

# 1 Introduksjon og sammendrag

Bladcentralen Ans (BC) distribuerer blader og bøker til ca. 8.000 forhandlere over hele landet på vegne av 7 eierforlag. BC distribuerer sine publikasjoner i kommisjon. Dette innebærer bl.a. at BC fastsetter leveransen på den enkelte tittel til hver forhandler hver uke. Hver uke beregnes ca. 350.000 leveranser. Forhandlere får på sin side kreditert alle publikasjoner som ikke selges. Returen representerer en betydelig verdi. For hvert prosent poeng returen kan reduseres uten at antall utsolgt-situasjoner øker, reduseres kostnadene betydelig for BC's eierforlag. Samtidig følger reduserte miljøkostnader.

Målet med leveranse-beregningene er å treffe etterspørselen best mulig slik at salget blir størst mulig med minst mulig trykkopplag. Kvaliteten i tildelingen blir normalt målt ved sammenhengen mellom utsolgt- og retur-prosent på den enkelte bladtype. I dag brukes ulike modeller i BC's leveranseberegning, den mest avanserte er utviklet av Petter Matthiessen A/S i Danmark.

Vi vil her beskrive og studere en ny metodikk for å finne en optimal leveranse av blader til utsalgssteder slik at kostnader i forbindelse med retur og utsolgt blir mindre enn for eksisterende metode. Dette innebærer å finne en god statistisk modell for potensielt salg.

Foreslått leveranse skjer i tre steg. Først predikeres potensielt salg for hvert salgssted basert på historiske data. Deretter bestemmes leveranse utfra den statistiske fordelingen til potensielt salg slik at retur-kostnader og utsolgt-kostnader balanseres. Til slutt gjøres en skalering av leveranse for hvert sted slik at summen av leveranse over alle steder er lik angitt totalopplag som Bladcentral bestemmer. Den nye metodikken er testet ut på historiske data fra 200 salgssteder og tre bladtyper.

Resultatene til Norsk Regnesentral (NR) er som følger: For samme utsolgt-volum som BC får vi reduksjon i returandel på 0.8, 2.0 og 2.4 % for tre bladtyper henholdsvis. En viktig kommentar til disse resultatene er at BC følger opp noen kundegrupper i større grad enn andre ved å ha leveranse to ganger i uken når dette påkreves, men selv for disse kundegrupper greier NR's metode å forbedre BC's metode.

## 2 Data

I dette forprosjektet har vi tilgjengelig salgsdata for tre bladtyper (kalt 800, 801 og 802) fra 200 salgssteder. For bladtype 800 og 802 er de historiske data fra uke 1 i 1995 til uke 25 i 1998. I noen av ukene er det ikke salg p.g.a. høytider. Tilsammen gir dette 174 uker med salgsdata for 200 salgssteder. Fra

bladtype 801 har vi data fra uke 1 i 1995 til uke 22 i 1998. Bladtype 801 har også 174 uker med salg. Bladtype 800 har størst salg fra de historiske data med total salg på 368665. Dernest har bladtype 801 totalt salg på 150099. Minst salg har bladtype 802 med totalt salg 69522. I tillegg til salgsdata har vi også tilgjengelig virkelige leveransedata til de forskjellige steder hver uke, der leveranse er basert på BC's metode. Bladcentralen deler sine salgssteder inn i 5 grupper: !, \*, A, B og C, der ! er de 200 største, \* er de 700 påfølgende, A er de 1500 neste, B er de 2000 deretter og C er de minste, ca 3500. I det foreliggende datamaterialet er det få salgssteder i gruppe ! og \*. Vi vil derfor behandle disse to gruppene som en gruppe. Videre vil vi kun bruke data fra salgssteder som har registrert salgsdata for minst 173 uker, se Tabell 1.

### 3 Metodikk

I dette kapitlet beskriver vi den nye metodikken. Vi vil kun gi en kort beskrivelse av metoden uten å gå veien om stringente matematiske formuleringer. For å finne den matematiske formulering av problemet vil vi referere til andre rapporter og artikler. Hoved-ideen i metoden kan deles opp i følgende tre punkter:

1. Lage modell for forventet potensielt salg, der potensielt salg er det virkelige salget vi ville kunne observert dersom det ikke var utsolgt noen steder.
2. Utifra en fordeling rundt forventet potensielt salg, finne en optimal leveranse for en "normal" uke, der kostnader i forbindelse med retur og utsalg er balansert.
3. Skalering av leveransene ut fra BC's vurdering av hva totalopplaget bør være.

