0.2088

O-83061

Statistikk-seminarer 1984

Arrangert på NIVA i samarbeid med

Norsk Regnesentral

Norsk institutt for vannforskning NIVA

Norsk Regnesentral NR

NIVA – RAPPORT

Norsk institutt for vannforskning NIVA

Hovedkontor
Postboks 333
0314 Oslo 3
Telefon (02) 23 52 80

Sørlandsavdelingen
Grooseveien 36
4890 Grimstad
Telefon (041) 43 033

Østlandsavdelingen
Rute 866
2312 Ottestad
Telefon (065) 76 752

Vestlandsavdelingen
Breiviken 2
5035 Bergen - Sandviken
Telefon (05) 25 97 00

Prosjektnr.: 0-83061

Undernummer:

Løpenummer: 2088

Begrenset distribusjon:

Rapportens tittel:
STATISTIKK-SEMINAR 1984
Arrangert på NIVA i samarbeid med Norsk Regnesentral

Dato: Januar 1988

Prosjektnummer: 0-83061

Forfatter (e):
Birger Bjerkeng (red.)
Dag Berge
Eivind Damsleth (NR)
Kim Esbensen (NR)
Arne Henriksen
Sigmund Kalvenes (NR)
Lars Kirkerud
Hans Vigge Sæbø (NR)
Rolf Volden (NR)

Faggruppe: Statistikk

Geografisk område:

Antall sider (inkl. bilag): 55

Oppdragsgiver: Statens forurensningstilsyn (SFT)

Oppdragsg. ref. (evt. NTNF-nr.):

Ekstrakt:
Rapporten er et referat fra 4 seminarer om bruk av statistiske metoder i overvåking av vannressurser. Seminarene tok opp følgende emner:

1. Bruk av statistikk for å tolke overvåkningsdata for tilstandskarakterisering og beskrivelse av trender, med eksempler fra Numedalslågen.
2. Analyse av tidsserier, med pH i vassdrag som eksempel.
3. Uformelle metoder, variabeltransformasjon, multivariat analyse.
4. Problemer rundt kluster- og regresjonsanalyse.

Rapporten refererer innledningene fra NIVA og NR og den etterfølgende diskusjonen.

4 emneord, norske:
1. Overvåkning
2. Statistiske metoder
3. Dataanalyse
4. Seminar

4 emneord, engelske:
1. Monitoring
2. Statistical methods
3. Data analysis
4. Seminar

Prosjektleder:
Birger Bjerkeng

For administrasjonen:
Merete Johannessen

ISBN - 82-577-1358-9

Norsk institutt for vannforskning (NIVA)
Norsk Regnesentral (NR)

0-83061

STATISTIKK-SEMINARER 1984

Arrangert på NIVA i samarbeid med
Norsk Regnesentral

Oslo januar 1988.

Prosjektleder: Birger Bjerkeng (NIVA)
Medarbeidere: Dag Berge (NIVA)
Eivind Damsleth (NR)
Kim Esbensen (NR)
Arne Henriksen (NIVA)
Sigmund Kalvenes (NR)
Lars Kirkerud (NIVA)
Hans Viggo Sæbø (NR)
Rolf Volden (NR)

FORORD.

Høsten 1984 arrangerte Norsk Regnesentral og Norsk institutt for vann-forskning en serie på fire statistikk-seminarer på NIVA. Seminarene ble arrangert i tilknytning til prosjektet "Bruk av statististiske metoder innen forurensnings-overvåkningen", som utføres i samarbeid mellom NIVA og NR på oppdrag for SFT.

Referatene fra seminarene er nå samlet i denne rapporten.

Hvert seminar presenteres i et eget kapittel.

En litteratur-liste finnes bakerst i rapporten.

Rapporten gir ikke noe ordrett referat fra seminarene. Den er basert på undertegnedes notater som referent, og tildels også på skriftlig materiale fra innlederne. Referatene er gjennomgått av innlederne, som alle har bidratt vesentlig til sluttredigeringen.

Seminarene samlet hver gang 15-20 NIVA-ansatte, mest forskere.

Birger Bjerkeng

Birger Bjerkeng

20.1.1988

I N N H O L D S F O R T E G N E L S E

1 STATISTISKE PROBLEMSTILLINGER TILKNYTTET OVERVÅKNINGSPROGRAMMET, MED EKSEMPLER FRA NUMEDALSLÅGEN

Tid:	25. september 1984 fra kl.9.30 til 11.30
Innledere:	Dag Berge, NIVA
	Hans Viggo Sæbø, NR.
Deltagere utenfor NIVA ellers:	
	Fra SFT: Anne Lill Gade.
	Fra NR: Kim Esbensen, Sigmund Kalvenes.

Hovedemnet for dette seminaret var hvordan en kan bruke statistikk til å tolke overvåkningsdata, både for å karakterisere tilstanden i en vannforekomst og for å beskrive endringer over tid (trender).

Seminaret knytter seg til en rapport fra Norsk Regnesentral (Sæbø, 1984), som bruker data fra NIVA's overvåkning av Numedalslågen som eksempel.

Overvåkningsprosjektetet for Numedalslågen er rapportert årlig, se f.eks. (NIVA 1984).

Seminaret hadde to innledere:

- NIVA's saksbehandler for overvåknings-prosjektet, Dag Berge, redegjorde først for prosjektet.
- Hans Viggo Sæbø fra Norsk Regnesentral presenterte så hovedpunktene i sin rapport.

Seminaret ble avsluttet med en diskusjon.

1.1 Dag Berge: Om overvåkningsprosjektet i Numedalslågen

Det har vært rutinemessig overvåkning i Numedalslågen siden 1977, som oppfølging av en noe ufullstendig "basis-undersøkelse", se kartskisse i figur 1.

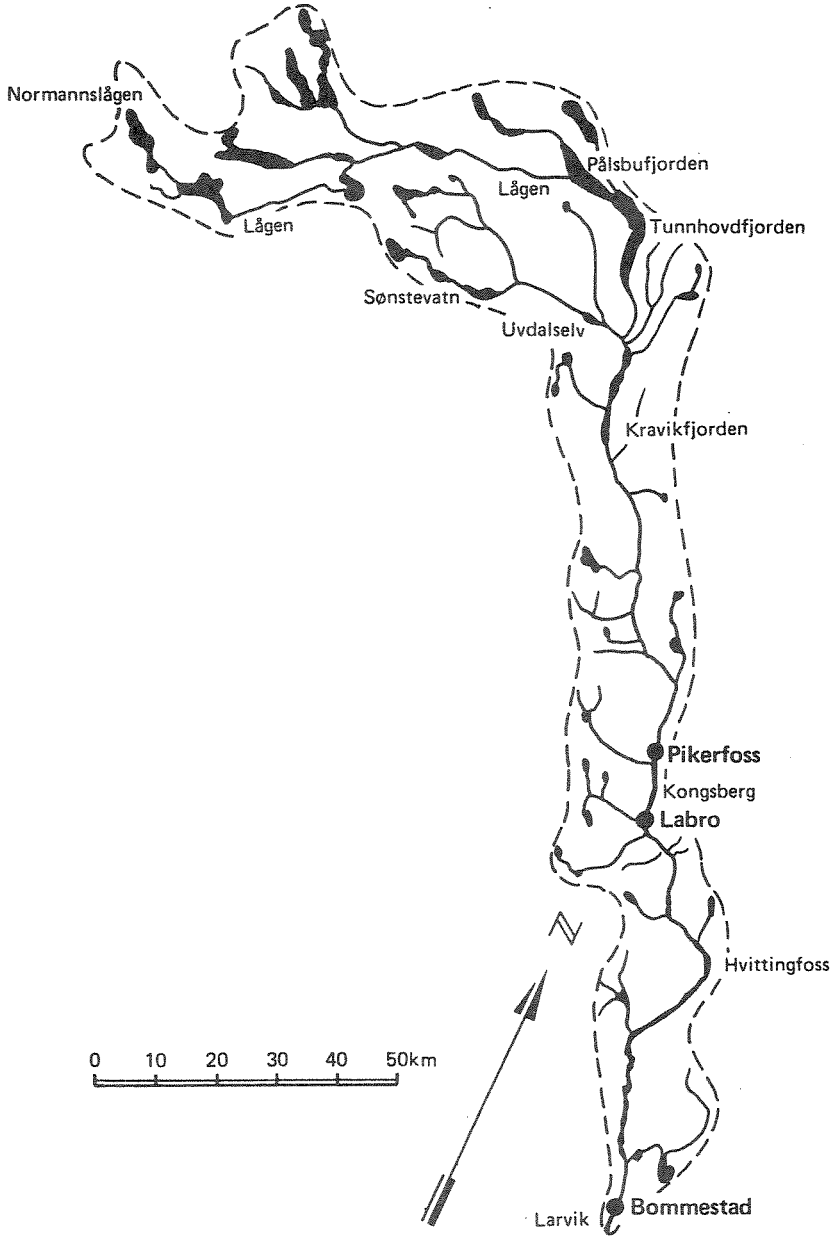


Fig. 1.Kartskisse over Numedalslågen med nedbørfelt. Stasjonene som inngikk i 1983 undersøkelsene er markert med sorte sirkler (Fra NIVA 1984)

Formålet er å holde forurensings-situasjonen under oppsikt over tid.

Planleggingen av et slikt prosjekt må avveie flere hensyn:

- Ønske om bestemte stasjoner
  (fra interessenter, lokale myndigheter)

- Ønske om bestemte parametre
  (ut fra de aktiviteter som foregår i vassdraget)

- Økonomiske rammer

- Krav til faglig forsvarlige utsagn om situasjonen.

Problemet blir ofte å få mest mulig ut av et magert materiale.

Det en vil beskrive er som regel:

- Forurensningssituasjonen for en gitt stasjon

- Sammenligning mellom flere stasjoner

- Utvikling over tid

Grafiske fremstillinger fra den siste overvåkningsrapporten (se f.eks. figur 2) viste at selv når en ser på årsmidler vil tallene som regel svinge frem og tilbake fra år til år gjennom en periode, og det kan være vanskelig å trekke konklusjoner om tidsutviklingen uten bruk av statistikk.

Forskjellen mellom øvre og nedre stasjoner i vassdraget var det lettere å se umiddelbart, særlig når det gjaldt bakterier.

Det kan være vanskelig å vite hva slags statistiske mål som bør brukes til å besvare forskjellige spørsmål. Skal stoff-transport beregnes virker det riktigst å bruke middelverdier, mens median kanskje kan være et bedre tall dersom en mest ønsker å beskrive typiske situa-

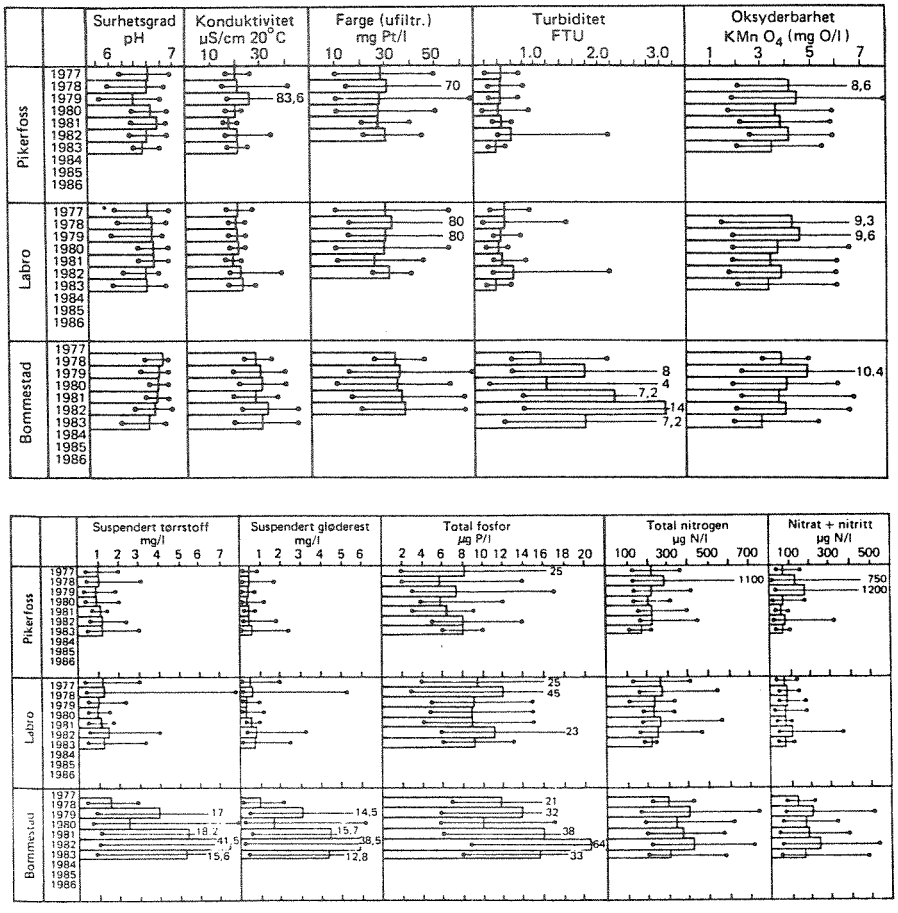


Fig. 2.Tidsveide årsmiddelverdier for en del fysisk/kjemiske variable fra Numedalslågen for perioden 1977-83. (Fra NIVA 1984)

sjoner. Det er også flere måter å beskrive variasjon på. Ekstrem-verdier kan ofte være interessante, bl.a kan det være de som gir sterkest uheldige virkninger (eks. fiskedød pga. lav pH).

Det er viktig at de beskrivelsesmåter som brukes blir forstått av politikere og forvaltning.

En viktig faktor er årstidsvariasjonene, som en må kunne korrigere for hvis en skal få noe ut av materialet.

Berge syntes rapporten fra NR var godt skrevet, og lett å forstå for ikke-statistikere, og at den besvarte hans spørsmål på en god måte.

Spørsmål: Hvor typisk er Numedalslågen som overvåkningsprosjekt?

Svar: Kanskje noe atypisk ved at det er en stor, gjennomregulert elv uten strykpartier (Bortregulerte).

Det er vanskelig/dyrt å gjøre biologiske observasjoner i slike elver, og en er stort sett begrenset til å måle kjemiske/bakteriologiske variable. To biologistasjoner er blitt ødelagt i løpet av undersøkelsen pga. regulering. Begroing/bunndyr ville ellers gi bedre bilde av middel-tilstand, også med hensyn til ekstremverdier.

Dessuten mangler vassdraget store innsjøer som ofte er de svakeste ledd i et vassdrag. Innsjøer hadde gjort det lettere å kvantifisere forurensningsutviklingen i vassdraget.

Selve programmet ellers, med kjemi/bakteriologi-prøver 12 ganger i året er ellers forholdsvis typisk for oversvåkningsprosjektene.

