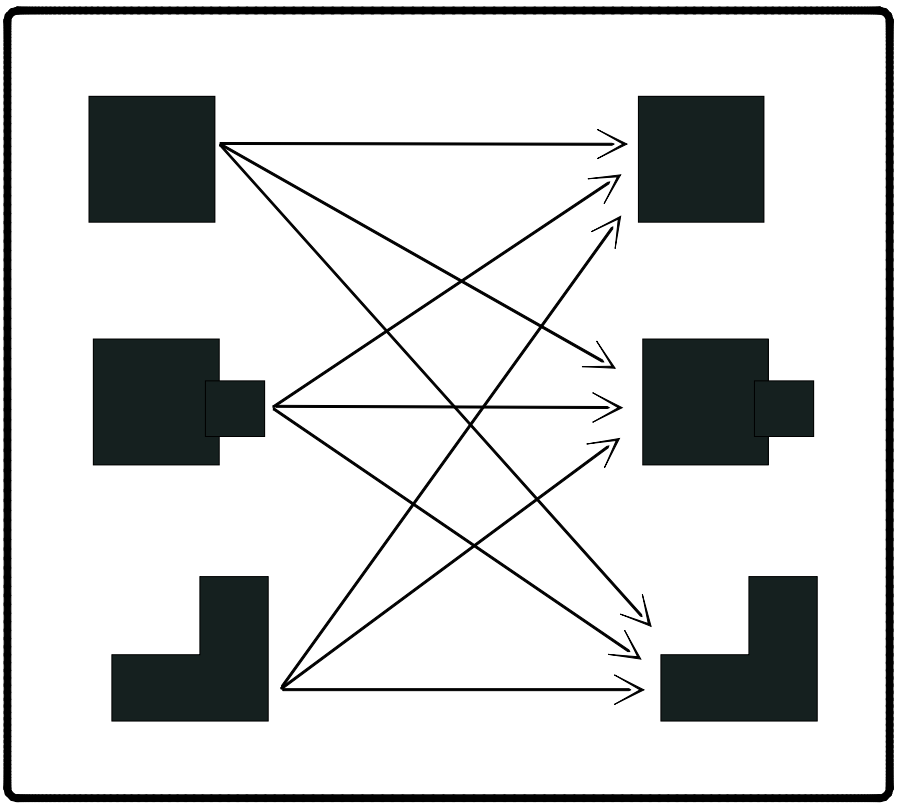
KARTOGRAFISK KOMMUNIKASJON



JAN TERJE BJØRKE

2005

Kartografisk kommunikasjon

Jan Terje Bjørke

5. oktober 2005

Forord

Denne læreboken ble opprinnelig skrevet for undervisning i digital kartografi for sivilingeniører ved NTH og NLH og ble første gang utgitt i 1997. De nevnte institusjonene har endret navn til NTNU og UMB og kartografi er nå en del av geomatikkfaget, *geomatics* på engelsk.

Boken behandler grunnleggende prinsipper for visualisering av geografisk informasjon og gir en innføring i det teoretiske grunnlaget for kartografisk kommunikasjon. Stoffvalget går lenger enn det som nok bør være pensum i kurs som retter seg mot bruk av geografiske informasjonssystemer (GIS). Tanken er imidlertid at boken skal kunne leses på flere kunnskapsnivåer og dels kunne benyttes som en referansebok for praktiserende sivilingeniører.

I den første utgaven fra 1997 hevdes det:

> *Utbredelsen av slagkraftige EDB-baserte tegneverktøy og høyoppløselig visualiseringsutstyr har gjort det mulig for store brukergrupper å produsere kart med et profesjonelt utseende, men som desverre ikke alltid formidler den informasjon de var tiltenkt å formidle. Utfordringen ligger i å benytte sunne kartografiske prinsipper i kombinasjon med god forståelse for hva som er viktig informajson i det foreliggende datamaterialet.*

Dette gjelder også fullt ut i 2005. Jeg håper at boken kan bidra til en bedre forståelse for hvordan kart kan brukes for å analysere utbredelsen til geografiske forekomster, samt gi en dypere forståelse for hvordan lage lett forståelige kart.