Punkt 1. over kan gjøres på flere forskjellige måter. Nedenfor vil vi gjøre rede for to mulige måter å modellere forventet potensielt salg. Merk at vi bruker begrepet potensielt salg. Vi definerer potensielt salg som hvor mange eksemplarer som ville blitt solgt dersom ingen steder ble utsolgt. Det er en sammenheng mellom virkelig salg og potensielt salg: Potensielt salg=virkelig salg + antall eksemplarer som kunne blitt solgt dersom man hadde gitt stor nok leveranse. Se (Aldrin 1995a) for lignende type data og metodikk. Det er kun mulig å registrere data for virkelig salg, men det vi ønsker å modellere er potensielt salg. Hvis man ser bort fra forskjellen mellom potensielt salg og virkelig salg vil man underestimere potensielt salg. Det er derfor svært viktig å korrigere for dette, men i dette forprosjektet har vi sett bort fra denne problematikken. Analysen er gjort i programpakken S-plus, se (Venables & Ripley 1994).

### 3.1 Potensielt salg modellert ved RRR/GAM

Salgstallene varierer over tid. Endel av denne variasjonen er systematisk, og kan predikeres frem i tid. Den systematiske variasjonen består typisk av en trend, en glatt sesongvariasjon over året, samt et spesielt stort eller lite salg i visse uker. Videre vil de ulike effekter typisk inngå multiplikatvt. Det vil f.eks. si at påskenummeret selger ca. 50% mer ved både store og små salgssteder. Den systematiske salgsmønsteret vil i stor grad være ganske likt for mange salgssteder. Det er derfor mulig å estimere salsmodeller simultant (multivariabelt) for flere salgssteder samtidig. Når en salgsmoell for et salgssted estimeres låner man styrke fra salgssteder med lignende salgsmønster. Til å forklare den systematiske variasjon bruker vi følgende 25 forklaringsvariable:

- Trend med lineært og kvadratisk ledd.
- Sesong variasjon over året bestående av 12 forklaringsvariable: 6 par sinus- og cosinus-funksjoner med periode på 1,1/2,1/3,1/4,1/5,1/6 år.
- To kryss-ledd mellom lineær trend og sinus- og cosinus med periode 1 år.
- Det er 9 spesielle uker indikert som indikator variable. Sommeruker: 1) uke 28, 2) uke 29, 3) uke 30, 4) uke 31, 5) førpåskeuke, 6) påskeuke, 7) førjul, 8) jul og 9) romjul.

De 25 forklarings-variable som er spesifisert over brukes til å beskrive den funksjonelle sammenhengen mellom forklaringsvariable og potensielt salg. Dette er mange forklarings variable i forhold til datamengde. Det betyr at estimering av en modell separat for hvert salgssted vil gi ustabile estimater. Vi behandler i stedet alle salgsstedene i en kundegruppe (en av de 4 beskrevet i kapittel 2) simultant. Dette gjøres for hver publikasjon separat. Videre gjør vi en såkalt redusert rang regresjon (RRR) for å lage nye, men langt færre forklaringsvariable for hver kundegruppe og hver publikasjon. Disse nye forklaringsvariablene kaller vi basiskurver og de er lineærkombinasjoner av de gamle forklaringsvariable, se (Aldrin 1995*b*). Redusert rang regresjon prøver å hjelpe steder med lite data ved å "låne" styrke fra andre steder. Egenskapene til disse basiskurvne er slik at de første forklarer mesteparten av variasjonen. Vi brukte i våre analyser 2 basiskurver som er langt færre enn 25 forklaringsvariable. Antall basiskurver er avhengig av hvor lang historisk informasjon som er tilgjengelig. Når kort historikk er tilgjengelig bør det brukes få basiskurver og ettersom mer historisk informasjon er tilgjengelig vil det være gunstig å bruke flere basiskurver.