1.2 Hans Viggo Sæbø: Resultatene fra det arbeidet NR har gjort i tilknytning til Numedalslågen

Utgangspunktet var NIVA's overvåkningsrapporter (NIVA 1981, 1982, 1983, 1984), og noen konklusjoner derfra som var spesielt avmerket fra NIVA's side for etterprøving.

Innholdet i Sæbø's innlegg vil stort sett finnes i (Sæbø 1984), og det refereres derfor bare kort her:

Rapporten er konsentrert om to aktuelle problemer:

-Når skal det måles?

-Hvordan skal målingene tolkes?

Analyse av endringer er sentralt. Det må avklares hvilke endringer som er interessante. Tidsperspektivet avgjør hvilke endringer som i en tidsrekkemodell betraktes som hhv. trend og tilfeldige svingninger.

Målte miljøindikatorer avhenger både av faste og variable naturlige forhold og av menneskelig aktivitet, oppgaven er å finne disse sammenhengene.

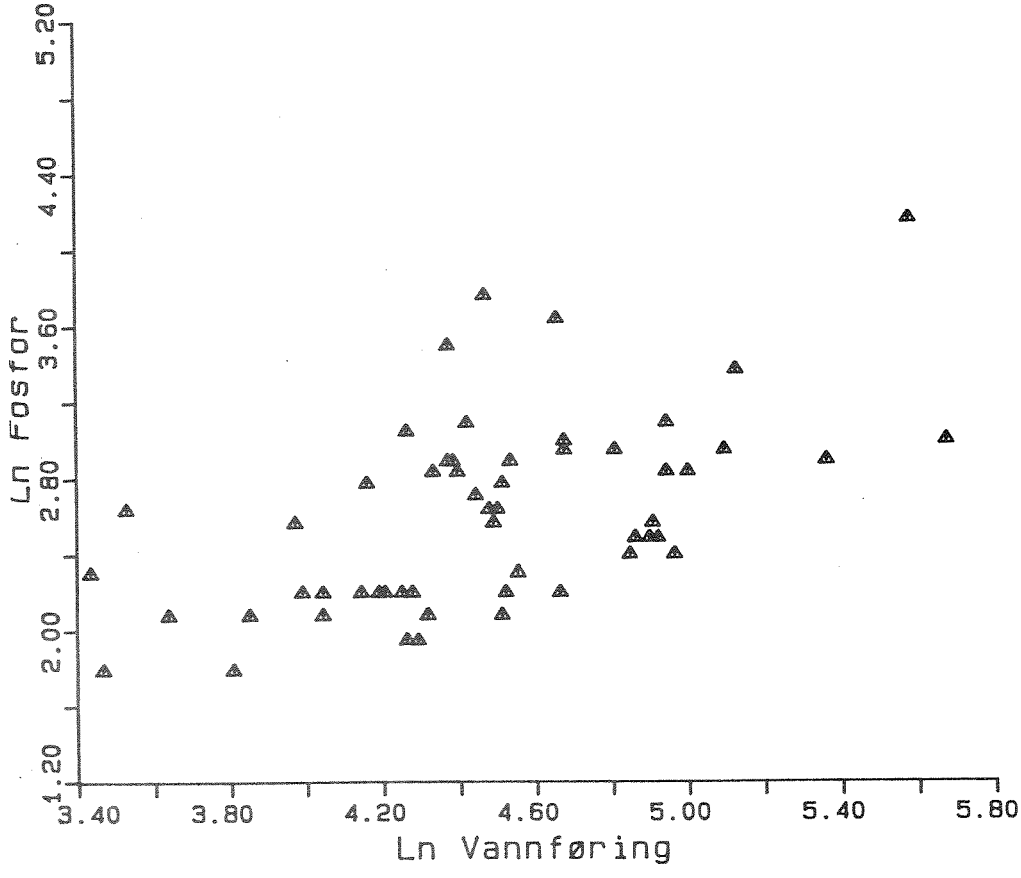


Fig. 3.Sammenhengen mellom total fosfor og vannføring ved Bommestad. Målinger utført 1980-1983 (Fra Sæbø 1984).

Et eksempel er sammenhengen mellom fosfor og vannføring i Numedalslågen, hvor det en klart signifikant positiv korrelasjon (figur 3).

Målingene tyder på lavere fosfor-verdier i 1983 enn i 1982, men en statistisk analyse ved multippel regresjon (kovarians-analyse) gir ikke noe klart svar, det ser ut til at modellen er for enkel, slik at restleddene er korrelerte.

Data fra 1983 viser imidlertid markert avvik fra det generelle mønsteret: Det er ikke høye fosforkonsentrasjoner ved store vannføringer. Dette tyder på at konklusjonene i overvåkningsrapporten kan være riktig: Den store vannføringen i 1983 skyldes rent vann fra fjellområder, og det gir en fortynning, i strid med hva en ville vente

ut fra den generelle sammenhengen. Ellers kan det ofte være farlig å kommentere endringer mellom to perioder som bygger på enkelte ekstremverdier.

Ved å modellere sammenhengen mellom vannføring og konsentrasjon bedre kunne en kanskje få mer presise estimater for virkningen av menneskelige aktiviteter.

Årstidsvariasjoner er viktige: De kan behandles ved å se på gjennomsnitt innenfor perioder av et år, eller ved periodiske funksjoner.

Alment bør det måles med faste intervaller innen hver periode, evt. med økt hyppighet i sommerhalvåret, men intervallene må ikke falle sammen med periode for naturlige svingninger (f.eks. over døgnet).

1.3 Fra den etterfølgende diskusjon

Bjørn Faafeng trakk frem hvor viktig det var å ta hensyn til døgnvariasjoner og tilfeldige korttidsvariasjoner. Ved hjelp av egne tette observasjonsserier fra mindre elver og bekker viste han på en overbevisende måte hvor viktig dette kan være. Variasjonene kan være mer eller mindre systematiske.

Sæbø (NR): Dersom variasjonene er tilfeldige er dette OK selv om en har få målinger, usikkerheten vil likevel fremgå av spredningen i dataene.

Hvis det er et mer fast variasjonsmønster, vil målinger tatt omtrent på samme tid hver dag gi et skjevt bilde av f.eks. forurensningstransport, men kan likevel gi et riktig bilde av relativ utvikling over tid.

Lasse Vråle viste tilsvarende eksempler fra målinger i avløpsnett, hvor det er systematisk variasjon gjennom døgnet. Dersom en vil ha et bilde av gjennomsnittlig transport er det helt nødvendig å bruke blandprøvetagere. Den samme variasjonen går igjen i resipienter som er direkte påvirket av avløpsvann over kort tid, men med noe forsinkelse.

For avløpsvann fra menneskelige aktiviteter er det også variasjoner med ukedag, som en må ta hensyn til. Kurver tegnet ut fra analyse av døgnvannprøver (Sydskogen) ga klar indikasjon på det.

Variasjonene påvirkes av faktorer som innvirker på menneskelig aktivitet, f.eks. været, og kan dessuten være ulike for forskjellige avløpsfelt.

Berge nevnte Telemarksvassdraget hvor det var gjort forsøk med blandprøvetagere for å se på betydningen av døgnvariasjoner. Ifølge Tjomsland viste resultatene vesentlig endring i forhold til målinger som bare var gjort på dagtid.

Berge: Et annet problem er plassering av målested i elvetverrsnittet. Ved Notodden, hvor elva renner rolig i et stort tverrsnitt, vil en få mye høyere verdier i prøver tatt nær land enn ute i hovedstrømmen, fordi avløpsvannet fra utslipp langs bredden ennå ikke har blandet seg effektivt inn i hele tverrsnittet. Utløpet av en innsjø er best hvis en vil måle endringer i transport over lengre tid, fordi innsjøen virker som blandebasseng og utjevner topper.

Arnesen: I tradisjonelle statistiske metoder forutsettes ofte at materialet er normalfordelt. I hvilken grad blir resultatene gale når dette ikke er tilfelle?

Kim Esbensen(NR): Et datum (måle-verdi) = signal + støy. Statistikk skal hjelpe oss med å isolere støyen, slik at vi får frem signalet. Statistikken gir oss imidlertid ikke signalet, men en "modell"=signal + feil, hvor feilen ønskes minst mulig. Middelverdi er et enkelt eksempel på en slik statistisk modell. Jo enklere modellen er, dess flere systematiske faktorer er tatt med i feilen. Modellen er tilfredsstillende dersom den reduserer støyen til hvit støy, hvor restledd er ukorrelert, og med normalfordeling rundt gjennomsnitt 0. Mer komplekse modeller kan redusere feilen, og få signalet klarere fram.

Systematiske tidsvariasjoner kan en ta hensyn til gjennom f.eks. tidsseriemodeller. Dette er emnet for det neste seminaret. Komplekse sammenhenger kan en få frem ved multivariat analyse, som blir tatt opp i det tredje seminaret.

Sæbø: Når det er få observasjoner er det ofte vanskelig å motbevise normalfordeling. Ved større/tettere serier må en bruke mer avanserte modeller for å få støyen normalfordelt. Det bør alltid undersøkes om restleddene (støyen) er korrelert, f.eks. dersom en bruker regresjon.

Ingen statistisk teknikk kan løse problemet med døgn-variasjoner hvis det systematisk er målt til visse tider av døgnet.

En må ta hensyn til slike variasjoner ved å plassere stasjoner slik at en får utjevning, eller ved automatisk prøvetaging e.l.

Sigmund Kalvenes(NR): Statistiske metoder og komplekse modeller kan ikke sikre at problemstillingene er formulert riktig - valg av modell må gjøres ut fra faglig kunnskap.

Holtan: Vassdraget Gaula i Trøndelag skal overvåkes fra 1985. Det er et komplisert vassdrag med mange sideelver, mye nedbør, og lite innsjøer. Vannføringene kan variere sterkt, fra $2m^3/s$ til $3000m^3/s$ ved Støren, og variasjonene kan være meget raske. Flommer kan en få også utenom vårperioden pga. regnskyll.

Oppgaven for overvåkningsprosjektet blir å kartlegge betydningen av forurensningstilførsler (fra 20 000 personer, store jordbruksområder, og utstrakt gruve-virksomhet) og finne ut i hvilken grad tiltak er nødvendige.

Budsjettrammer tilsier at en må nøye seg med et begrenset antall enkelt-observasjoner, automatiske prøvetagere er ikke aktuelt. Hvordan kan statistikerne hjelpe til med planleggingen, slik at en kan få maksimalt ut av små ressurser? Dersom NR kan tenke på dette til neste gang ville det være fint.

Arnesen presenterte diagrammer som viste målt sammenheng mellom vannføring og konsentrasjon av fosfor og turbiditet i løpet av enkelte flom-episoder. Konsentrasjonene er systematisk høyere ved stigende vannføringer. For å beskrive dette skikkelig kreves stor innsats.

Sæbø: I NR's rapport for Numedalslågen er det vist sammenheng mellom vannføring og konsentrasjon, men den kan være ulik fra år til år. Ved å skille mellom flere typer vannføring, f.eks. tilførsel fra høyfjellet kontra avrenning fra lavlandsområder, vil en kunne få bedre tilpasning. Men mer komplekse modeller krever også et bedre og større datamateriale.

Esbensen: Statistikk og data-analyse kan hjelpe til å undersøke mulige sammenhenger mellom observasjoner og forklaringsvariable. Ved å lage meget komplekse modeller kan en alltid få god tilpasning, men det er et faglig problem å velge de rette forklaringsvariable (faktorer)

Arnesen: Tidsseriemodeller krever tradisjonelt like tidsintervall mellom målingene. Mot dette står hensynet til at en vil ha karlagt flest mulige situasjoner. Hva er viktigst? I hvilken grad kan en håndtere målinger med ulike tidsintervall?

Sæbø: Enkeltstående unntak kan håndteres, men dersom det gjennomgående er ulike tidsintervall kan en vanskelig benytte tidsserie-modeller. En mulighet er å ta ekstra prøver i visse perioder, og se bort fra dem i tidsseriemodellene.

Berge: Budsjettene setter ofte stramme grenser for hvor mange ganger en kan ta prøver.

Esbensen: Hva med bruk av blandprøver?

Holtan: Hvis man er ute etter ekstremverdier hjelper det ikke med blandprøver. Egentlig ville en trenge kontinuerlige målinger. Når det er umulig er spørsmålet når en skal ta prøver for å få maksimalt igjen for innsatsen.

I Gaula har vi tenkt å engasjere en fra lokalbefolkningen til å ta prøver. Prøvetidspunktet velges ut fra variasjonene i vannføringen, ut fra en fastlagt strategi.

Esbensen: Hvis NR skal bidra i planleggingen må vi få noe bakgrunnsmateriale.

Anne Lill Gade (SFT): SFT har laget en problemanalyse for Gaula-vassdraget som viser hvilke forurensningskilder som finnes, og hva vi ønsker svar på. Den bør NR kunne få som en del av dette materialet.

Traaen: Det er viktig at en på forhånd definerer hvilke differanser en bør kunne påvise, f.eks. når en skal sammenligne et år med et annet. Her må vannforskerne bli flinkere. Ofte har vi urealistiske forestillinger om hva man kan få ut av et materiale. I et tilfelle viste det seg at datagrunnlaget bare ville gi grunnlag for å påvise en økning i tilførsler som tilsvarer en seksdobling av befolkningen ved bruk av årsmidler.

Esbensen: Ved å bruke modeller kan en redusere den nedre grensen for påviselige endringer ut fra en gitt prøvefrekvens. Det er her "fagmannens" og statistikerens felles ansvar at modellen er både realistisk og praktisk anvendelig.

2 Analyse av tids-serier, med pH i vassdrag som eksempel

Tid: 18. oktober 1984 fra kl.9.00 til 11.30

Innledere: Arne Henriksen, NIVA
Eivind Damsleth, NR

Deltagere utenfor NIVA ellers:
Fra SFT: Morten Svelle
Fra NR: Sigmund Kalvenes.

Emnet for dette seminaret var analyse av tids-utvikling.

Arne Henriksen redegjorde for en analyse av pH-trender i vassdrag, basert på regresjon, hvor sesong-variasjoner ble modellert som sinusledd.

Einar Damsleth orienterte om tids-serie-analyse mer generelt, med spesiell vekt på ARIMA-modeller.

2.1 Arne Henriksen: Studium av trender for pH i vassdrag innenfor sur-nedbør programmet

Det arbeidet som omtales her er presentert i en rapport (Henriksen et.al. 1981).

Nedenfor omtales derfor bare hovedlinjene kortfattet.

Datainnsamlingen ble opprinnelig startet i 1965 av nå avdøde vitenskapelig konsulent Einar Snekvik i Direktoratet for viltstell og ferskvannsfiske (DVF). Det ble opprinnelig gjort målinger i 11 elver. DVF's rutineopplegg omfatter idag 38 elver med 88 lokaliteter. Fra 1970 kom sur-nedbør prosjektet (SNSF) inn i bildet.

Fra 1980 er det også samlet inn data under SFT's overvåkningsprogram, på ialt 20 elver, hvorav 13 fra DVF's måleprogram.