Ved neste revisjon av boken vil kartografisk generalisering bli gjenstand for en omfattende behandling. I de senere årene har det foregått en forskning innen dette området som vil bli dekket. Undertegnede tar gjerne imot reaksjoner på boken (emnevalg, skrivefeil, logiske feil, konstruktive forslag til forbedringer o.l.).

UMB 5. oktober 2005

Jan Terje Bjørke[1]

[1]Jan Terje Bjørke er professor i geografisk informasjonsvitenskap ved UMB, 1432 Ås og forsker ved Forsvarets forskningsinstitutt.
epost: jtb@ffi.no

Innhold

Kapittel 1

Introduksjon

1.1 Kartografi

Begrepene kart (av gr. *khartes* = papyrusblad) og kartografi har vært og er fremdeles gjenstand for diskusjon. Drivkraften bak dette tilskrives den teknologiske utviklingen, en økende kunnskap om kartografiske prosesser og framveksten av nye kartografiske produkter. I tiden like etter 2. verdenskrig adobterte FN (The United Nations) følgende definisjon av kartografi [AO93]:

> Cartography is considered as the science of preparing all types of maps and charts, and includes operation from original survey to final printing of maps.

På internasjonalt nivå finnes det en organisatorisk overbygning, ICA (*International Cartographic Association*), som Norge er medlem av (representert ved Norges karttekniske forbund). Da ICA ble opprettet i 1959, var et av de viktigste spørsmål å definere det fagfeltet organisasjonen skulle ivareta [AO93]. Professor F. J. Ormeling, senior, sier i sitt skrift "ICA 1959-1984: The First Twenty-Five Years of the International Cartographic Association":

> Contrary to the broad UN concept of cartography, the Committee of Six concentrated on a more restricted field, exluding surveying and photogrammetry and all primary data gathering by other disciplines such as geology, statistics, etc.

Under sitt møte i Amsterdam i 1967 adopterte ICA følgende definisjon av kartografi [Int73]:

> *Cartography*: The art, science and technology of making maps, together with their study as scientific documents and works of art. In this context maps may be regarded as including all types of maps, plans, charts and sections, three-dimensional models and globes representing the Earth or any celestial body at any scale.

Denne definisjonen av kartografi nevner ikke moderne teknologi knyttet opp mot automatiserte prosesser. Dette skapte etter hvert betydelig diskusjon innen fagmiljøene og Morrison [Mor86] foreslo derfor i 1986 følgende definisjon:

> *Cartography*: An information transfer process that is centered about a spatial data base which can be considered, in itself, a multifaceted model of geographic reality. Such a spatial data base then serves as the central core of an entire sequence of cartographic processes, receiving various data inputs an dispersing various type of information products.