Vi har så langt ikke tatt hensyn til at salgsdata er heltallige. Dette tar vi hensyn til ved å gjøre en Poissonregresjon for hvert salgssted. Vi bruker såkalt log-linkfunksjon, som gir oss en multiplikatv modell. Vi bruker her 2 basiskurver som forklaringsvariable, og disse ivaretar systematisk struktur som

er felles på tvers av salgssteder. I tillegg tar vi med en fleksibel, med glatt trendfunksjon. Denne metoden går under fellesbetegnelsen GAM (generaliserte additive modeller). Den glatte trendfunksjonen er en såkalt “smoothing spline” der vi har brukt 4 parametere. Vi kaller metoden RRR/GAM.

## 3.2 Potensielt salg modellert ved nevralt nett

Nevrale nett er en fleksibel, ikke-lineær regresjonsmetode. Vi trenger ikke i samme grad som i forrige kapittel å spesifisere detaljert sammenhengen mellom responsvariable (salgsdata) og forklaringsvariable som definerer hovedstrukturen i problemet. Som i forrige kapittel behandler vi publikasjoner hver for seg, men analyserer alle salgssteder innen hver kundegruppe simultant. Vi betrakter logaritmen til salget ved de ulike salgssteder som en multivariabel responsvariabel (logaritmen gjør at modellen blir multiplikativ på original skala). Vi bruker følgende 12 forklaringsvariable:

- Trend.
- Sesong variasjon over året bestående av 2 forklaringsvariable: 1 par sinus- og cosinus-funksjoner med periode på 1 år.
- Det er 9 spesielle uker indikert som indikator variable. Sommeruker: 1) uke 28, 2) uke 29, 3) uke 30, 4) uke 31, 5) førpåskeuke, 6) påskeuke, 7) førjul, 8) jul og 9) romjul.

Vi har brukt et såkalt feed forward nettverk med ett skjult lag, se (Ripley 1994). Valget av såkalte noder i det skjulte lag (“hidden layer”) av metoden bestemmer antall parametere. Mange noder gir en god tilpasning til data, men også mange parametere å estimere. Antall noder spiller noe av samme rolle som antall basiskurver i RRR/GAM. En parameter, “weight decay” straffer bruk av mange mange noder. Det er her en overhengende fare for overtilpassning med for mange noder og for liten “weight decay”. Etter et forstudie valgte vi 6 noder og “weight decay” lik 0.01. Metoden vil bli betegnet som NN, der NN står for nevralt nett. Ved bruk av nevralt nett er det viktig å “standardisere” responsvariabelen. Vi valgte å lage ny responsvariabel der vi tok den gamle å dividerte på  $3 \times \text{standardavvik}$ .

## 3.3 Sammenligning av RRR/GAM og NN

Figur 1 viser forventet salg for RRR/GAM metode og NN metode respektivt. To steder fra Gruppe A og bladtype 800 med salg (heltrukken linje) er vist over en periode fra uke 1 1995 til uke 25 1998. Fra uke 1 1995 til og med uke 24 1997 er de stiplede kurvene estimert salg ut fra alle data i denne perioden. Fra uke 25 1997 starter prediksjon angitt med vertikal stiplet linje. Forventet salg er predikert ved hjelp av historiske data som er minst 4 uker gamle. Vi ser ut

fra figuren at forventet salg tilpasser seg bra til salg i begge av NR's metoder i estimeringsperioden. Sesongsvingninger fås med på en tilfredstillende måte. I prediksjons-perioden derimot ser vi at tilpasningen er dårligere. Det er vanskelig ut fra kun disse plottene å fastslå hvilken metode som fungerer best. Vi kommer tilbake til dette i kapittel 4.