Data foreligger stort sett med månedlige målinger, hver uke under snøsmeltingen om våren.

I rapporten nevnt ovenfor presenteres en bearbeidelse av de data DVF har samlet inn til og med 1979. Bare pH og tildels hardhetsdata er behandlet.

Data-analysene er utført av Rolf Volden ved Norsk Regnesentral.

Data fra 1965-70 for 19 elver ble analysert i 1972. For å se på tidsutviklingen ble det først prøvd med vanlig regresjon mot tid.

Data-seriene hadde imidlertid innebygget støy i form av naturlige variasjoner gjennom året. Generelt var det forholdsvis høye pH-verdier sommer og vinter, mens verdiene minsket under vårflom og høstflom. Det ble derfor laget en multippel regresjonsmodell, hvor årstidsledd ble tatt med i form av sinus-ledd og cosinus-ledd med helårlig og halvårlig periode. Dette ga en klar forbedring av modellens forklaringsevne.

Ved den fornyede analysen, som omfattet alle data frem til og med 1979, ble problemet forsøkt løst på to måter:

1. Ved enkel regresjon mot tid for årsmidler.

2. Ved å ta med årstidsledd i en multippel regresjonsmodell for enkeltobservasjoner.

Begge metodene ga signifikante lineære trender for pH i nesten alle elver, med reduksjoner i pH på ca. 0.02 pr. år.

Det ble også gjort en sammenligning av den kumulative frekvensfordeling mellom to perioder, 1970-72 og 1977-79. For hver elv ble median, 10% og 90%-persentil sammenlignet for de to tidsperiodene. Resultatet er vist i figur 4 (tallene langs horisontal-aksen angir elv og lokalitet).

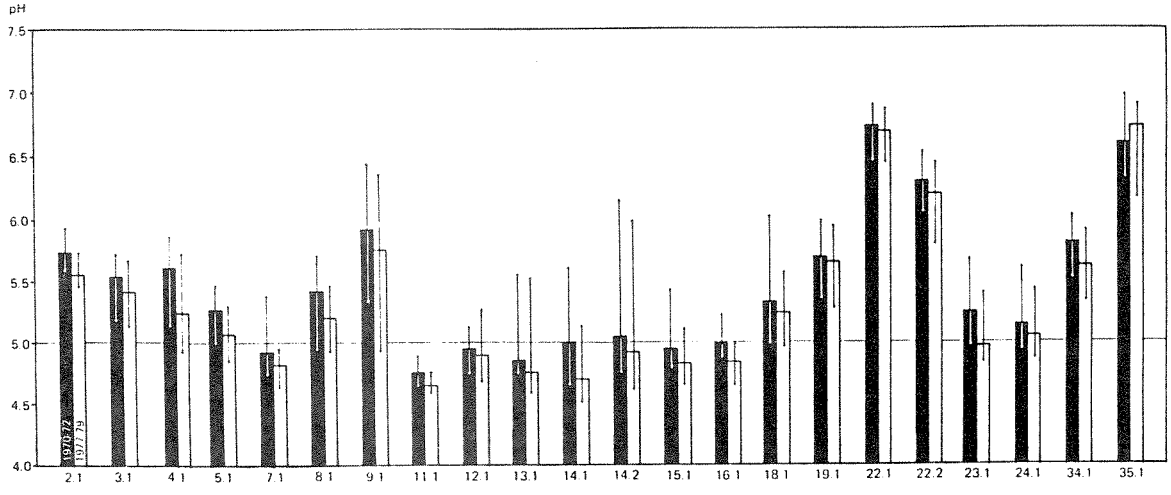


Fig. 4.Kumulativ hyppighet av pH-observasjoner i elver med regelmessige observasjoner fra 1970 for periodene 1970-72 og 1977-79. Median og 10% og 90% fraktiler er vist. Fra Henriksen et.al. 1981

Det fremgår at alle elver unntatt en (Namsen, 35.1) har redusert pH i siste periode.

Residualene fra regresjonsmodellene ble analysert. Et eksempel på en slik analyse er vist i figur 5.

Resultatet av disse analysene kan oppsummeres slik: Svært mange observasjoner avviker fra modellen på en tilfeldig måte, og det er ønskelig. Mange observasjoner avviker imidlertid også systematisk over sammenhengende perioder, og det er mer alvorlig. Systematiske avvik ser delvis ut til å skyldes unormale regnværs- eller tørkeperioder, eller atypiske år. Dette gjør at en må være forsiktig med å tolke signifikans-sannsynligheter fra analysene.

Analysen styrkes imidlertid av at både regresjon på enkeltdata og enkel regresjon på årsmidler viser en rekke felles tendenser, med reduksjon i pH over tid, og at dette går igjen i svært mange av elvene.

Damsleths tidsrekkeanalyse på data fra 1970 til 1980 ga i motsetning til dette ingen signifikant trend når det ble tatt hensyn til vann-føringen. Hans analyse gir indikasjoner på en lokal nedgang i 1977, 1978 og tildels også 1979, etterfulgt av høyere verdier i 1980. Dette kan forklare forskjell i konklusjon for de to analysene: Når 1980 ikke er med fremtrer nedgangen 1978-79 som en trend.

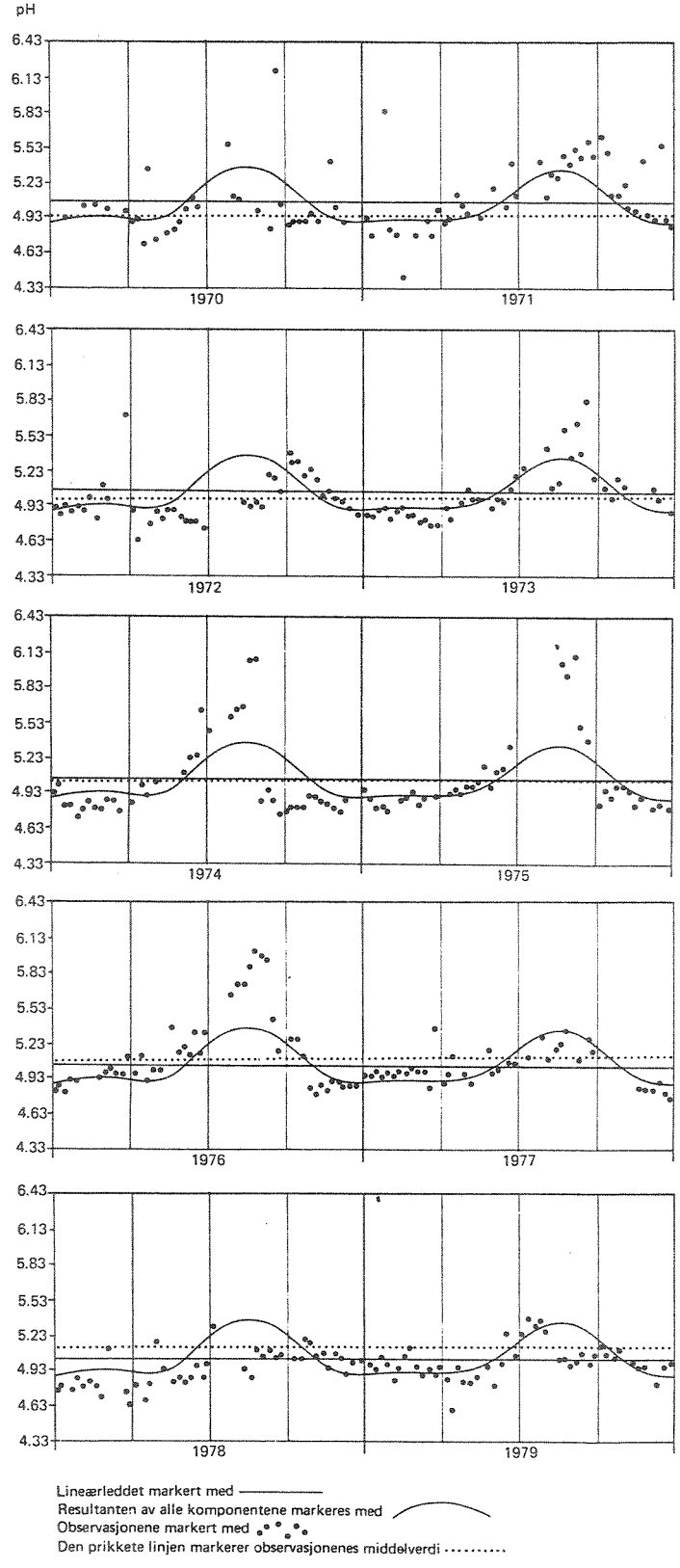


Fig. 5.Residualanalyse av data fra Tovdalselva.
(Fra Henriksen et.al. 1981)

2.2 Eivind Damsleth (NR): Tidsrekke-analyse

En formell definisjon av tidsrekker kan gis som nedenfor.
(Fra Box & Jenkins 1970)

> *Time series.* A time series is a set of observations generated sequentially in time. If the set is continuous, the time series is said to be *continuous.* If the set is discrete, the time series is said to be *discrete.* Thus, the observations from a discrete time series made at times $\tau_1, \tau_2, \ldots, \tau_t, \ldots, \tau_N$ may be denoted by $z(\tau_1), z(\tau_2), \ldots, z(\tau_t), \ldots, z(\tau_N)$. In this book we consider only discrete time series where observations are made at some fixed interval h. When we have N *successive* values of such a series available for analysis, we write $z_1, z_2, \ldots, z_t, \ldots, z_N$ to denote observations made at equidistant time intervals $\tau_0 + h, \tau_0 + 2h, \ldots, \tau_0 + th, \ldots, \tau_0 + Nh$. For many purposes the values of τ_0 and h are unimportant, but if the observation times need to be defined exactly, these two values can be specified. If we adopt τ_0 as the origin and h as the unit of time, we can regard z_t as the observation *at time t.*
>
> Discrete time series may arise in two ways.
>
> (1) By *sampling* a continuous time series; for example, in the situation shown in Figure 1.2, where the continuous input and output from a gas furnace was sampled at intervals of nine seconds.
>
> (2) By *accumulating* a variable over a period of time; examples are rainfall, which is usually accumulated over a period such as a day or a month,

Det forutsettes at målingene er ekvidistante. Det er mulig, men vanskelig, å behandle målinger som ikke er ekvidistante.

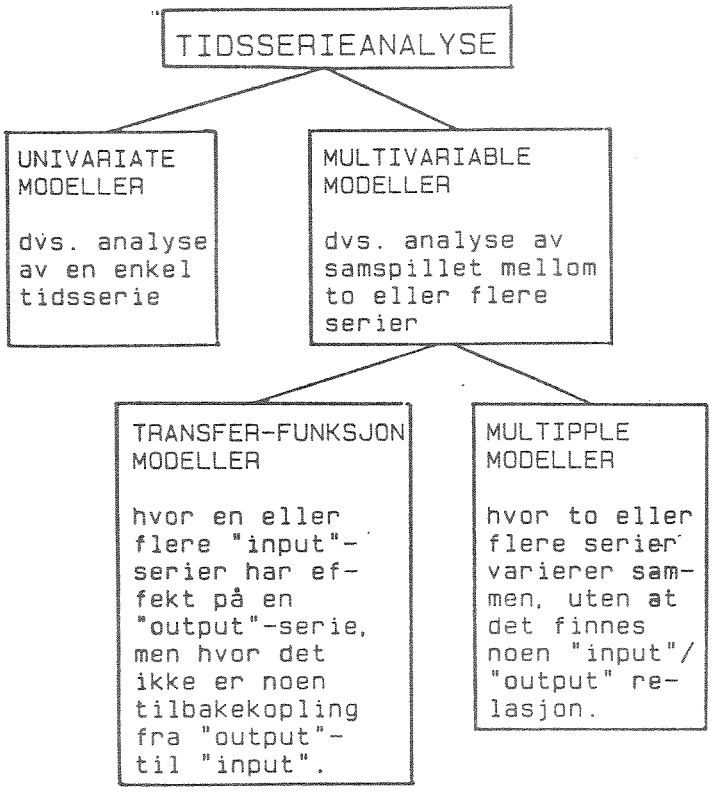
Formålet med å bruke kvantitative prognosemetoder for miljødata er like mye analyse som prognose.

Kvantitative prognose-metoder kan deles inn slik:

Tidsrekke-analyse:	- dekomponering	
	- spektralanalyse	
	- ARIMA-modeller	
Regresjon:	- fast trend	
	- kausale modeller	
	- struktur-modeller	brukes av
	- økonometri	økonomer

Her skal vi snakke mest om ARIMA-modeller, som beskriver en kovarians-struktur.

Det finnes forskjellige modelltyper, som kan systematiseres slik:



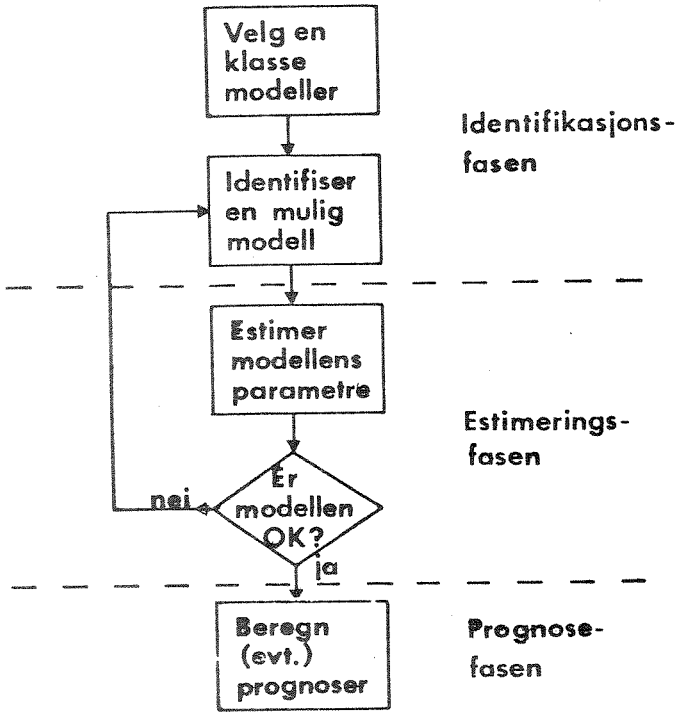
Som et eksempel på en transfer-funksjon modell kan nevnes pH som funksjon av vannføring. Det er ingen tilbakekobling fra pH til vannføring.

Når vi tilpasser modeller til en tidsrekke, gir modellen forslag til "hukommelse" i serien (f.eks. middel av siste uke). Målet er å komme frem til en modell hvor residualene er uavhengige stokastiske variable, dvs. representerer hvit støy. Dersom residualene har en kovarians-struktur kan det gi ideer til forbedring av modellen.

Ved bruk av regresjon får vi riktige parametre selv om restledd er avhengige, men beregnet standardavvik og tester med konfidensnivåer vil ikke være riktige.

Ved tidsrekker modelleres strukturen i restleddene, og vi prøver å komme over i uavhengige observasjoner.