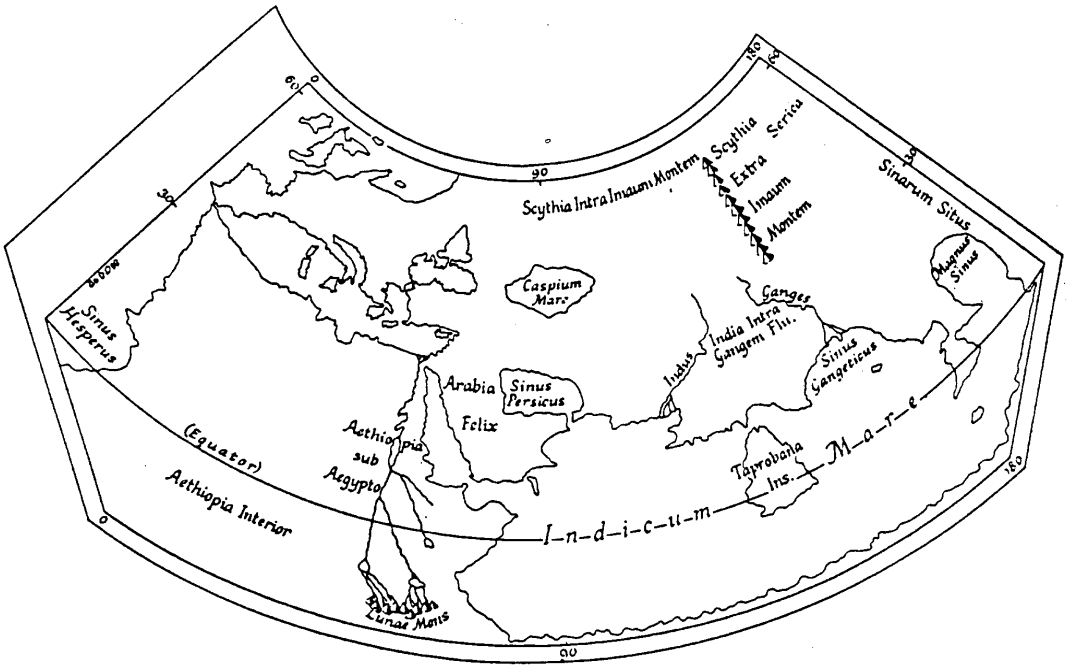
I Morrisons definisjon har automasjon kommet inn som et viktig aspekt. På grunn av de nye muligheter som gis innen grafisk databehandling, har også kartbegrepet blitt diskutert og er nå gitt et moderne innhold. En arbeidsgruppe under ICA (Working Group on Cartographic Definitions), presenterte i 1991 ved Dr. C. Board, følgende definisjoner [AO93]:

> *Map*: A conventionalized image representing selected features or characteristics of geographical reality designed for use when spatial relationships are of primary relevance.
>
> *Cartography*: The discipline dealing with the conception, dissemination and study of maps.
>
> *Cartographer:* A person who engages in cartography.

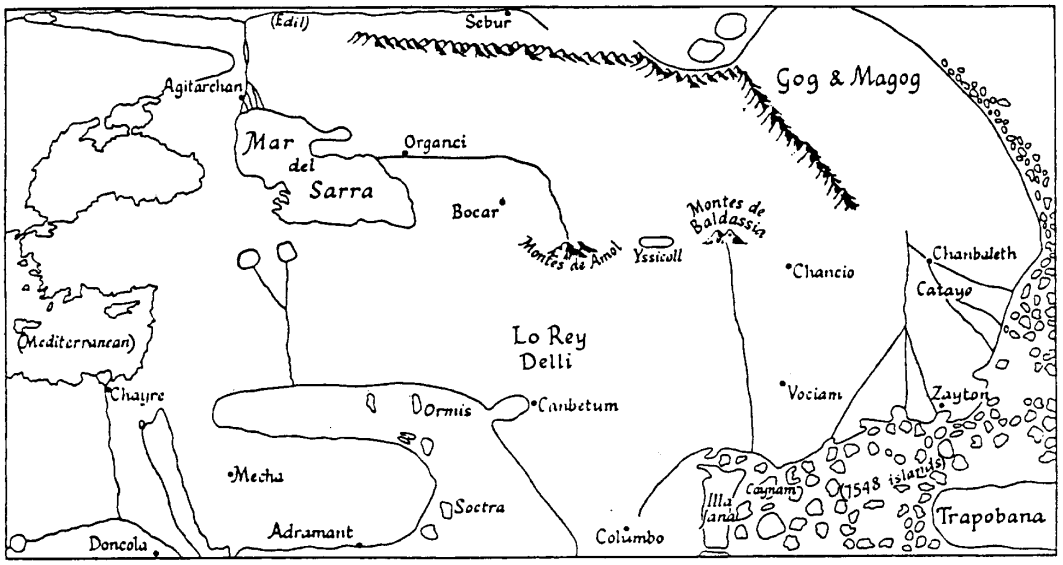
Den ovenfor gitte definisjonen av kart er ganske vidtfavnende og begrenser ikke kartbegrepet til bestemte projeksjoner. Det stilles heller ikke krav om en enhetlig målestokk innen hele bildet. Kriteriet på et kart etter definisjonen, er at bildet inneholder et *utvalg* geografiske forekomster og at bildet er designet slik at de *romlige relasjoner* til de geografiske forekomster kommer tydelig fram. Etter definisjonen er et perspektivbilde over utvalgte geografiske forekomster et kart. Derimot er ikke et fotografisk bilde av et terrengområde et kart, fordi dette bildet ikke inneholder et utvalg av geografiske forekomster. Definisjonen er ikke presis og det vil eksistere gråsoner der vi ikke klart kan si om bildene kan klassifiseres som kart.