### 3.4 Beregning av leveranse for “normal” uke

Over har vi beskrevet metoder for å beregne forventet potensielt salg. Nå er imidlertid ikke bestemmelse av forventet potensielt salg vår primære oppgave, men bestemmelse av leveranse. Utfra forventet potensielt salg er det mulig å bestemme leveranse ved å anta en sannsynlighetsfordeling for potensielt salg. Sannsynlighetsfordelingen kan bestemmes på ulike måter, men det mest nærliggende er å anta en Poisson fordeling (diskret fordeling) omkring forventet potensielt salg. En av egenskapene til Poisson fordelingen er at variansen for er den samme som forventningsverdien. En kunne tenke seg en mer generell fordeling som tar hensyn til at variansen ikke er den samme som forventningen. Se konklusjon for forslag om forbedringer. Anta nå vi bruker en Poisson fordeling rundt forventet potensielt salg. For å finne leveransen vil vi bruke sannsynlighetsfordelingen rundt forventet potensielt salg på følgende måte: Bestem “normal” leveranse slik at et gitt forhold ( $f = k_1/k_2$ ) mellom retur-kostnader ( $k_1$ ) og utsolgt-kostnader ( $k_2$ ) minimeres ved å ta hensyn til sannsynlighet for retur og utsolgt.

### 3.5 Skalering av “normal” leveranse

En “normal” leveranse sier noe om hva total leveransen bør være for en normal uke. Dersom BC mener at total leveranse bør nedjusteres eller oppjusteres av ulike grunner foretas det en skalering. Denne justeringen av leveranse foretas multiplikativt.

## 4 Test av metodikk på data

For å finne ut hvor godt den nye metoden fungerer i forhold til BC's eksisterende metode vil vi gjennomføre et simulerings-eksperiment basert på tilgjengelige data. Vi er interessert i å måle utsolgtandel og returvolum i fremtiden for den nye metoden. Dette kan vi ikke gjøre, men vi kan prøve ut metoden på de historiske data og sammenligne med BC's eksisterende metode. Eksperimentet vil bli utført på følgende måte:

1. Vi bruker salgs-data fra uke 1 1995 til uke 25 1997 til å lage en modell for salg, som brukes til å predikere forventet salg 4 uker i fremtiden (uke 29 1997). Deretter oppdateres modellen ved å bruke salgs-data fra uke 1 1995 til uke

26 1997 til å predikere 4 uker frem i tid (uke 30 1997). Denne oppdateringen fortsetter helt frem til uke 25 1998, slik at vi har predikert nytt forventet salg med ny metode i ett helt år. Vi har nå oppnådd å simulere predikert salg for ett helt år.

2. Neste steg er å finne optimal leveranse for en “normal dag”, slik at kostnader i forbindelse med retur og utsolgt er balansert. Dette gjøres som angitt i kapittel 3.4 for ulike verdier av styringsparameteren  $f$ .

3. Siste steg består i å skalere leveransene slik at total leveranse for hver uke ligger omtrent på samme nivå som det BC har.

Punktene 1, 2 og 3 gjøres for metode RRR/GAM og for NN. For hver verdi av styringsparameteren  $f$  kan vi beregne returandel og utsolgtvolum. Vi velger en  $f$  slik at vi får nøyaktig samme utsolgtvolum som BC, slik at returandelene er sammenlignbare.

## 5 Resultater

En viktig faktor som spiller inn på resultatene for noen av gruppene, spesielt kundegruppe gruppe !,★ er at BC følger dem opp i større grad. Dette vil i praksis bety at man har konsulenter som ringer opp stedene og hører om påfyll trengs. En slik oppfølging medfører ekstra kostnader for BC. Dersom NR's metode gjør det bedre enn BC's metode med oppfølging vil således slik oppfølging kunne kuttes med de besparelser det medfører. I tillegg vil vi ha besparelser forbundet med mindre returandel. I Tabell 1, 2 og 3 finner man bladtype 800, 801 og 802 respektivt. Innen hver bladtype har vi fire gruppetyper, den første er !,★, dernest har vi A-C, se kapittel 2. Tabell 1, 2 og 3 viser returandel (kolonne 3) og utsolgtvolum (kolonne 4) tabellert ved bruk av forskjellige metoder. Kolonne 5 angir forbedring i returandel i forhold til BC's metode med samme utsolgtvolum. Positivt tall betyr her forbedring, mens negativt tall betyr forverring. Nederst i tabellene angis total veid gevinst ved å bruk av RRR/GAM istedet for bruk av BC's metode. Dette tallet er veid i forhold til totalvolum for de forskjellige grupper. Totalt veid gevinst for alle de tre bladtypene målt i returvolum i forhold til BC's metode er på ca. 1.2 prosent. Det må imidlertid tas hensyn til at de tre bladene kanskje ikke er representativt for alle publikasjoner. Trenden i resultatene viser at for små publikasjoner og små salgssteder er det best resultater.