Fasene i en tids-serie analyse kan fremstilles slik:



Kriteriet på om modellen er OK er at restleddene er uavhengige.

NR har brukt metodene både for prognose- og analyse-formål:

Prognose-prosjekter:
- NVE
- Televerket
- Elkem
- Postgirokontoret
- SSB
- Norges Bondelag
- PFI

Analyse-prosjekter:
- NIVA
- SIFA
- Rendalsprosjektet
- NVE
- Televerket
- Postgirokontoret

En viktig ting å være oppmerksom på er at tilfeldige svingninger i seriene kan gi tilsynelatende signifikante trender sett over kortere perioder.

Et eksempel på dette er en analyse av temperaturdata som ble gjort i forbindelse med vannkraftutbygging i Rendalen. (Pleym, 1980).

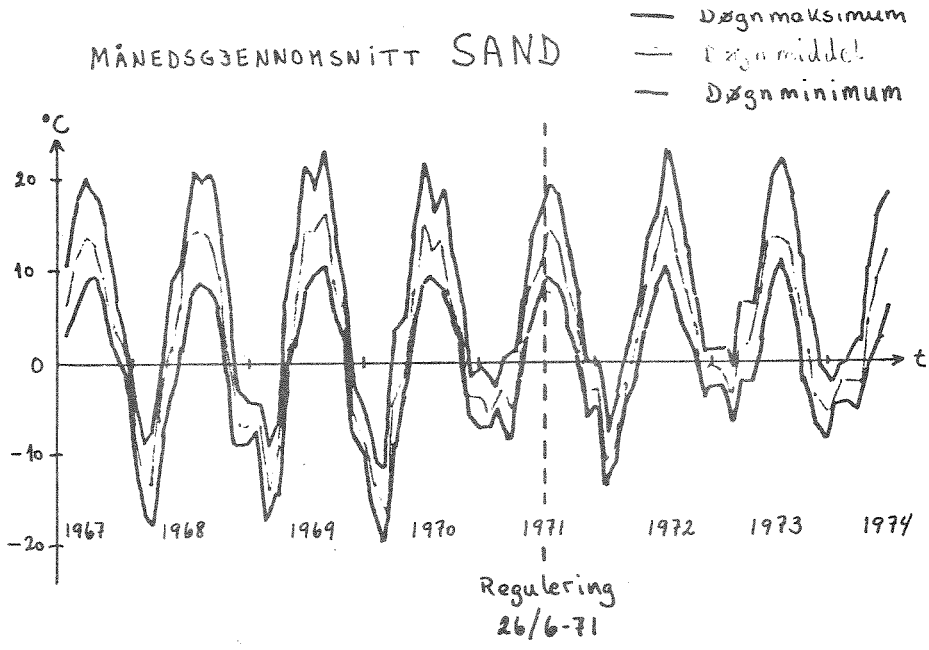


Fig. 6.Månedsgjennomsnitt for temperatur, Sand stasjon, Rendalen (Etter Pleym, 1980)

Det visuelle inntrykket av dataene for Sand stasjon i det regulerte vassdraget var at vintertemperaturene var høyere enn før i de tre årene etter reguleringen (se figur 6).

Data-serien ble analysert ved en intervensjonsmodell, hvor serien ble beskrevet som en vanlig tidsserie + ledd som kan beskrive en endring etter regulering uavhengig for hver årstid:

$$Y_t = w_1 X_{1t} + w_2 X_{2t} + w_3 X_{3t} + w_4 X_{4t} + n_t$$

hvor

Y_t = Normaliserte pentader,(minimum, maksimum og middel).

Tidsserie-modellen ble funnet å ha formen:

$$n_t = a_t - \Theta_1 a_{t-1} - \Theta_2 a_{t-2} - \Theta_3 a_{t-3}$$

hvor a_t er residualene.

X_{1t} = 1 i vintermånedene etter regulering,
= 0 ellers

X_{2t}, X_{3t} og X_{4t} er tilsvarende for vår, sommer og høst.

Ved modelltilpasningen estimeres modellparametrene, som her er w_i og Θ_i, og samtidig får en beregnet residualene a_t.

Analysen viste signifikant temperaturøkning i vinter- og vår-månedene: w_1 og w_2 var signifikant større enn null.

En analyse av temperatur-differansen mellom stasjonen i Rendals-vassdraget og en referansestasjon (Tynset) med en tilsvarende modell (n_t ble nå et enklere uttrykk) ga imidlertid det motsatte resultat: det var ingen signifikante endringer.

Konklusjonen på dette ble (sitat fra Pleym 1980):

"Det kan ikke påvises at reguleringen har ført til signifikante forandringer i temperaturklimaet på pentade- og måneds-midler på de analyserte 9 stasjonene i Rendalen. De påviste forandringene i temperaturklimaet lokalt skyldes endringer i klimaet på regional skala. Klimaet i prosjektperioden kan karakteriseres ved signifikant midlere vintermåneder etter reguleringen sammenlignet med perioden før reguleringen."

En skal altså være forsiktig med å trekke konklusjoner ut fra en analyse av en enkel rekke.

NR har gjort en metode-studie for SFT angående bruk av tidsrekke-analyse på forurensningsdata. Arbeidet er beskrevet i en egen rapport (Damsleth 1984). For detaljer henvises til rapporten.

Rapporten tar for seg to eksempler:

A) pH i vassdrag.
B) Sulfat i luft.

Formålet med denne studien var å se i hvor stor grad tidsrekke-analyse er egnet til å beskrive slike tidsrekker.

2.2.1 EKSEMPEL A: pH I vassdrag

Data for Nidelva, Tovdalselva og Mandalselva ble analysert.

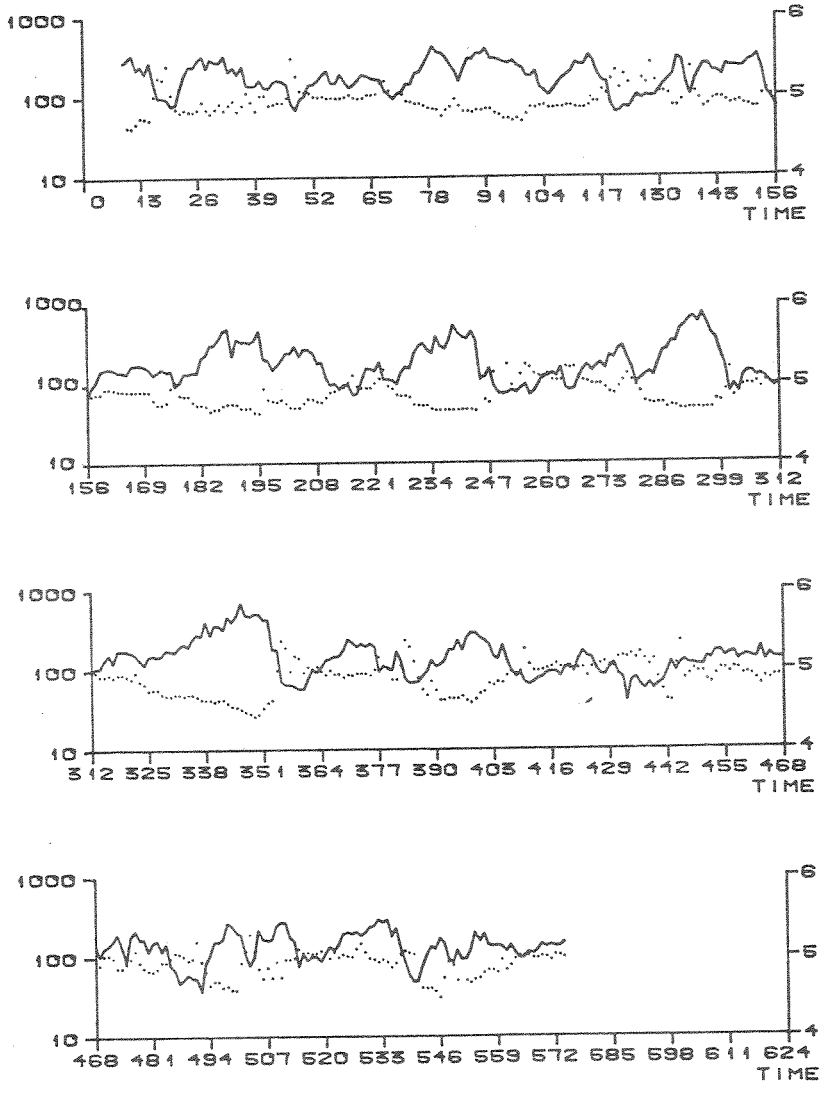


Fig. 7.pH (stiplet) og log-transformert vannføring for Nidelva 1972-1982. (Fra Damsleth 1984)

I figur 7 er vist en av de tidsseriene som ble analysert, pH og vannføring for Nidelva 1972-1982.

pH-seriene ble analysert i en univariat modell, og etter noen iterasjoner ble det identifisert en modell hvor pH-verdien på et gitt tidspunkt ble beskrevet som en lineær kombinasjon av pH i foregående uke (kort-tids korrelasjon), og pH ett år tidligere (beskriver årssyklus), i tillegg til residual-ledd.

Når residualene ble analysert ble det ikke funnet endring over tid, bortsett fra tendenser til lavere verdier i 1977-1979, men det ble funnet signifikante forskjeller mellom årstidene.

Det ble også tilpasset en modell hvor log av vannføring ble antatt å styre pH ved en transferfunksjon, og korrigerte pH-verdier analysert som tidsrekke. Også her ble det signifikante forskjeller i residualer mellom årstidene, modellene er altså ikke gode nok til å beskrive alle årstider.

2.2.2 EKSEMPEL B: Sulfat i luft

Utgangspunktet var tidsserier for sulfat-konsentrasjoner ved tre stasjoner, Birkenes, Skreådalen og Hummelfjell.

Disse tidsseriene ble analysert sammen med data for vindretning. Vindretningene ble gruppert i 8 sektorer.

Det ble så estimert en modell hvor sulfatkonsentrasjonen ble analysert som en tidsserie, men med avhengighet av vindretning modellert ved hjelp av en intervensjonsmodell analogt med den som er vist ovenfor fra Rendalsprosjektet.

Det viste seg at variasjonen med vind-retning var nokså nær sinus-formet rundt sirkelen, og det ble derfor også tilpasset en forenklet modell, hvor retningsavhengigheten ble beskrevet direkte som en sinusfunksjon. Modellene ga ingen signifikant endring over tid.

2.3 Fra den etterfølgende diskusjonen

Torgeir Bakke, NIVA: Er multiple modeller mer kompliserte enn vanlige transferfunksjon-modeller?

Damsleth: Det er mer kompleks avhengighet, en får fort mange parametre. (Variasjon mellom flere variable og flere tidsperioder)

Henriksen: Hva med multippelregresjon på vannføring?

Damsleth: Hovedforskjell er at tidsrekkeanalyse tar hensyn til avhengighet i restledd. Slik avhengighet må en ta hensyn til ved signifikans-tester. Hvis ikke blir det for lett signifikans i testene.

Flere variable er ikke kompliserende sålenge innvirkningen av dem kan uttrykkes ved en tranferfunksjon for den variable man vil studere.

Bakke: Hva ligger i restleddene?

Damsleth: Restleddene er avvik mellom modell og måling. Dette er tildels målefeil, som oftest er uavhengig, men inneholder også det som modellen ikke tar hensyn til (F.eks. virkning av andre ting enn vannføring i pH-eksemplet).

En "ideell" modell har restledd ned mot null.

For industriutslipp kan det være treghet i restleddene pga. periodiske effekter som ikke modelleres.

For å forbedre en modell kan en trekke inn nye variable, eller endre tidsavhengigheten. Hvis formålet er prognose kan det være vel så bra å modellere avhengighets-struktur i restledd som å trekke inn nye variable.

Rolf Tore Arnesen, NIVA: Når det gjelder pH-eksemplet: Ville en få bedre modell ved å trekke inn andre variable? Er f.eks. nedbør bedre, sammen med vindretning evt.?

Damsleth: Vanskelig å vite om det dreier seg om sanne kausalrelasjoner eller tilfeldige variasjoner.

Henriksen: Modellen kunne forbedres ved å trekke inn kalsium og sulfat som bestemmende for pH. pH er en avhengig variabel.

Damsleth: For prognoseformål kan vi også bruke sammenhenger som ikke er kausale.

Arnesen: En kan neppe regulere pH ved å variere vannføring.

Damsleth: En kan imidlertid lage prognoser ved å forutsi vannføring (f.eks. ved snesmelting).

Bakke: Er arbeidet med å finne riktig modell vesentlig maskinelt, eller krever det matematisk innsikt?

Damsleth: Identifikasjonsfasen i ARIMA-modellering er et samspill mellom maskin og menneske. En får beregnet endel ting som kan brukes som hjelpemidler, f.eks. autokorrelasjon, som viser sesongsvingninger. Ved å vekselvis estimere en modell og studere residualene kan en nærme seg en løsning trinnvis.

NR har tidsrekke-programmer både for univariate modeller, og for multiple/transferfunksjons modeller opp til 5 variable.

Programmene kan håndtere opp til 6 rekker, hver med 500-600 verdier.

Sigmund Kalvenes, NR: Det gjelder å finne ut hvor mye statistikk som bør brukes i NIVA's daglige arbeide. En må vurdere og prøve ut metodene, gjerne i samarbeid med NR. Hva bør NIVA ha som eget verktøy, og hva er det naturlig å samarbeide med andre om?

Arnesen: En må finne de naturlige grensene mellom instituttene, og få til samarbeid. En må iallfall vite nok til å kjenne til avansert verktøy og bruken av det.

Bakke: Det er bruk for kunnskap om metoder for tidserie-analyse i planleggingen - vi bør velge verktøy før vi måler.

Egil Støren, NIVA: Ang. problemet med ekvidistante målepunkter: Kan avstand mellom målepunkter brukes som forklaringsvariabel?

Damsleth: Problemet kan være av to typer:

A. Variabel avstand mellom målepunktene. Dette gjør det vanskeligere å gjennomføre en identifikasjonsfase - en må velge en modell a priori i større utstrekning.

B. Huller i data. Her kan en eventuelt interpolere data inn i hullene. En forutsetning for dette er at modellen er stabil, dvs. at det er samme korrelasjon mellom alle tall.

Det vil normalt ikke være noe poeng i å bruke avstanden mellom observasjonene som forklaringsvariabel. I enkelte situasjoner, hvor det er sammenheng mellom prosessen og målehyppighet, f.eks. ved episodestudier i spesielle perioder, kan det likevel være av interesse.

Henriksen: Uansett hvor lang tidserie vi har kan det alltid opptre langbølgede svingninger. Det kan vel også være ulik korrelasjons-struktur i ulike deler av observasjonsperioden?

Damsleth: Dette ble analysert i Rendalprosjektet, og en fant forskjeller. Det interessante er ofte nettopp å se om underliggende prosesser forandrer seg.

