på surveydata. Vi antar alltså att surveydata-skattningarna ligger närmast de "sanna" proportionerna. Vilka ekologdata-skattningar leder till de "bästa" skattningarna? Vi redovisar först utfallet av skattningarna av partival och vi använder oss därvid av de två "likhets"-mått vi tidigare beskrivit; beträffande dessa mått kan sägas att ju lägre värde de visar, desto större likhet mellan jämförelseobjekten. Jämförelseobjekten för partivalen är tabellerna 5 och 6. Utfallet av jämförelsen mellan tabellerna redovisas i tabell 11 nedan.

De två måtten – T och NS – överensstämmer sinsemellan nästan fullt ut. Jämförs som exempel viktade och oviktade surveydata är likheten dem emellan .050 respektive .048. Därmed vet man att värden i storleksordningen .05 innebär en hög grad av överensstämmelse. Nu förhåller det sig dock så att dessa värden är beroende av antalet celler i de tabeller som jämförs; med fler celler tenderar överensstämmelsen att bli mindre varav följer att värdena blir högre. Detta kan illustreras med motsvarande beräkningar av överensstämmelsen mellan survey- och ekolog-skattningar som redovisats av Thomsen (1987a:32). Därvid har vi dels utgått från Thomsens ursprungliga 66

Tabell 11. Jämförelser mellan surveydata-skattningar och ekologdata-skattningar av partival 1979: grad av överensstämmelse

	Survey-skattningar			
	Viktade		Oviktade	
	T	NS	T	NS
Survey-skattningar				
Viktade			.050	.048
Oviktade	.050	.048		
Ekolog-skattningar				
Modell [10] Viktad	.276	.304	.276	.315
Modell [10] Oviktad	.220	.236	.223	.248
Modell [11] Viktad	.285	.308	.281	.305
Modell [11] Oviktad	.219	.234	.223	.246
Modell [12] Viktad	.261	.273	.256	.268
Modell [12] Oviktad	.182	.168	.184	.178
Modell [13] Viktad	.238	.202	.246	.227
Modell [13] Oviktad	.219	.166	.228	.193
Modell [14] Viktad	.119	.130	.128	.147
Modell [14] Oviktad	.168	.159	.172	.176
Modell [15] Viktad	.103	.112	.117	.139
Modell [15] Oviktad	.161	.149	.170	.174

celler, dels grupperat om tabellerna så att vi erhållit 24 celler. Vi har då erhållit följande värden: .142 och .155 (T) respektive .149 och .164 (NS) för tabeller med 66 celler medan motsvarande värden för tabeller med 24 celler var .079 och .075 (T) respektive .086 och .081 (NS). Således bör man kunna säga att värden som ligger i närheten av .100 representerar en acceptabel grad av överensstämmelse mellan olika skattningar när man jämför tabeller med 24 celler. Med denna norm som bakgrund kan konstateras att flertalet ekolog-skattningar har värden kring .200 eller däröver. Inte någon av de skattningar som baseras på regression utan intercept [10] eller bivariat regression [11] ger acceptabla värden. Det bästa utfallet får vi för de modeller som beaktar olika omgivningsfaktorer [14] eller modifieringar därav [15] och utfallet är bättre för de viktade än för de icke viktade varianterna. Kännetecknande för dessa modeller är dels att de vanligen uppvisar den högsta förklarade variansen, dels att de ingående variabelernas parametrar är signifikanta. Däremot är det inte alltid nödvändigt att multikolliniariteten i modellen måste vara låg. Trots att de skattningar som görs utifrån dessa ekologiska regressionsmodeller är acceptabla leder de ändå i några fall till klara över- respektive underskattningar. Så överskattas proportionen sysselsatta inom jordbruk som röstar för VPK medan

samma proportion underskattas för CP; likaså överskattas proportionen sysselsatta inom industri för CP och MOD medan denna proportion underskattas för VPK och SAP.