1.2 Fra kartografiens historie

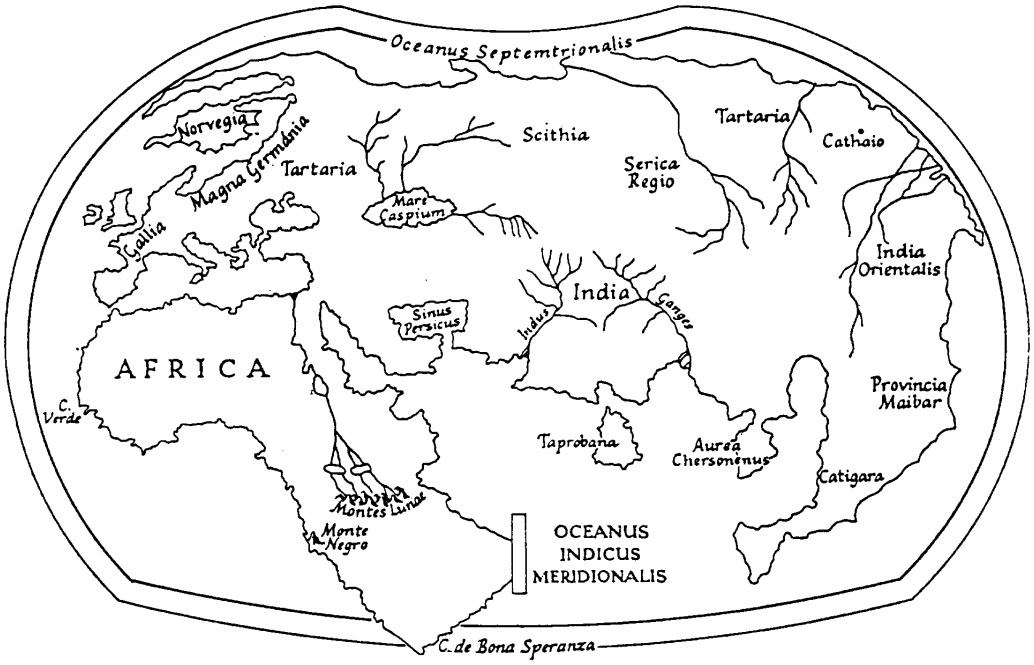
Spor av karttegning finnes allerede tidlig i oldtiden og hos naturfolk. En karttegning basert på vitenskapelige metoder begynner først med grekerne. Anaximander blir vanligvis tilskrevet konstruksjonen av det første greske kart tidlig i det sjette århundre f.Kr. I tidsrommet 400-200 år f.Kr. innførte greske vitenskapsmenn gradinndelingen og de første projeksjonsmetoder. Vårt kjennskap til greske kart skriver seg fra nedtegninger av blant annet geografen og vitenskapsmannen Ptolomaios (ca.