Morten Svelle, SFT: For å få bedre kjennskap til verktøyet kunne en f.eks. ta et prosjekt, og planlegge hvordan metoden skulle brukes i et 5 års prosjekt. En burde da få til et samarbeid mellom en gruppe saksbehandlere på NIVA og NR.

3 UFORMELLE METODER, VARIABELTRANSFORMASJONER, MULTIVARIAT ANALYSE.

Tid: 1. november 1984 fra kl.9.00 til 11.30

Innledere: Sigmund Kalvenes, NR.
Kim Esbensen, NR.
Lars Kirkerud, NIVA.

3.1 Sigmund Kalvenes,NR: Generell innledning

Statistisk metodikk nyttes om en vid skala av problemer.

Etter hvert har den fått særlig status når det gjelder å formidle resultater og konklusjoner av egne undersøkelser til en kritisk vitenskapsverden av kolleger. Ved hjelp av statistikk kan en formulere vurderbare premisser for konklusjoner. Blir disse akseptert, kan påliteligheten i konklusjonene kvantifiseres, og man har nådd en slags objektivisering i presentasjonen.

Tukey kaller denne statistikken for konfirmerende statistikk. Det å kunne bruke dette verktøyet er ofte et godt mål.

For å kunne bruke statistikk på denne måten må man ofte ha stor kunnskap om sitt spesifikke mål på forhånd. Denne kunnskapen må være rimelig formaliserbar, slik at matematiske, realistiske modeller kan formuleres.

Som alle vet har vi ofte mye kunnskap om vårt problem som det er vanskelig å formalisere, det vi noe diffust kaller "erfaring".

Ganske ofte vet vi også forbausende lite om det problemkompleks vi undersøker. For å komme videre samler vi inn data, og leter etter verktøy til å analysere dem.

Tradisjonelle (eller litt galt "elementære") statistikkpakker tilbyr metoder og teknikker i hovedsak tilpasset den konfirmerende statistikk. Brukes metodene herfra ukritisk i situasjoner hvor man vet lite om problemet, kan man komme til å presentere konklusjoner med en pålitelighets- status som er uberettiget. Man kan naturligvis bruke dette verktøyet også i slike mer diffuse situasjoner, men bruken må være annerledes.

Denne bruken av verktøyet må heller gjenkjennes som det Tukey kaller "explorative statistics", eller oppdagende statistikk.

Problemet her er ikke å få kvalitetsvurdert konklusjonene man er kommet til, men heller å oppdage sammenhenger. Dvs. vi er ute etter å bli oppmerksom på mulige sammenhenger i problem- komplekset basert på analyse av data.

Eksempler på slike metoder er såkalt deskriptiv statistikk, med histogrammer og kurver. Dette åpner for muligheter og gir ideer, men overbeviser ikke noen, fordi vi vet at usikkerheten i materialet er stor, og at vi på forhånd vet lite om den.

I den noe nyere tid er det utviklet en mengde metoder av vesentlig mer avanserte art, som kan hjelpe en til å strukturere data på ulike vis med tanke på å få hjelp i sin faglige spekulasjon.

Kluster-analysen utgjør her et ytterpunkt. Det er en nesten "modell-fri" metode, som gir forslag til klassifikasjon og gruppering. Den er ikke objektiv, forskjellige metoder gir forskjellige løsninger, dvs. forskjellige strukturer.

I en mellomstilling mellom Kluster-analyse og tradisjonell statistikk finner vi Diskriminant-analyse, Faktor-analyse og Prinsipal-komponent-analyse.

I dag skal vi se litt nærmere på Prinsipal-komponent-analyse, som riktignok forutsetter en slags modeller, men en type meget "mykere" modeller enn før; modeller som utvikles i samspill med data.

3.2 Kim Esbensen, NR: Om multivariat teknikk, bruk av Prinsipal Komponent analyse

Oftest finnes det mange variable som kan brukes til å studere et problem. Et problem med stringent konfirmerende statistikk er at en må velge en eller noen få variable som en vil studere. De multivariate teknikker tillater bruk av alle tilgjengelige variable. Resultatene kan senere kombineres med stringente teknikker, og gi pålitelighetsgrad i svarene.

Multivariat analyse kan visualiseres ved kart, plott og tabeller.

Utgangspunktet for en multivariat analyse er en tabell, eller matrise, av data. Tabellen viser ingen sammenhenger, men multivariat data-analyse illustrerer strukturer i materialet. Fra tabellen henter vi fram informasjon (data struktur), systematisk informasjon (trender) og informasjon om kovarians.

Data kan betraktes som bestående av to komponenter: systematisk struktur (en data modell) og usystematisk støy. Vi antar at de underliggende mekanismer vi er ute etter å beskrive kommer til uttrykk i den systematiske strukturen.

Som første skritt i en analyse vil vi ofte gjøre plott for å se sammenhenger:

- plotte hver variabel mot tid

- plotte variabel nr. i mot variabel nr. k.

Hvorfor så ikke plotte flere variable mot hverandre? Det er nettopp hva som gjøres i multivariat data-analyse, se fig.9.

Hvis vi f.eks. vil klassifisere observasjoner i to grupper, vil vanlig univariat analyse innebære at vi prøver med en variabel av gangen. Noen variable kan da vise seg å være bra til å skille de to gruppene fra hverandre, mens andre er dårlige, se fig.8.

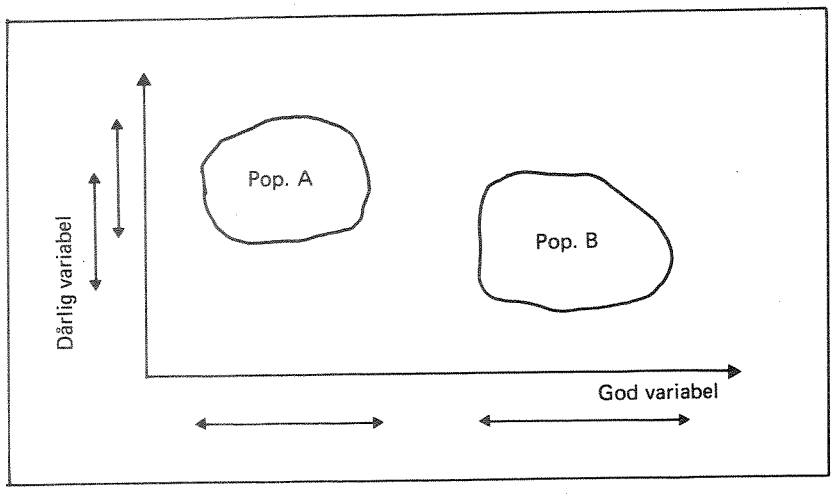


Fig. 8.Eksempel på gode og dårlige diskriminerende variable

Det er imidlertid ikke alltid mulig å finne en "god" variabel, og dette er derfor generelt en dårlig strategi. Multivariat teknikk bruker alle variable samtidig, se fig.9.

Projeksjon:

Hvis vi har observasjoner som består av n variable, kan vi tenke oss observasjonene som punkter i et n-dimensjonalt rom. Vi kan da finne det plan som minimerer avstand til punktene i dette rommet. Projeksjonen av punkter inn på dette planet gir den største variasjon vi kan finne i punktsvermen. Projeksjonen = bilineær modellering.

Punktenes plassering i planet representerer modellen (den systematiske informasjon).

Vi kan plotte objektene i variabel-rommet, men vi kan også snu matrisen rundt (transponere), og plotte variablene i objektrommet. I objektrommet finner vi en akse for hvert observert objekt i datamaterialet. Et punkt defineres av alle verdiene for en variabel, ved at verdi nr. i er satt av langs den aksen som gjelder objekt nr. i. Et slikt plott vil vise hvordan variablene er korrelert med hverandre: variable med stor korrelasjon vil ligge nær hverandre i objektrommet.

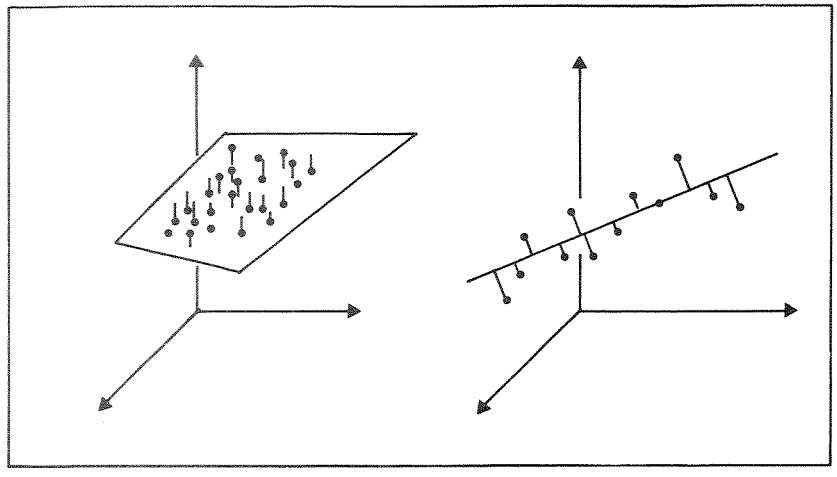


Fig. 9.Eksempel på projeksjon av punkter inn på et plan og linje

Bilineære projeksjoner:

Faktor analyse/ Prinsipal Komponent analyse
Korrespondanse analyse
Eigen-vector analysis
Singular value decomposition.
..

For å operasjonalisere data-analysen går vi ut fra disse forutsetninger:

Likhet mellom prøver (data-vektor)
↓
Likhet mellom objekter
↓
Geometrisk likhet (vektoriell kovarians-struktur)

I SIMCA-metoden (Soft Independent Modelling of Class Analogy) modellerer vi en punktsverm som lavt-dimensjonale modeller, f.eks. linjer + et konfidens-volum (en sylinder), eller en mer kompleks modell, f.eks. et plan i rommet omgitt av en "konfidenskasse". Med mange dimensjoner får vi en "hyperkasse", men matematikken er den samme, se fig.10.

Analysen vil vise hvor mange dimensjoner som er nødvendig.

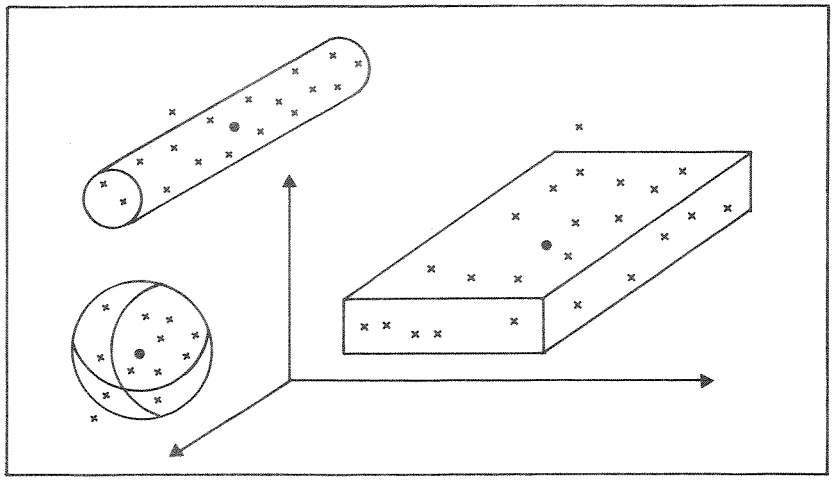


Fig. 10.Eksempler på punktsverm med "konfidenskasse"

Modellen har støy, som uttrykkes ved konfidensvolumet. Støyen bør være hvit (normalfordelt), rundt modell-linje eller -plan etc.

Programmet finner riktig kompleksitet (signal/støy forhold).

En modell av en objektgruppe, med konfidensvolum, kan brukes som klassifikasjonsredskap.

Ved hjelp av modellen lager vi dataklasser. Nye prøver kan sammenlignes med klassene, og kan gi vurdering av f.eks. endringer i miljøet.

Klassifiseringen er asymmetrisk: Dersom en bestemt betingelse er oppfylt (innenfor konf.volum) antas prøven å tilhøre en klasse. Ligger prøven utenfor klassen er den i en udefinert tilstand (= forskjellig fra den klassen).

I vanlig lineær diskriminant-analyse derimot blir nye objekter alltid tilordnet en bestemt klasse (symmetrisk klassifisering).

Ofte er vi ute etter ekstrem-verdier (anomalier, "outliers"). Vi konsentrerer oss da om å beskrive det normale, og uttrykker anomali f.eks. som avstand fra det normale. Punkter som ligger langt fra det normale variasjonsområde betraktes da som anomalier.

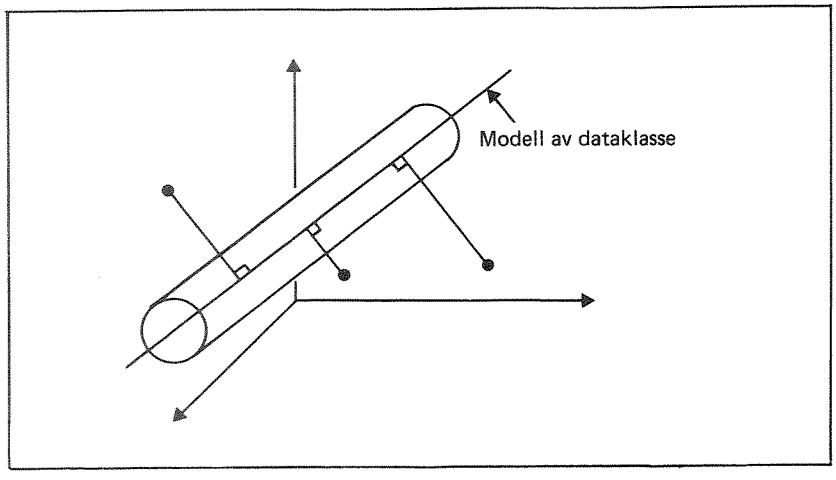


Fig. 11.Eksempler på multivariate anomalier

3.3 Et data-eksempel - miljøgiftdata for fisk

3.3.1 Lars Kirkerud, NIVA: Bruk av klassiske teknikker

Det aktuelle data-materialet er fra overvåkningen av Nordsjø-forurensningen under Oslo-Paris-konvensjonen.

Det er brukt data for torsk fra Oslofjorden. Oppstillingen nedenfor viser når og hvor fisken er fanget:

	Vinteren 1982/83	Høsten 1983
Solbergstrand (Drøbak) Færder	X X	X

For hvert individ er registrert følgende variable:

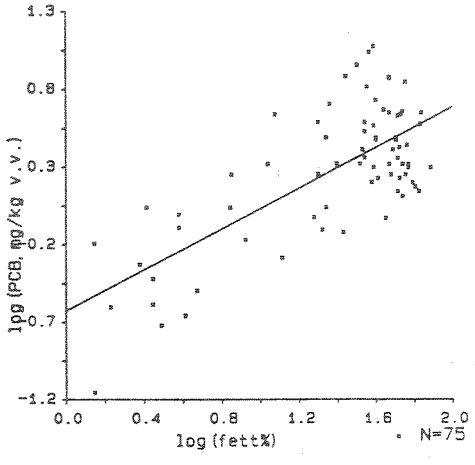
Beskrivende:	Kjønn, alder, vekt, lengde.
For lever:	Vekt fett-prosent, tørrvekts-prosent, innhold av kadmium (Cd) og klorforbindelsene HCB, DDE og PCB.
For muskel:	Tørrvekst-prosent og kvikksølv-innhold (Hg)

For en fullstendig data-tabell, se vedlegg 1.