En motsvarande jämförelse mellan survey- och ekolog-skattningar avseende blockval redovisas i tabell 12; i denna tabell är det tabellerna 7 och 8 som tjänar som jämförelsepunkter. Eftersom det är frågan om blockval minskar antalet celler i tabellerna från 24 till 9.

Tabell 12. Jämförelser mellan surveydata-skattningar och ekologdata-skattningar av blockval 1979: grad av överensstämmelse

	Survey-skattningar			
	Viktade		Oviktade	
	T	NS	T	NS
Survey-skattningar				
Viktade			.037	.040
Oviktade	.037	.040		
Ekolog-skattningar				
Modell [16] Viktad	.143	.137	.143	.140
Modell [16] Oviktad	.129	.111	.129	.118
Modell [17] Viktad	.078	.077	.075	.073
Modell [17] Oviktad	.133	.117	.134	.122
Modell [18] Viktad	.199	.128	.191	.128
Modell [18] Oviktad	.156	.107	.154	.127
Modell [19] Viktad	.166	.135	.168	.152
Modell [19] Oviktad	.141	.130	.147	.147
Modell [20] Viktad	.158	.133	.162	.150
Modell [20] Oviktad	.138	.124	.143	.142

Med ett mindre antal celler i jämförelsetabeller tenderar värdena att minska; så är nu likheten mellan tabell 7 och 8 .037 (T) och .040 (NS) gentemot motsvarande för tabell 5 och 6 (.050 respektive .048). De lägre "likhets"-värdena är alltså en följd av det minskade antalet celler, och man kan inte utan vidare hävda att skattningarna av blockvalet är bättre än skattningarna av partivalen. Dessa värden säger endast att vissa skattningar överensstämmer bättre med survey-skattningar än vad som är fallet med andra skattningar. Andra skillnader kan också noteras i jämförelse med tabell 11. Den bästa skattningen uppvisar den viktade bivariata regressionen, medan både regression utan intercept och olika varianter av multipel regression leder till sämre skattningar. Vad gäller den multipla regressionen leder dock successiva modifieringar av modellerna till förbättrade skattningar. Tydligt är ändå att ekologiska inferenser som bygger på blockval i stället för partival inte utan vidare leder till bättre skattningar.

Avslutning

Vi har haft ett flertal syften med vår uppsats där det gemensamma temat har varit att försöka reda ut hur man på bästa sätt skall uppnå rimliga och acceptabla ekologiska inferenser. Vi har därvid berört följande problem: aggregeringsnivå, viktningförfarande, typ av regression och modellspecifikation. Vi har inte kommit fram till entydiga svar i alla avseenden, men den redovisade analysen ger antydningar om var svaren kan vara att finna.

Ett resultat är att viktningförfarandet inte har så stor betydelse för de skattningar som görs. Vanligen leder viktade skattningar till rimligare resultat, men skillnaderna mellan olika viktningförfaranden är många gånger marginell. Inte heller tycks val av aggregeringsnivå vara något stort problem. Analyserna av valet 1944 pekar i varje fall mot att skillnader mellan skattningar som baseras på databärande enheter omfattande ett $N=283$ eller ett $N=2479$ är förhållandevis små. Om man däremot minskar antalet databärande enheter till $N=24$ som gäller för länsnivå erhåller man avvikande skattningar. Det märkliga är att med till synes underspecificerade modeller kan man på denna aggregeringsnivå erhålla rimliga skattningar. Detta till trots är vi tveksam till användningen av län som databärande enhet, eftersom den variation som finns inom dessa enheter överstiger den variation som finns mellan enheterna.