Til nå har vi fokusert på hvor godt NR's fremgangsmåte har fungert i forhold til BC's metode ved å se på returandel og utsolgtvolum for de forskjellige

bladtyper og grupper. En viktig fase i modellbyggingen er å sjekke slike ting som tegn på overtilpasning, korrelasjon i såkalte residualer osv. I Tabell 1, 2 og 3 har vi også tatt med noen mål på hvor godt modellen tilpasser salgsdata i estimeringsperioden og prediksjonsperioden.  $rmsee$  står for “root mean squared error” i estimeringsperiode, mens  $rmsep$  står for “root mean squared error” i prediksjonsperiode. Feilledd for hvert sted ( $s$ ) til tid ( $t$ ) skriver vi som:

$$e_{s,t} = \text{salg}_{s,t} - \text{estimert salg}_{s,t}. \quad (5.1)$$

Vi definerer “root mean squared error” som:

$$rmse = \sqrt{\frac{1}{ST} \sum_{s=1}^S \sum_{t=1}^T e_{s,t}^2}, \quad (5.2)$$

der  $S$  er antall steder og  $T$  er antall uker, se kapittel 2.  $T$  vil være forskjellig for estimerings- og prediksjonsperiode. For å sammenligne de forskjellige tallene har vi skalert ved å dividere på forventet salg (med metode RRR/GAM) der alle data er brukt. I stedet for å tabellere  $rmsep$  har vi tabellert  $rmsep/rmse$  som sier noe om hvor bra prediksjonen er i forhold til estimeringen. Den typiske konklusjon vi kan trekke ut fra tabellene er at metode RRR/GAM estimerer bedre og predikerer bedre enn metode NN. I noen situasjoner er det faktisk slik at NN metoden bommer kraftig på prediksjonen, noe som sier oss at metoden er lite robust.

## 6 Konklusjon

Redusert rang regresjon med 2 basis kurver + Poisson regresjon og med Poisson fordeling rundt forventet salg har vist seg å gjøre det generelt bra i forhold til en del andre metoder vi har prøvd ut. Her er det imidlertid mulig å bruke 3 eller flere basis kurver ettersom mer historisk informasjon er tilgjengelig.

Resultatene fra kapittel 5 forteller oss om reduksjon i returandel på 0.8, 2.0 og 2.4 for tre bladtyper når vi holder veid utsolgtvolum på samme nivå som BC's. Best resultater får vi for små publikasjoner og små salgssteder.

Nevrale nett har vist seg å gi betydelig dårligere resultater enn både metoden over og BC's metode. Mulige forbedringer/utvidelser:

1. I sammenligninger av NR's metode har vi ikke i tilstrekkelig grad tatt hensyn at vi observerer kun potensielt salg i ca. 85% av tilfellene og i de resterende tilfellene har vi utsolgt. I en videreføring må dette tas hensyn til både når vi estimerer og predikerer.
2. Vi har kun sett på Poisson fordeling rundt forventet salg. Denne fordelingen forutsetter den sterke antagelsen om at variansen av salg er lik forventet salg. En mulig forbedring er å bruke en fordeling der også variansen kan modelleres mer fleksibelt.

3. I stedet for å balansere returandel og utsolgtvolum ved å balansere kostnader knyttet til retur og utsolgt, kan man tenke seg å angi akseptabelt utsolgt-andel pr. salgs-sted.
4. Ta med flere spesielle uker slik som: Vinterferie, maidager, høstferie.
5. Utvidelse til publikasjoner der utgivelse ikke er for hver uke.
6. Ta hensyn til at det er korrelasjon mellom bladtyper.
7. En ganske forskjellig fremgangsmåte enn de metoder som er sett på i dette notatet er bruk av dynamiske modeller. Her kan vi også nevne såkalte “Mixed models”.