Tilsvarende data finnes også for blåskjell og vannprøver.

Formålet med analysen var å sammenligne PCB og Hg for å se hvordan de kunne brukes som miljø-indikatorer. Data må da korrigeres for fettinnhold, vekt, lengde etc.

PCB
Torsk



Kvikksølv
Torsk

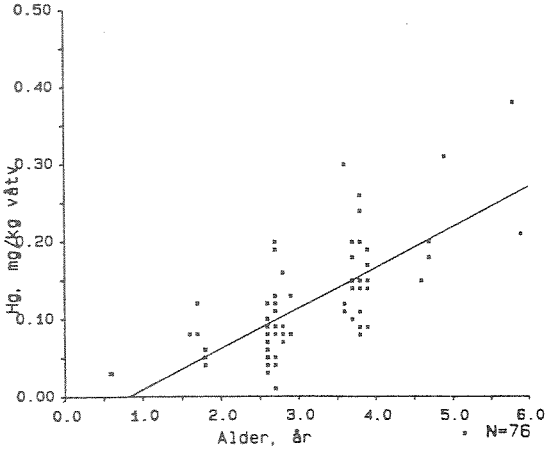


Fig. 12.Eksempler på plott av miljøgiftvariable mot beskrivende variable.

For å se på disse sammenhengene laget jeg plott av to og to variable, PCB mot fett-innhold, Hg mot alder osv. Se figur 12.

Jeg brukte dessuten multippel regresjon på miljøgiftinnhold som funksjon av beskrivende variable.

Analysen viste at kjønn og tid på året hadde liten betydning, og de beskrivende variable som sto igjen var alder, vekt, lengde og fett-% i lever.

Konsentrasjonene av miljøgift ble så ved hjelp av regresjonsligningen korrigert for disse variablene, slik at verdiene ble omregnet til å gjelde for 1 kilo's fisk, med gitte normal-verdier for lengde etc.

Korreksjonen ble gjort for hvert individ etter følgende formel:

$$X_0 = X - (a-a_0)\beta_a - (w-w_0)\beta_w - (l-l_0)\beta_l - (f-f_0)\beta_f$$

hvor

X = log av målt miljøgiftinnhold
a = log av alder
w = log av vekt
l = log av lengde
f = log av fettinnhold i lever

tilsvarende symboler med index $_0$ er normaliserte verdier
(f.eks. w_0 = log av vekt 1 kg)

og β-verdiene er koeffisienter i regresjonsligningen.

De korrigerte resultatene ble så sammenlignet med t-test, for å se på eventuelle forskjeller mellom steder og tidspunkter.

Resultatene viste ingen signifikante forskjeller, hverken mellom de to stedene eller fra det ene tidspunktet til det andre.

En fullstendig redegjørelse for data-analysen er gitt i (NIVA 1985).

3.3.2 Kim Esbensen: Analyse med multivariat teknikk

Det samme datasett er også analysert med multivariat teknikk.

Jeg fikk de samme data som Lars Kirkerud hadde brukt, men ingen konklusjoner.

Observasjoner var a priori delt i tre grupper:

A. Færder 1982
B. Drøbak 1982
C. Færder 1983

10 variable ble brukt i den multivariate analysen.

Et første to-dimensjonalt plott i projeksjonsplanet for den tilpassede modellen viste at gruppe C avvek fra A og B, mens A og B var blandet.

Plottet viste også at 4 prøver fra gruppene (C,A,A,B) henholdsvis avvek fra resten. De avvikende prøvene ble tatt bort, og materialet analysert på nytt.

De to første prinsipal-komponentene forklarte henholdsvis 37% og 21% av variansen i materialet i denne nye analysen.

Noen variable (1 og 9) viser variasjon i overenstemmelse med variasjon i lengde og vekt (variabel 2 og 3).

Prinsipal komponentdekomponering gir uavhengig varians i de to retningene (kovarians =0 mellom retningene).

Variabel 4 varierer på tvers av dette til den ene siden, og markerer altså en "avviks-retning" på tvers av den ovennevnte "vekst-retningen", mens variablene 7,8 og 10 markerte avvik til den andre siden. En kan da f.eks. bruke verdien på variabel 4 dividert på 7,8 eller 10 for å vise hva som representerer dette avvik fra "normal" utvikling i veksten. For variabel 5 og 6 har begge faktorer betydning.

Variablene analysert i objektrommet viser visuelt hvilke variable som har høy vekt i de ulike faktorer (prinsipal-komponenter) (Plott vist på seminaret).

Typisk for metoden er at:

- Data bestemmer modellens kompleksitet, istedet for at en velger en bestemt dimensjon på modellen a priori.
- Utsagn kan kvantifiseres (skåring langs de prinsipale komponenter).
- Modellen kan visualiseres (skåring og "loading plots").

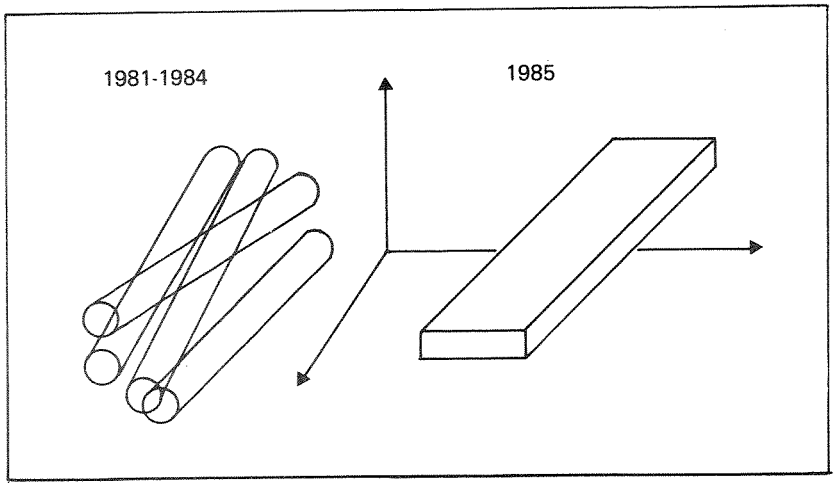


Fig. 13.Eksempel på modeller av ulik kompleksitet

Ved analyse av to ulike deler av et data-sett kan en godt få forskjell i kompleksitet mellom de to modellene, ikke bare i beliggenhet i variabel-rommet, f.eks. som illustrert i figur 13.

De projiserte, komprimerte data en får ut av metoden kan videre brukes i eventuelle statistiske standard-metoder, andre faglige presentasjoner etc.

3.4 Fra den etterfølgende diskusjon

Det ble spurt om tilgjengelighet av slike metoder. SIMCA-pakken finnes nå (jan. 1988) bl.a i Microsoft BASIC under CP/M, samt en svensk FORTRAN versjon som også går på PC. Den er også implementert på NORD-500og VAX i en versjon for "store datasett". Det finnes også en norsk-produsert pakke (UNSCRAMBLER), som i tillegg de her beskrevne metoder også omfatter multivariat regresjon <1> (PLS) m.m. Også denne pakken er beregnet til PC-bruk.

Problemet med hull i datasettet ble tatt opp. Mens standard-pakker oftest kaster bort alle observasjoner med hull, så vil SIMCA-programmet sette inn "modell-verdier" der verdier mangler. Det betyr at alle tilgjengelige observasjoner blir utnyttet så langt som mulig, også de som er ufullstendige.

Det ble spurt etter "kokebok" i bruk av metoden. Det er imidlertid viktig å forstå metoden for å bruke den. Den kan ikke brukes slavisk etter en oppskrift. Til gjengjeld rekker det (langt) med bare en geometrisk forståelse, jfr. avsnitt 3.2.

Et videre samarbeid ble diskutert, og det ble foreslått å følge opp seminaret med kurs og opplæring i bruk av endel metoder, basert på konkrete problemer.

<1> Generalisering av vanlig lineær regresjon (se avsnitt 4.2) til flere, samtidige y-variable m.v.

4 PROBLEMER RUNDT ANVENDELSE AV KLUSTER- OG REGRESJONSANALYSE

Tid: 27. november 1984 fra kl.9.00 til 11.30

Innledere: Sigmund Kalvenes, NR.
Rolf Volden, NR.

4.1 Sigmund Kalvenes, NR: Klusteranalyse

Formålet med klusteranalyse er å avdekke strukturer i multivariable datamengder. Dette søkes oppnådd ved ulike grupperinger av datamaterialet.

Datasettet består av et antall variable som er målt for et antall objekter og er gjerne presentert i tabellform med en linje for hvert objekt og en kolonne for hver variabel:

$$\begin{array}{cccc} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{array}$$

x_{ij} i tabellen er en målt/registrert verdi for variabel nr. j i tilknytning til objekt nr. i.

Ved klusteranalyse søker en å gruppere slik at variable og/eller objekter som ligner hverandre mest mulig i en gitt forstand, blir holdt sammen. Variable som blir gruppert sammen, viser likeartet variasjon fra objekt til objekt. Objekter i samme gruppe er karakterisert ved at at variabelverdiene for objektene ligger nær hverandre. Hva vi mener med "likeartet" og "nær" er imidlertid ikke gitt i utgangspunktet, og friheten til å presisere disse forholdene på

ulike måter kan gi grunnlag for ulike grupperingsprinsipper og ulike grupperinger for konkrete datamaterialer.

Klusteranalyse konkurrerer med mange andre prinsipper og metoder -og stort sett på vikende front jo mer kunnskap og forståelse om objektene og variablene man sitter inne med på forhånd. Selv med forholdsvis begrenset forhåndskunnskap kan f. eks. faktoranalyse og prinsipal komponent analyse tenkes å være vel så egnet til gruppering av variable som klusteranalyse er det.

Klusteranalyse er i første rekke et datanalytisk hjelpemiddel. Ved liten forhåndskunnskap skal den hjelpe til å se/oppdage strukturer i foreliggende datatabeller. En bakenforliggende hensikt - og håp -er imidlertid at vi derigjennom skal bli hjulpet til nye ideer og innfall og derved kanskje til bedre forståelse av den virkelighet som har gjort at vi har interessert oss for nettopp disse objekter og variable.

I det følgende vil vi begrense oss til å se på gruppering av objekter.

Klusteranalyse klarer alltid å trekke frem strukturer i data-materialer, men verdien av disse kan være både usikker og tvilsom når hensikten egentlig er å finne strukturer i det reelle fenomen eller virkelighet som datamaterialet antas å representere. Verdien av en utført klusteranalyse kan først estimeres i etterhånd ved innføring av utenforliggende momenter som forutforståelse og kobling til andre og beslektede undersøkelser.

Vi kan imidlertid prøve å stille oss i en så gunstig posisjon for etterbehandlingen som mulig ved å nærme oss selve klusteranalysen med en viss forsiktighet.

- Først må vi bruke omtanke ved valg av både objekter og variable for å få til en best mulig representasjon av de interessante fenomener vi vil vite mer om.

- Vi må også prøve å mene noe om hvor stor vekt - relativt sett - hver enkelt variabel skal tillegges.

- Dernest må vi konstruere - velge- et mål for likhet eller avstand mellom objektene.

- Endelig må det velges grupperingsprinsipp eller metode. I dette inngår bestemmelse av likhet/avstand mellom grupper av objekter.

Når det gjelder variabelvalg og -vekting kan det være et problem med ukjent korrelasjon (samvariasjon) mellom variablene. Uten kjemisk innsikt kunne man f.eks. utilsiktet komme til å vektlegge en reell havvannspåvirkning dobbelt ved å la både Na^+ og Cl^- inngå på lik linje med alle andre variable i analysen.

Alle de valg som gjøres før selve klusteranalysen påvirkes av måleskalaen til de enkelte aktuelle variable. Denne kan være:

1. Nominell. F.eks. A, B, C eller "firkantet", "sirkelformet", "elliptisk".

2. Ordinal. Da har vi en ordning av verdiene til variabelen. Vi kan si at x_{ik} er mindre enn x_{jk}, men vi kan ikke si noe om hvor mye mindre den første verdien er enn den siste. (liten, mindre, minst)

3. Intervall. I tillegg til ordning av verdiene får nå differansen $x_{ik} - x_{jk}$ mening.
   Eks.: Temperatur gitt som grader Celcius.

4. Kvotient. Nå gir det også mening å snakke om verdien av brøken x_{ij}/x_{kj}. Eks.: Temperatur gitt som grader Kelvin.

De fleste klustreringsmetodene forutsetter at alle variablene registreres i samme skala. Det er alltid forbundet med noe vilkårlighet å redefinere en verdi i en skala til en "finere", f.eks. å gå fra ordinal til intervall skala. Den omvendte vei er vanligvis mindre problematisk. Man ender derfor gjerne opp med den "dårligste" skalatype blant de typene som forekommer blant de variablene vi har valgt å ta med i datatabellen.

Eksempler på avstands- og likhetsmål:

La A_i og A_j være to objekter med variabelverdier på intervallskala: x_{ik} og x_{jk}. Er antall variable = K, kan aktuelle avstandsmål for avstanden mellom A_i og A_j være bl.a.:

$$\text{Euklidsk: } D(A_i, A_j) = \left[\sum_{k=1}^{K} (x_{ik} - x_{jk})^2 \right]^{1/2}$$

$$\text{Minkowski: } D_2(A_i, A_j) = \left[\sum_{k=1}^{K} (x_{ik} - x_{jk})^q \right]^{1/q}$$

$$\text{Chebychev: } D_3(A_i, A_j) = \max_k | x_{ik} - x_{jk} |$$

Et meget brukt likhetsmål for variable brukes undertiden for objekter, men da med tvilsom tolkning:

$$\text{Pearson-korrelasjon: } S(A_i, A_j) = \frac{\sum_{k=1}^{K} (x_{ik} - \bar{x}_{i.})(x_{jk} - \bar{x}_{j.})}{\sum_{k}^{K} (x_{ik} - \bar{x}_{i.})^2 \sum_{k}^{K} (x_{jk} - \bar{x}_{j.})^2}$$

$$\text{Her er } \bar{x}_{m.} = \frac{1}{n} \sum_{k=1}^{K} x_{mk}$$

I flere NIVA-arbeider hvor klusteranalyse har vært nyttet på biologiske data, har disse vært gitt i nominell skala (binær).

Et likhetsmål som da har vært nyttet, går under flere navn.

PÅ NIVA heter det

Sørensens indeks: $\frac{2c}{a + b}$

Her er: a = totalt antall arter registrert ved en stasjon (objekt)
b = totalt antall arter registrert ved en annen stasjon
c = antall arter som er felles ved de to stasjoner.