I andra frågor är våra slutsatser mindre säkra. Val av typ av regression har konsekvenser för skattningarna. Ofta finns skäl anta att regressioner utan intercept eller bivariata regressioner är underspecificerade varför vissa efterfrågade proportioner tenderar att överskattas. Dock tycks det inte alltid förhålla sig så. Vid blockval där man arbetar med stora proportioner tycks en bivariat regressionsmodell kunna leda till en skattning som är väl så rimlig som den som erhålls med en multipel regression. Vid partival förefaller emellertid den multipla regressionen leda till de bästa skattningarna. Det stora problemet som är förknippat med en multipel regression är huruvida en modell är väl specificerad eller ej. En ledtråd därvid utgör den förklarade variansen såtillvida att ju högre förklarad varians, desto rimligare tycks de skattade proportionerna bli. Problemet är bara att man inte heller med en hög förklarad varians kan vara säker på att den skattade proportionen blir rimlig. Ibland kan multikollinariitet inverka, men som något stort pro-

blem upplever vi inte att kollinariiteten är. Viktigast är att få en modell rätt specificerad är en ingående kännedom om den underliggande teoretiska modellen och kunskap om indikatorer på aggregatnivå som kan fånga upp de teoretiskt relevanta storheterna. Vad vi vill ha sagt är att man inte automatiskt kan komma fram till något mått som ger uttryck för hur väl specificerad en modell är utan att man måste ta ställning till detta i det enskilda fallet.

En av utgångspunkterna i uppsatsen har varit att man genom väl specificerade regressionsmodeller skäll kunna göra rimliga och acceptabla ekologiska inferenser. Till viss del har vi visat att så kan vara fallet. Med en bredare databas, bättre individuellt anpassade modeller för enskilda partier eller partiblock är det sannolikt att vi kunnat erhålla ännu bättre skattningar som legat närmare de survey-baserade skattningarna. Andå återstår problem som är förknippade med den multipla regressionsanalysen. Ett sådant är att de skattade proportionerna nästan aldrig summerar till ett och ibland avviker ganska mycket från ett. Det är inte heller alltid säkert att en linjär regressionsmodell svarar mot den bästa anpassningen till data. Därför kan det vara på sin plats att förutom arbetet med att förbättra modellspecifikationerna utnyttja andra alternativ inom regressionsanalysen. Så kan det vara lämpligt att använda sig av restriktioner eller justeringar så att skattade proportioner alltid summerar till ett. Ett annat alternativ är att pröva regressionsanalysen på ett mindre antal sinsemellan mer homogena regioner från vilka man kan skatta proportioner som avser Sverige som helhet. Först när man prövat ett stort antal varianter av regressionsanalys som även bygger på flera val och med en mer omfattande databas kan man mer definitivt uttala sig om möjligheten att göra ekologiska inferenser. Så här långt tycker vi oss ha kunnat visa att ekologisk regression under vissa omständigheter ger underlag för ekologiska inferenser som kan sägas vara såväl rimliga som acceptabla i jämförelse med motsvarande inferenser byggda på survey-data.

Noter

¹ Vi har genom tillmötesgående från Sören Holmberg kunnat utnyttja data från 1979 års väljarundersökning. Varken han eller andra inblandade i väljarundersökningen har något ansvar för de analyser och tolkningar vi i denna uppsats gör av nämnda data.