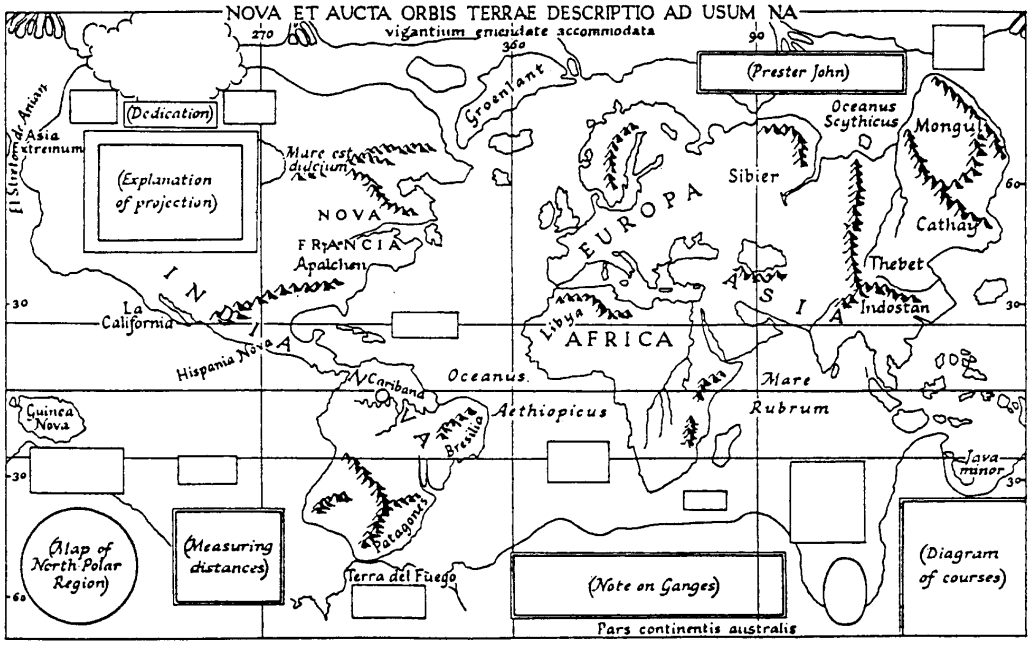
Figur 1.1: Ptolomaios' verdenskart, en Romersk utgave fra 1490



Figur 1.2: Østre del av det Catalanske atlas, 1375



Figur 1.3: Henricus Martellus' verdensatlas fra 1489



Figur 1.4: Mercators verdensatlas fra 1569

140 e.Kr). I disse nedtegningene fortelles det at grekerne lagde kart over sine viktigste reiseruter i det Østre Middelhavsområdet. På disse kartene ble det forsøkt lagt inn noen fundamentale linjer. En slik linje ble lagt langs en av hovedveiene mens en annen ble lagt langs Nilen. Siden disse objektene ikke var rettlinjet, ble feilene på kartet betydelige. Ideen om at jorden er en kule og ikke en flat skive, ble først lansert av filosofer av Pytagoras' (570-500 f.Kr.) skole, men ble senere formulert av Platon (427-347 f.Kr.).

En videre praktisk utvikling fikk karttegningen ved Marinus fra Tyre (ca. 100 år e.Kr.), som innførte gradnett på sine plankart. Marinus trakk meridianene og parallellsirklene som rette linjer som skar hverandre under rette vinkler. Siden han ikke tok hensyn til meridianenens konvergens, kunne bare små landområder kartlegges uten store forvanskninger. På dette punkt ble han kritisert av Ptolomaios (ca. 140 e.Kr.), som utviklet to kartprojeksjoner. Ptolomaios er kjent for sine vitenskapelige arbeider innen geografi og astronomi. Han oppdaget også den såkalte Ptolomaios' sats om forholdet mellom diagonaler og sider i en firkant.

Sin høyeste vitenskapelige utvikling nådde kartografien i oldtiden under Klaudios Ptolomaios. Ptolomaios' sine karter er imidlertid mest kjent som utgaver tegnet flere hundre år etter hans død, se figur 1.1 som er en romersk utgave fra 1490. Romerne bygde sin kartografi på Ptolomaios sine ideer, men med Romerrikets fall ble Ptolomaios' sitt berømte verk *Geographica* etter hvert glemt. Den vitenskapelig baserte karttegning forsvant og det gikk over tusen år før den på nytt ble gjenopptatt. Selv om Araberne gjenopptok Ptolomaios' arbeid på 1100 tallet, nådde ikke deres karttegnere (Edrisi) høyt.