Et annet likhetsmål som NIVA har studert for samme type data er

Diee's assosiasjonsmål: $\frac{c}{\min(a, b)}$

Vekting

For klustermetodene, som skal lete frem grupper av objekter som ligger nær hverandre, spiller det ofte stor rolle hvilken relativ vekt som blir tillagt de enkelte variablene. Det kan være viktig å være bevisst dette problemet og tenke gjennom om man har spesiell faglig innsikt som man kan ha nytte av å bruke i denne sammenheng. Egentlig slipper man aldri utenom vektingproblemet, men man finner det ofte bekvemt å velge en av flere standardmåter å vekte på som i en eller forstand gir variablene lik vekt. To måter å gjøre dette på er knyttet til skalering:

- Alle variable skaleres slik at de får samme variasjonsbredde.
- Alle variable skaleres slik at deres standardavvik blir likt.

Det er naturligvis ofte på sin plass å velge slike indirekte vektinger, man må bare være klar over hva man gjør og ikke tro at man derved objektiviserer bort det faglige ansvar.

Klustermetoder deles gjerne inn i byttemetoder og hierarkiske metoder.

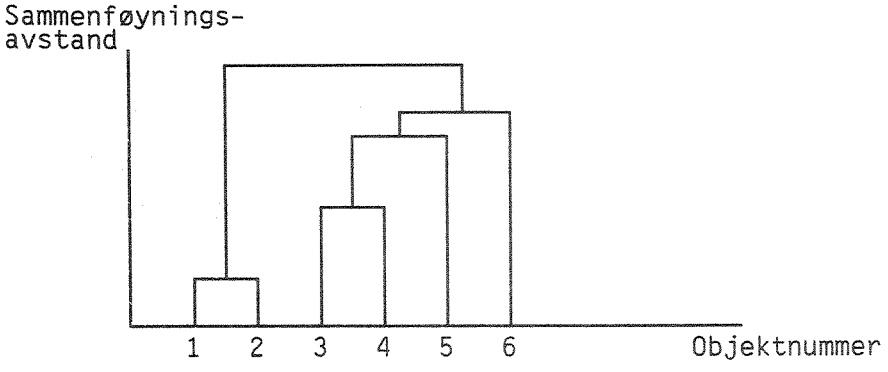
- Byttemetoder starter med en ferdiggjort gruppering som utgangspunkt. Den kan være automatisk generert, eller være basert på et tentativt faglig skjønn. Byttemetodene prøver derefter å forbedre grupperingen etter et valgt prinsipp ved å bytte objekter mellom gruppene. Det kan f.eks. være at metoden prøver å redusere den totale varians innen grupper. Aktuelle metoder er bl.a. McQueen og Konvergent K-means.

- Hierarkiske metoder er vanligvis enten splittemetoder eller sammenføyningsmetoder.

  . Splittemetoder starter ut med å se alle objektene som ett kluster. Dette deles i to slik at de to delene blir mest mulig forskjellige etter et nærmere bestemt mål. Delingen fortsetter så etter samme prinsipp inntil hvert enkelt kluster består av kun et enkelt objekt.

  . Sammenføyningsmetodene går på en måte motsatte vei i forhold til splittemetodene. Man tar utgangspunkt i det settet av klustre som fremkommer når man sørger for at hvert kluster består av ett og kun ett objekt. Fremgangsmåten videre er at man etter tur slår sammen de klustre som ligner hverandre mest mulig etter et nærmere definert kriterium. Til slutt ender man opp med alle objektene i ett kluster.

  Sammenføyningsmetoden er det vanlig å anskueliggjøre ved et tre som er snudd på hodet og gjerne kalles dendrogram.



Dendrogrammet viser avstanden mellom klustre som blir slått sammen og dermed også rekkefølgen for sammenslåingene ved at sammenføyningsprosedyrene slår sammen de klustre som har minst avstand blant de klustre som til enhver tid eksisterer.

I praksis interesserer man seg ikke primært for hele det hierarkiske tre eller dendrogram, men heller for de klustre som er blitt slått sammen under en viss, gjerne problemavhengig bestemt, avstand. Ved å legge en horisontal linje ved denne sammenføyningsavstand i dendrogrammet får man kuttet de sammenslåtte klustre i delklustre som oppfyller den avstandsbetingelsen man har valgt.

Blant de hierarkiske metodene har sammenføyningsmetodene fått en god del større utbredelse enn splittemetodene, og vi vil i det følgende se litt nærmere på en videre oppsplitting av denne metodegruppen og kort omtale: link metoder, sentroide metoder og optimeringsmetoder.

- Link metoder

  Den vanligste link metode er "Single linkage". Denne metoden definer sammenføyningsavstanden mellom to klustre som den minste avstand som eksisterer mellom objekter i det ene og objekter i det andre.

  "Complete linkage" definerer sammenføyningsavstanden som den største avstanden som eksister mellom objekter i det ene og objekter i det andre.

  Man kan også si at den første metoden interesserer seg for den største likhet som eksisterer mellom objekter fra det ene og fra det andre klustre mens den andre metoden er komplementær i den forstand at sammenføyningsavstanden defineres ved de to objekter som er minst like fra hvert kluster.

Begge metoder slår etter tur sammen de klustre som har minst sammenføyningsavstand.

Mange varianter av link metoder er blitt konstruert, og nye blir stadig dannet, ved å velge ulike kompromisser mellom de to hovedmetodene "Single og Complete linkage". Disse varianter betegnes ofte "Average linkage".

- Sentroide metoder

Disse metodene bygger på at man for hvert kluster danner middelvektorene, kalt sentroider, for klustrene. Forskjellig veiing ved dannelsen av sentroidene gir opphav til ulike metoder. Ved utregning av sammenføyningsavstander og ved sammenføyning av klustre lar man klustrene representeres ved sine klustre.

Ved NR har vi ikke nyttet sentroide metoder i nevneverdig grad, men særlig i biologiske miljøer er metodene mange steder populære.

- Optimerings metoder

Disse metoder definerer gjerne sammenføyningsavstanden som en global størrelse, som en straffefunksjon, definert over alle etablerte klustre.

Den optimeringsmetoden vi på NR har hatt størst erfaring med går under betegnelsen av Wards metode. Den søker ved hver aktuell sammenføyning å slå sammen de to klustrene som gir den minste økningen i den totale summen av kvadratavvikene innen klustre.

De ulike klustermetodene som er tilgjengelig kan gi meget forskjellige resultater på konkrete datamaterialer. Dette så vi under seminaret et eksempel på ved bruk av Single linkage og Wards metode på Kirkeruds fiskedata. Da valg av metode i konkrete tilfeller er noe vilkårlig,

betyr dette at overbevisningskraften ved bruk av klusteranalyse er noe begrenset. Overbevisningskraft er imidlertid ikke primærhensikten ved klusteranalyse. Skal man formidle forskningsresultater på en kontrollerbar måte til andre forskere, er andre metoder mer egnet. Spesielt er klassisk statistikk med usikkerhetsvurderte konklusjoner et sterkt hjelpemiddel. Klusteranalyse derimot kan være til uvurderlig hjelp når man prøver å formulere problemer og hypoteser, som det kan være verd å arbeide videre med, i en forholdsvis ustrukturert situasjon med diffus forståelse, mange variable og mange muligheter. Da kan vi også se det positive at ulike metoder kan gi ulike resultater og dermed flere forslag til strukteringer for videre bearbeiding.

4.1.1 Programpakker

Flere av de store statistiske programpakkene har gode tilbud på klusteranalyse metoder, f.eks. er både BMDP og SAS vel verd å se nærmere på.

På NR har Vidar Berteig utarbeidet en interaktiv programpakke som både gir mange muligheter for metodevalg, og som gir anledning til statistiske karakteriseringer av de klustre man kommer frem til og velger å interessere seg for: middeltall, median, standardavvik min. og max. med mere. Programpakken er benevnt NCLUST og er særlig godt vedlikeholdt på Nord-500. På Nord-100 går en noe redusert versjon, men på mindre datamaterialer kan også denne være tilfredsstillende.

4.2 Rolf Volden, NR: Regresjons-analyse

Utgangspunktet er den lineære regresjonsmodellen:

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_i \quad ; \qquad i = 1,..,n = \text{observasjons-nr.}$$

De avhengig variable y forutsettes altså å bestå av en deterministisk del, uttrykt som en lineær kombinasjon av p uavhengige variable x_j + en støydel ε_i.
At regresjonsmodellen er lineær vil si at β inngår lineært i modellen. X-variablene kan være transformert eller beregnet på grunnlag av flere andre variable.

Det kreves at antall observasjoner n > p+2 for at en estimering av alle ukjente parametre i modellen skal være mulig.

Tilpasningen av en slik regresjonsmodell til et datasett består i å beregne verdier på β slik at modellen passer best mulig til data.

Beregningen bygger på følgende antagelser:

1. $E(y_i) = \beta_0 + \sum_j x_{ij}\beta_j$ for alle i
2. $var(y_i) = var(\varepsilon) = \sigma^2 \stackrel{=}{} \text{konstant}$
3. ε_i er uavhengige av hverandre
4. ε_i er normalfordelt $N(0,\sigma)$

Antagelse nr. 3 holder f.eks. ikke hvis dataene er fra en tidsserie hvor det er autokorrelasjon.

Avvik fra antagelse 4 er vanligvis ikke alvorlig.

Formålet med regresjonsanalysen kan være:

- Prediksjon av y ut fra x ved nye observasjoner.

- Optimere en ytelse (F.eks. ved prosess-styring, hvor x kan være design-variable, og en leter etter den kombinasjon av x som gir høyest y).

- Teste hypoteser på β-verdiene (Dette stiller høye krav til at forutsetningene i modellen skal være oppfylt.)

- Approksimere en funksjon y, dvs. bruke regresjons-ligningen som en 1.ordens tilnærmelse for å beskrive hvordan y varierer med x.

- Deskripsjon av y. Hvis en f.eks. har en tidsserie med sterke variasjoner, kan en modellere års- eller halvårsvariasjon ved hjelp av regresjon, og derved korrigere for disse.

- Kalibrering

- "Forstå dataene"

Det er ofte et viktig spørsmål når data skal skaleres om ved:

Sentralisering, dvs. at gjennomsnitt = 0.

Normaliseres, dvs. at gjennomsnitt = 0 og standard-avvik = 1.

Det avgjørende her er hva slags variable en har.

Hvis en f.eks. skal tilpasse polynomer i x, så bør dataene normaliseres for at en skal unngå numeriske problemer.

Hvis x-variablene er f.eks. blandingsprosenter er det best å ikke sentralisere eller normalisere, for å unngå å gjøre tolkningen av resultatene komplisert.

Estimeringen av β gjøres vanligvis ved minste kvadraters metode. Dette krever at forutsetningene er oppfylt. Hvis de ikke er oppfylt, kan det tenkes at det er bedre å bruke robuste metoder e.l.

Kontroll av modell-forutsetningene:

I forbindelse med regresjonsberegningen bør en vurdere om modellforutsetningene er oppfylt. Det kan gjøres ved:

- Plott, som ofte fungerer bra, og som gir informasjon om hva som eventuelt er galt med dataene,
- Statistiske tester.

Det bør undersøkes om det er kolinearitet i x-verdiene. X-matrisen kan være defekt ved at en av x-variablene er sterkt korrelert med en lineær kombinasjon av noen av de andre x-variablene:

$$\text{F.eks.:} \quad x_p = 1 - \sum_{j=1}^{p-1} x_j$$

Dette gir numeriske problemer selv om sammenhengen bare er tilnærmet. Problemet viser seg ved at partiell korrelasjonskoeffisient er nær 1, og en får ustabile koeffisienter, hvor usikkerheten -> ∞.

Problemet med kolinearitet kan løses på ulike måter:

Ved "Ridge"-regresjon vil en løse problemet på en måte som gir forventningsskjeve estimater.

Ved å ta en komponent-analyse på x-ene kan en erstatte x-ene med komponent-verdier (scorer) og deretter foreta regresjons-analyse av y med komponentverdiene som uavhengige variable.

Hvis x-variablene uttrykker blandingsforhold mellom forskjellige komponenter kan en bare sparke ut overflødige x, som i eksemplet ovenfor hvor summen nødvendigvis er 1.

Hvis kolineariteten skyldes det eksperimentelle design, f.eks. fordi x-verdiene er for begrenset, kan en gjøre flere observasjoner på nye nivåer av designet, slik at kolineariteten forsvinner.

Når regresjonsberegningen er gjort, bør residualene $\hat{\varepsilon}_i$, dvs. avvik mellom målte y_i og beregnede $\hat{y}_i$, plottes mot beregnet $\hat{y}_i$. Det er en kontroll på om antagelse 1 og 2 ovenfor er oppfylt. Hvis de er det, skal avvikene ligge tilfeldig fordelt rundt verdi 0, uten sammenheng med y.

Hvis residualene viser en økende tendens med økende y, tyder det på at y bør transformeres, f.eks. ved en log-transformasjon.

Hvis residualene varierer systematisk som en funksjon av y tyder det på ikke-lineære relasjoner mellom y og x-variable.

Plott av x_{ij} mot $\hat{\varepsilon}_i$, ett for hver j, kan være til hjelp for å vurdere om antagelse 1 er oppfylt.

Normalfordelingsplott av $\hat{\varepsilon}_i$ gir grunnlag for å vurdere om antagelse 4 er oppfylt. Avvik fra forutsetningene kan vise seg ved tunge haler og skjevheter i fordelingene.

Dersom observasjonene er gjort som en ordnet serie bør residualene også plottes mot løpenummer i serien. Det kan avsløre avvik fra antagelse 3 (uavhengige obs.).

Testing av residualer kan også gjøres med f.eks. Durbin-Watson estimator.

I tillegg til dette bør en se på om det er "outliers", eller datapunkter som har spesielt stor innflytelse.

Det finnes her 10-12 forskjellige tester, noen for små sampler, andre for store. Det kan være problematisk å vurdere hvilke metoder som skal brukes. (Hawkins).

Hovedkonklusjonen er at regresjonsanalyse er vanskelig hvis en har store krav.

Et spesielt problem er der en har støy ikke bare i y, men også i x-variablene. Da brukes "betinget regresjon". Bruk av slike regresjoner til å predikere y forutsetter at en måler x med samme feil også senere. Det er uenighet i fag-kretser om bruk av regresjon med støy-beheftet x. Regresjonen blir mer usikker.

Det er likevel mulig å bruke regresjon til å se på sammenhenger, dvs. "forstå" data.

For å undersøke hvilke x som er relevante, er det aktuelt å bruke trinnvis regresjon, hvor en prøver seg frem, og suksessivt tar med nye eller kutter ut x-variable inntil prosessen stopper ut fra visse kriterier. Dette er en subjektiv metode, dvs. hvor en må velge mellom varianter. Ulike metoder gir forskjellig svar mht. hvilke x-variable som skal være med i regresjonen som forklaringsvariable til y. Trinnvis regresjon er altså subjektiv på lignende måte som Kluster-analyse.