Referenser

- Alker, H A (1969) "A typology of ecological fallacies" s 69–86 i Dogan M & Rokkan, S (red) *Quantitative Ecological Analysis in the Social Sciences*. Cambridge, MA: MIT U. P.
- Brown, C (1982) "The nazi vote: a national ecological study." *American Political Science Review*, 76, s 285–302.
- Burstein, L (1978) "Assessing differences between grouped and individual-level regression coefficients: alternative approaches." *Sociological Methods & Research*, 7, s 5–28.
- Crewe, I & Payne, C (1976) "Another game with nature: an ecological regression model of the British two-party vote ratio in 1970." *British Journal of Political Science*, 6, s 43–81.
- Deegan, J (1976) "The consequences of model misspecification in regression analysis." *Multivariate Behavioral Research*, 11, s 237–248.
- Feisenmaier, DR (1985) "On modelling aggregation impact in spatially distributed data." *Quality and Quantity*, 19, s 71–82.
- Firebaugh, G (1978) "A rule for inferring individual-level relationships from aggregated data." *American Sociological Review*, 43, s 557–572.
- Flanigan, WH & Zingale, NH (1981) "Alchemist's gold: inferring individual relationships from aggregate data." Paper presented to the ECPR joint sessions of workshop at Lancaster.
- Goodman, L A (1953) "Ecological regressions and behavior of individuals." *American Sociological Review*, 18, s 663–664.
- Goodman, L (1959) "Some alternatives to ecological correlation." *American Journal of Sociology*, 64, s 610–625.
- Hammond, J L (1973) "Two sources of error in ecological correlations." *American Sociological Review*, 38, 764–777.
- Hannan, M T (1971) *Aggregation and Disaggregation in Sociology*. Lexington, MA: Lexington Books.
- Hannan, M T & Burstein, L (1974) "Estimation from grouped observations." *American Sociological Review*, 39, 374–392.
- Hanushek, E A, et al. (1974) "Model specification, use of aggregate data, and the ecological correlation fallacy." *Political Methodology*, 1, s 89–107.
- Hanushek, E A & Jackson, J (1977) *Statistical Methods for Social Scientists*. New York: Academic Press.
- Holmberg, S & Nordlöf, H (1982) *Valundersökning 1979: teknisk rapport*. Stockholm: SCB.
- Ingulfson, H & Hagman, R. (1950) "De svenska partiernas sociala och andra grundvalar" s 123–181 i Håstad, E et al. "Gallup" och den svenska väljarkåren. Stockholm: Gebers.
- Jones, E T (1972) "Ecological inference and electoral analysis." *Journal of Interdisciplinary History*, 2, s 249–262.
- Kelley, J & McAllister, I (1983) "The methodology of aggregate analysis: errors in traditional procedures

- and suggestions for improvement." *Quality and Quantity*, 17, s 461–474.
- King, G (1986) "How not to lie with statistics: avoiding common mistakes in quantitative political science." *American Journal of Political Science*, 30, s 666–687.
- Klevmarcken, A (1970) "Olikhetskoefficienter: några deskriptiva mått för utfallsanalys av prognoser." *Statistisk tidskrift*, 8, s 261–273.
- Langbein, LI & Lichtman, AJ (1978) *Ecological Inference*. Beverly Hills: Sage.
- Lemieux, P (1976) "Heteroskedasticity and causal inference in political research." *Political Methodology*, 3, s 287–316.
- Lewin, L et al. (1972) *The Swedish Electorate 1887–1968*. Stockholm: Almqvist & Wiksell.
- Lichtman, AJ & Langbein, LI (1978) "Ecological regression versus homogenous units: a specification analysis." *Social Science History*, 2, s 172–194.
- Lohmöller, J-B & Falter, JW (1986) "Some further aspects of ecological regression analysis." *Quality and Quantity*, 20, s 109–125.
- Robinson, WS (1950) "Ecological correlations and the behavior of individuals." *American Sociological Review*, 15, s 351–357.
- Schadee, HMA (1987) "A regression model for the disaggregation of aggregate (electoral) data." Paper presented to the ECPR joint sessions of workshops at Amsterdam.
- Selle, P (1983) *Norges Kommunistiske Parti 1945–1950*. Bergen: Universitetsforlaget.
- Stokes, DE (1969) "Cross-level inference as a game against nature", s 62–83 i Bernd, JL (red) *Mathematical Applications in Political Science: IV*. Charlottesville: The University Press of Virginia.
- Thomsen, SR (1987a) "A latent structure approach to ecological inference." Paper presented to the ECPR joint sessions of workshop at Amsterdam.
- Thomsen, SR (1987b) *Danish Elections 1920–79: A Logit Approach to Ecological Analysis and Inference*. Århus: Politica.
- Wörlund, I & Ersson, S (1987a) "Valdataarkiv 1911–1985: Kodbok 1: Kommunkoder och Kartor." Umeå universitet: Statsvetenskapliga institutionen.
- Wörlund, I & Ersson, S (1987b) "Valdataarkiv 1911–1985: Kodbok 2: Variabelförteckning över politiska och socioekonomiska data." Umeå universitet: Statsvetenskapliga institutionen.