## Referanser

- Aldrin, M. (1995*a*), Reduced rank regression for multivariate response with missing observations., Technical Report NR-Note STAT/07/1995, Norwegian Computing Center.
- Aldrin, M. (1995*b*), ‘A statistical approach to the modelling of daily car traffic.’, *Traffic Engineering & Control* .
- Ripley, B. D. (1994), ‘Neural networks and related methods for classification’, *Journal of Royal Statistical Society, Series B* **56**, 409–456.
- Venables, W. N. & Ripley, B. D. (1994), *Modern Applied Statistics with S-plus*, Statistics and Computing, Springer Verlag.



bladtype	Gruppe- type	Antall salgssteder	Antall salgssteder med salgsdata for minst 173 uker
800,801,802	!,*	15	15
800,801,802	A	33	32
800,801,802	B	50	49
800	C	102	88
801	C	102	79
802	C	102	77

Tabell 1: Antall salgs-steder med og uten fjerning manglende salgs-data.

Metode	Gruppe- type	Retur- andel i %	Utsolgt- volum i %	Gevinst i retur- andel i forhold til BC's metode i %	rmsee	<u>rmsep</u> rmsee
BC	!,*	17.3	3.9			
RRR/GAM	!,*	17.2	3.9	0.1	1.15	1.36
NN	!,*	19.3	3.9	-2.0	1.13	1.57
BC	A	16.5	1.0			
RRR/GAM	A	16.2	1.0	0.3	0.85	1.38
NN	A	16.9	1.0	-0.4	0.89	1.40
BC	B	18.7	10.2			
RRR/GAM	B	17.2	10.2	1.5	0.82	1.19
NN	B	19.2	10.2	-0.56	0.88	1.40
BC	C	22.1	18.3			
RRR/GAM	C	20.3	18.3	1.8	0.75	0.80
NN	C	20.6	18.3	1.4	0.75	0.73
Total veid gevinst for RRR/GAM metode i retur- andel i forhold til BC's metode i %				0.8		

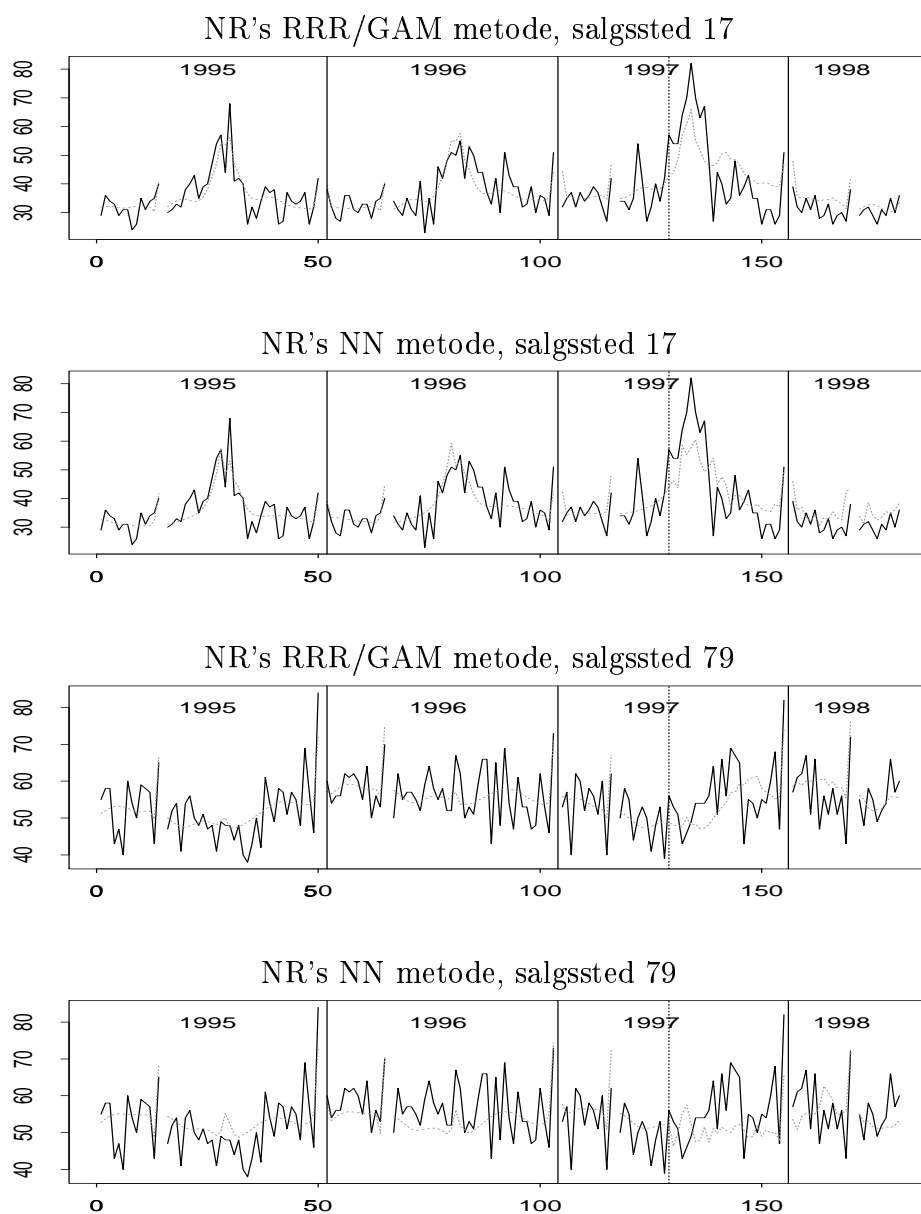
Tabell 2: Returandel og utsolgtvolum for bladtype 800.