Momenter fra diskusjon:

Det er viktig å kjenne dataenes bakgrunn når en skal bruke regresjon. Bl.a. må det være faglig mening i å trekke ut en variabel som avhengig variabel y.

LITTERATUR:

1. Box & Jenkins 1970: "Time Series Analysis, Forecasting and Control." Holden Day, San Francisco.

2. Damsleth, E. 1984: "Tidsrekkeanalyse av forurensningsdata - en metodestudie". Rapport nr. 745, Norsk Regnesentral. Oppdragsgiver: Statens forurensingstilsyn.

3. Hawkins, D.U, Bradu, D. and Kass, G.V.: "Location of Several Outliers in Multiple Regression Data Using Elemental Sets", Techometrics Vol.26, No.3.

4. Henriksen, A. (NIVA), Snekvik, E. (DVF), Volden, R. (NR) 1981: "Endringer i pH i perioden 1966-1979 for 38 norske elver". NIVA-rapport nr. 2/81, 0-80006-02, utgitt innenfor "Statlig program for forurensningsovervåkning". Oppdragsgiver: Statens forurensningstilsyn.

5. NIVA 1981: "Rutineundersøkelse i Numedalslågen." Årsrapport 1980. Overvåkningsrapport 11/81. Norsk Institutt for vannforskning. Saksbehandler: Dag Berge. Oppdragsgiver : SFT

6. NIVA 1982: "Rutineovervåkning i Numedalslågen 1981." Overvåkningsrapport 34/82. Norsk Institutt for vannforskning. Saksbehandler: Dag Berge. Oppdragsgiver : SFT

7. NIVA 1983: "Rutineovervåkning i Numedalslågen 1982." Overvåkningsrapport 100/83. Norsk Institutt for vannforskning. Saksbehandler: Dag Berge. Oppdragsgiver : SFT

8. NIVA 1984: "Rutineovervåkning i Numedalslågen 1983." Overvåkningsrapport 150/84. Norsk Institutt for vannforskning. Saksbehandler: Dag Berge. Oppdragsgiver : SFT

9. NIVA 1985: "Overvåkning av PCB, kvikksølv og kadmium i sjøvannsmiljø i Oslofjordområdet 1982-1983." Lars Kirkerud (NIVA), Beate Enger og Kari Martinsen (SI), Tore Håstein og Gunnar Norheim (Vet.inst.). 0-80106. Oppdragsgiver:SFT

10. Pleym, H. 1980: "Rendalsprosjektet. En lokalklimatisk undersøkelse i forbindelse med vannkraftutbygging i Rendalen. Analyse av temperaturdata." NLVF.

11. Sæbø, H. V., 1984: "Statistiske metoder i forurensingsovervåkningen - Eksempler fra analyse av utviklingen i Numedalslågen." Rapport nr. 753, Norsk Regnesentral.

Analyseresultater for torskeprøver fra Drødaksundet (Solbergstrand) vinteren 1982/1983, mg/kg våtv.

Nr.	Dato	Kjønn 1=M 2=F	Alder, år	Vekt g	Lengde cm	LEVER vekt g	LEVER Hg	LEVER Cd	LEVER Se	LEVER DDE	LEVER PCB	LEVER Fett %	LEVER Tørrv. %	MUSKEL Hg	MUSKEL Se	MUSKEL PCB	MUSKEL Fett %	MUSKEL Tørrv. %
1	821020	2	2+	640	40	10	<0.01	0.04	1.26	0.13	2.10	47	56	0.06	0.32	<0.05	0.3	22
2	821020	1	1+	530	38	12	<0.01	0.04	1.23	0.11	1.30	55	61	0.08	0.25	<0.05	0.3	20
3	821020	1	2+	710	42	17	<0.01	0.02	1.21	0.16	2.10	55	57	0.07	0.30	<0.05	0.3	22
4	821020	2	2+	760	42	17	0.04	0.02	1.40	0.17	1.80	57	62	0.09	0.27	<0.05	0.3	21
5	821020	1	2+	740	42	15	<0.01	0.04	1.24	0.17	2.10	59	51	0.10	0.25	<0.05	0.3	23
6	821020	2	2+	790	44	19	0.02	0.02	1.21	0.16	1.60	62	60	0.05	0.26	<0.05	0.4	22
7	821020	1	2+	710	44	12	0.04	0.04	1.16	0.37	4.40	54	51	0.05	0.31	<0.05	0.4	23
8	821020	1	2+	900	45	30	0.03	0.02	1.12	0.17	2.00	59	61	0.08	0.30	<0.05	0.4	22
9	821020	1	2+	840	44	26	0.01	0.04	1.48	0.14	1.40	67	62	0.03	0.25	<0.05	0.4	22
10	821020	2	2+	730	43	19	0.03	0.03	1.17	0.12	1.50	64	62	0.04	0.26	<0.05	0.4	22
11	821020	2	2+	1140	51	19	0.05	0.02	1.37	0.35	4.70	44	50	0.08	0.28	<0.05	0.2	21
12	821020	2	2+	1020	48	17	0.04	0.08	0.95	0.32	3.40	35	46	0.09	0.26	<0.05	0.2	20
13	821020	1	2+	1070	51	17	0.04	0.07	1.34	0.74	7.60	47	50	0.09	0.29	<0.05	0.2	22
14	821020	2	2+	990	49	22	0.05	0.11	1.60	0.75	3.80	68	57	0.09	0.24	<0.05	0.2	22
15	821020	2	2+	1270	53	20	0.09	0.04	1.15	0.52	5.40	40	51	0.12	0.26	<0.05	0.2	21
16	821105	2	2+	904	41	19	0.04	0.04	1.77	0.18	2.00	39	52	0.08	0.32	<0.05	0.5	21
17	821105	1	4+	1717	58	-	0.05	0.04	1.43	0.37	3.90	35	64	0.15	0.32	<0.05	0.5	23
18	821105	1	3+	1678	61	-	0.13	0.06	1.13	0.16	2.10	25	67	0.11	0.35	<0.05	0.5	22
19	821105	2	2+	817	46	-	0.07	0.10	1.98	0.39	5.10	23	39	0.08	0.38	<0.05	0.5	20
20	821105	2	3+	2252	63	-	0.16	0.04	1.84	0.66	7.70	28	62	0.30	0.42	<0.05	0.5	22
21	821105	2	3+	1359	58	-	0.10	0.17	1.72	0.86	11.00	37	46	0.12	0.35	<0.05	0.5	20
22	821105	-	0+	64	-	-	0.05	0.02	-	0.07	0.94	45	-	0.03	0.32	<0.05	0.2	20
23	821215	1	3+	3711	75	61	<0.01	0.02	1.42	0.30	2.80	58	63	0.15	0.27	<0.05	0.2	21
24	821215	1	3+	1750	60	40	0.05	0.03	1.73	0.55	7.50	47	60	0.14	0.43	<0.05	0.2	21
25	830223	2	5+	3500	74	-	0.25	0.13	1.55	1.22	12.00	39	54	0.21	0.40	<0.05	0.2	24
26	830223	2	3+	2640	69	-	0.13	0.04	1.73	0.69	7.10	57	55	0.15	0.37	<0.05	0.2	23
27	830223	2	3+	2290	69	-	0.15	0.06	3.02	0.70	6.60	36	54	0.14	0.35	<0.05	0.2	24

Analyseresultater for torskeprøver fra Færder vinteren 1982/1983, mg/kg våtvekt.

Nr	Dato	Kjønn 1=M 2=F	Alder, år	Vekt g	Lengde cm	LEVER Vekt, g	Hg	Cd	Se	DDE	PCB	Fett %	Tørrv. %	MUSKEL Hg	Se	PCB	Fett %	Tørrv. %
1	830202	-	1+	600	39	-	0.04	0.04	-	0.06	1.10	22	41	0.04	0.25	<0.05	0.1	19
2	830202	1	2+	460	37	-	-	0.13	-	<0.05	0.79	21	-	0.09	0.57	<0.05	0.1	19
3	830202	1	1+	480	37	-	0.14	0.06	-	0.15	1.80	20	-	0.06	0.36	<0.05	0.1	20
4	830202	1	1+	600	39	-	0.01	0.05	2.08	0.07	0.95	19	-	0.06	0.30	<0.05	0.1	19
5	830202	2	2+	670	43	-	<0.01-	0.11	-	<0.05	0.52	13	28	0.08	0.52	<0.05	0.1	30
6	830301	1	3+	910	46	-	0.13	0.15	2.06	0.10	1.60	38	-	0.09	0.40	<0.05	0.1	17
7	830301	2	2+	720	44	-	-	0.14	-	0.05	0.76	27	-	0.08	0.46	<0.05	0.1	18
8	830301	2	2+	1020	47	-	0.10	0.06	1.47	0.24	2.30	35	37	0.13	0.46	<0.05	0.1	18
9	830202	1	1+	780	44	-	0.04	0.04	0.88	0.23	2.70	53	62	0.05	0.37	<0.05	0.1	22
10	830202	2	3+	1110	50	-	0.13	0.12	1.94	0.30	3.90	20	36	0.15	0.44	0.06	0.1	18
11	830202	2	3+	1090	50	-	0.05	0.08	1.75	0.27	3.10	22	35	0.20	0.52	<0.05	0.0	20
12	830202	1	2+	1080	48	-	0.10	0.06	1.69	0.11	2.00	77	35	0.07	0.46	<0.05	0.0	21
13	830301	2	3+	1230	53	-	0.15	0.14	-	0.08	1.10	7	-	0.17	0.34	<0.05	0.0	19
14	830301	1	2+	1040	50	-	0.10	0.07	1.62	0.19	2.60	34	33	0.08	0.52	<0.05	0.0	19
15	830202	2	3+	1790	59	-	0.12	0.06	1.40	0.32	4.50	69	46	0.09	0.38	<0.05	0.0	20
16	830202	2	2+	1720	56	-	0.08	0.04	0.87	0.13	1.70	41	52	0.16	0.28	<0.05	0.4	21
17	830202	2	3+	1590	58	-	0.08	0.08	1.49	0.31	3.70	39	46	0.15	0.53	<0.05	0.4	21
18	830202	2	3+	2090	58	-	0.09	0.05	0.94	0.18	2.10	33	46	0.09	0.30	<0.05	0.4	21
19	830202	2	3+	1900	61	-	0.11	0.13	1.92	0.37	4.40	12	29	0.24	0.49	<0.05	0.4	21
20	830202	2	3+	2410	65	-	0.03	0.01	1.00	0.22	2.30	52	54	0.08	0.32	<0.05	0.4	22
21	830301	2	4+	2610	67	-	0.18	0.21	2.60	0.16	2.10	11	19	0.31	0.39	0.06	0.2	18
22	830301	2	3+	3350	69	-	0.20	0.04	1.74	0.39	4.60	55	52	0.19	0.53	<0.05	0.2	21
23	830202	1	3+	3680	72	-	0.11	0.05	2.31	0.35	3.10	51	56	0.14	0.44	<0.05	0.2	22
24	830202	1	3+	2700	66	-	0.07	0.03	0.93	0.26	3.00	51	53	0.26	0.40	<0.05	0.2	22
25	830202	2	3+	2700	67	-	0.09	0.03	1.92	0.29	2.60	49	56	0.11	0.45	<0.05	0.2	23
26	830202	2	5+	3400	77	-	0.14	0.12	1.47	0.96	9.10	32	43	0.38	0.40	<0.05	0.2	20
27	830202	1	1+	430	35	-	-	0.13	-	-	-	-	-	0.04	0.31	<0.05	-	20

Analyseresultater for torskeprøver fra Færder høsten 1983, mg/kg våtvekt.

Nr	Dato	Kjønn 1=♂ 2=♀	Alder år	Vekt g	Lengde cm	LEVER Vekt g	LEVER Cd	LEVER HCB	LEVER DDE	LEVER PCB	LEVER Fett %	LEVER Tørrv. %	MUSKEL Hg	MUSKEL Tørrv. %
1	831201	2	8.7	7330	100	163.3	0.16	0.13	1.09	14.20	41.0	43.4	1.31	17.8
2	831201	2	2.7	710	39	4.6	0.68	<0.01	0.06	0.81	3.8	27.0	0.08	18.3
3	831201	2	3.7	5000	76	140.0	0.03	0.05	0.19	1.70	53.0	35.9	0.20	20.9
4	831201	1	2.7	540	37	4.1	0.25	<0.01	<0.05	0.47	2.4	-	0.09	18.8
5	831201	2	2.7	500	38	3.9	0.24	<0.01	<0.05	0.25	1.7	40.1	0.12	18.9
6	831201	1	2.7	1400	52	19.7	0.40	0.01	0.06	0.68	8.4	26.5	0.19	19.3
7	831201	2	2.7	730	42	10.2	0.31	0.04	0.18	3.00	40.0	50.2	0.20	19.0
8	831201	2	2.7	795	44	8.1	0.23	<0.01	<0.05	0.64	1.4	30.9	0.13	18.7
9	831201	2	3.7	1620	58	9.0	0.21	<0.01	0.07	0.99	3.8	26.3	0.14	18.5
10	831201	2	2.7	550	36	5.7	0.09	<0.01	<0.05	0.07	1.4	22.7	0.09	16.8
11	831201	1	2.7	1270	50	36.0	0.04	0.05	0.13	1.80	48.0	25.1	0.13	19.1
12	831201	2	2.7	1240	48	8.2	0.12	<0.01	0.07	1.10	7.0	42.0	0.04	17.5
13	831201	1	4.7	3040	70	19.9	0.33	<0.01	0.06	1.10	2.6	25.8	0.18	17.4
14	831201	2	3.7	3050	58	42.2	0.04	0.06	0.31	4.30	52.0	51.9	0.18	18.6
15	831201	2	2.7	320	33	2.9	0.23	<0.01	<0.05	0.26	2.8	-	0.11	17.8
16	831201	1	4.7	2470	64	37.7	0.13	0.07	0.53	4.50	47.0	47.0	0.20	18.2
17	831201	2	1.7	600	40	5.4	0.11	<0.01	<0.05	0.32	4.7	26.1	0.08	18.3
18	831201	2	3.7	1040	48	18.3	0.42	<0.01	<0.05	0.19	3.1	26.8	0.10	18.1
19	831201	1	2.7	710	42	6.2	0.26	<0.01	<0.05	0.38	2.8	25.6	0.12	19.4
20	831201	2	3.7	1620	70	23.3	0.10	0.04	0.26	3.10	40.0	45.6	0.15	19.8
21	831201	2	2.7	650	49	5.8	0.22	0.01	0.17	1.80	7.1	24.4	0.05	18.7
22	831201	1	1.7	650	42	17.2	0.06	0.04	0.08	1.40	52.0	54.1	0.12	18.4
23	831201	1	2.7	500	38	6.4	0.36	<0.01	<0.05	0.22	4.1	29.1	<0.01	19.9