Mot slutten av middelalderen, etter oppblomstringen av de italienske byers skipsfart og kompassets innføring, kommer en periode med de berømte italienske og katalanske sjøkart, de såkalte portulankart (det eldste fra 1311). Bruken av kompasset førte til at portulankartene ble bygd på retningsmålinger. Det mest kjente kartet fra denne perioden er *Catalan atlas* av 1375, se figur 1.2. Atlaset som finnes i det nasjonale biblioteket i Paris, er tegnet på skinn. Kystlinjen er tegnet med svart strek. Kartet inneholder en rekke navn på havner og byer. Viktige havnebyer er skrevet med rødt, de øvrige i svart. Små øyer og elvedeltaer er tegnet i kraftige farger, rød eller gul. Fjell eller grunner er avmerket med små kryss eller prikker i rødt eller svart. Kartene er meget dekorative. Ingen av portulankartene er forsynt med graddeling og meridianenens konvergens er ikke tatt hensyn til.

Etter hvert som de store oppdagelsesreiser tok til, ble behovet for nøyaktige karter påtrengende. Dette kombinert med at Ptolomaios' berømte verk Geographica ble oversatt til latin, gav støtet til fornyelsen av den vitenskapelig baserte karttegning. Kartografien begynte nå å blomstre i Italia, Tyskland og særlig Nederlandene (Otelius' kartsamling av 1570, Mercators virksomhet og den yngere Mercators atlas av 1595). Kartet i figur 1.3 er et verdenskart etter Martellus fra 1489. Vi legger merke til at kartet ikke har graddeling. Derimot har Mercators verdenskart, figur 1.4, fra 1569 graddeling. Etter oppdagelsen av de nye verdensdeler gikk karttegningen hurtig framover, nye og forbedrede projeksjonsmetoder ble introdusert.

Triangulering ble tatt i bruk i oppmålingens tjeneste i 1624-35 og ble allminnelig i løpet av 1700-tallet. Astronomiens framskritt muliggjorde nøyaktige stedsbestemmelser (Frankrike og England på 1600- og 1700-tallet). På 1800-tallet ble reproduskjonsmetodene forbedret (ved fotograferingskunsten).

Det eldste kjente kart utført i Norge er Nordfjorkartet fra 1598, utført etter foranledning av Tycho Brahe. I 1773 ble Norges geografiske oppmåling opprettet, og en systematisk kartlegging av Norges land- og sjøområder tok til.

1.3 Karttyper

Kart kan klassifiseres etter ulike kriterier som kartets bruksområde, kartmålestokk, kartinnhold, kartets persepsjonsnivå osv.. I denne boken vil klassifikasjonen etter persepsjonsnivå være sentral. En klassifikasjon på dette grunnlaget gir

1. lesekart (lokalt persepsjonsnivå),
2. sebare kart (globalt persepsjonsnivå),
3. kommuniserbare kart (globalt persepsjonsnivå over flere informasjonsvariable der informasjonen er sterkt generalisert).

Karttypene ovenfor vil bli utdypet i et senere kapittel.

En annen klassifikasjon legger til grunn om kartet er på digital form eller om kartet er på visuell form (analogt kart). Selv om begrepet digitalt kart er innarbeidet, brukes likevel begrepet i noe ulik betydning: (1) i en snever betydning der vi oppfatter det digitale kartet som en direkte digital versjon av et visuelt kart og (2) i betydningen en geodatabase som ikke nødvendigvis inneholder regler om hvordan dataene skal tegnes ut i form av kart. Begrepet geografisk database (geodatabase) bør imidleritd reserveres for en mere omfattende betydning enn begrepet digitalt kart, i det en geodatabase bør ha databasens kjennetegn som sømløse geodata, strukturerte data, flerbruk, sikkerhet, konsistenskontroll osv..