Metode	Gruppe- type	Retur- andel i %	Utsolgt- volum i %	Gevinst i retur- andel i forhold til BC's metode i %	rmsee	<u>rmsep</u> rmsee
BC	!,*	28.8	3.1			
RRR/GAM	!,*	27.2	3.1	1.6	1.18	1.43
NN	!,*	33.0	3.1	-4.3	1.27	15.24
BC	A	25.9	7.3			
RRR/GAM	A	24.3	7.3	1.6	0.90	1.33
NN	A	24.7	7.3	1.2	0.92	1.54
BC	B	27.9	8.5			
RRR/GAM	B	26.2	8.5	1.7	0.83	1.29
NN	B	27.5	8.5	0.4	0.88	1.46
BC	C	30.3	16.5			
RRR/GAM	C	27.0	16.5	3.3	0.74	1.18
NN	C	27.9	16.5	2.4	0.76	1.23
Total veid gevinst for RRR/GAM metode i retur- andel i forhold til BC's metode i %				2.0		

Tabell 3: Returandel og utsolgtvolum for bladtype 801.

Metode	Gruppe- type	Retur- andel i %	Utsolgt- volum i %	Gevinst i retur- andel i forhold til BC's metode i %	rmsee	<u>rmsep</u> rmsee
BC	!,*	36.0	3.3			
RRR/GAM	!,*	32.1	3.3	3.9	1.05	1.19
NN	!,*	36.6	3.3	-0.6	1.09	1.28
BC	A	33.9	10.8			
RRR/GAM	A	32.8	10.8	1.1	0.96	1.18
NN	A	36.1	10.8	-2.2	1.04	1.46
BC	B	36.9	12.8			
RRR/GAM	B	34.9	12.8	2.0	0.88	1.50
NN	B	38.7	12.8	-1.8	0.95	4.00
BC	C	39.2	17.8			
RRR/GAM	C	36.3	17.8	2.9	0.78	1.20
NN	C	37.3	17.8	1.9	0.87	1.26
Total veid gevinst for RRR/GAM metode i retur- andel i forhold til BC's metode i %				2.4		

Tabell 4: Returandel og utsolgtvolum for bladtype 802.



Figur 1: Denne figuren beskriver salget for 2 steder (salgssted 17 og 79), spesifisert som heltrukken linje. Forventet salg utfra NR's 2 modeller er spesifisert som stiplet linje. Estimerings-perioden er fra uke 1 1995 til uke 25 1997. Prediksjonen er beregnet ut fra de historiske data opptil 4 uker før prediksjonen.