1.4 Det geografiske rom

Ved karttegning eller ved modellering av geografiske forekomster generelt, er det nødvendig at vi klargjør hvilke egenskaper ved de aktuelle objekter vi ønsker at kartet (modellen) skal representere. Dataenes målenivå er en egenskap til dataene en geodatabase må kjenne til. Vi skiller mellom fire målenivåer.

Nominalnivå På dette nivået er det kun de kvalitative egenskaper til dataene som kommer til uttrykk. Kvalitative egenskaper sier noe om grunnleggende egenskaper til dataene som skog, fjell, vann, vei, elv osv.. Begrepet nominell måleskala kan lett gjøre at vi forbinder nominell med at det benyttes en tallinje, men nominalnivået er basert på mengdelære, et heller enkelt matematisk

konsept. Mellom elementer i en mengde består ingen ordning. Elementer er medlem av en mengde i kraft av sine kjennetegn. En nominell skala introduserer derfor kun kjennetegn som setter oss i stand til å gruppere geografiske objekter.

Ordinalnivå På dette nivået foreligger informasjon om dataene slik at de kan ordnes i en logisk rekkefølge. For eksempel om vi har informasjon knyttet til geografiske enheter av typen høy, middels og lav befolkningstetthet, kan enhetene sorteres på det nevnte grunnlaget. Informasjon på ordinalnivå må ikke benyttes til å beregne differanser eller relative verdier.

Intervallnivå En intervallskala benytter et tilfeldig valgt nullpunkt og en vilkårlig valgt intervallbredde (valg av enhetsavstand). Vi skal imidlertid være forsiktige med å beregne relative verdier for data på intervallnivå, fordi slike tall lett kan feiltolkes. La oss for eksempel anta at vi har temperaturmålinger over et område og at dataene er gitt i grader Celsius. Dersom vi tar to målepunkter og regner temperaturforskjellen som $c_2 - c_1$, får vi et tall vi kan tillegge et fornuftig meningsinnhold. Derimot om vi regner den relative verdien c_2/c_1, får vi et tall som lett kan mistolkes. Det gir for eksempel ikke god mening å si at 10°C er dobbelt så varmt som 5°C. Det relative tallet sier i vårt tilfelle at 10 ligger dobbelt så langt fra 0 som 5. Årsaken til en mulig feiltolkning ligger i at våre målinger ikke refererer seg til et absolutt nullpunkt.

Forholdsnivå (eng. *ratio*) En forholdsskala benytter et absolutt nullpunkt og en vilkårlig valgt intervallbredde (for eksempel lengden av en meter, buen til en gon osv.). Siden dataene på forholdsnivået er referert til et absolutt nullpunkt, gir forholdstall mellom dataene en umiddelbar tolkning. Dersom temperaturmålingene i eksemplet ovenfor var gitt i grader Kelvin, ville vi hatt en referanse til et absolutt nullnivå. Alle målinger utført med vinkelmålende og avstandsmålende instrumenter vil som regel foreligge på forholdsnivået.

Data som foreligger på ordinal-, intervall- eller forholdsnivå, forutsetter også informasjon på nominal nivå. Det er for eksempel lite meningsfullt å snakke om et areal på 12kvKm dersom vi ikke vet hvilken type geografisk fenomen arealet referer seg til. Derimot er utsagnet "12kvKm skog"menigsfylt, men det er først når vi også har informasjon om beliggenheten til objektene at vi kan snakke om geografiske data (stedfestede data).

Vi må i våre kartografiske produkter unngå å gi inntrykk av et høyere målenivå enn det dataene gir grunnlag for. På den annen side er det uten videre klart at vi kan transformere data fra et høyere målenivå til et lavere målenivå. Dette er noe vi ofte gjør i forbindelse med kartografisk generalisering. La oss ta et eksempel knyttet til en teknikk som kalles seriering (beskrevet senere i boken). Metoden baserer seg på at dataene foreligger på nominalnivå. La oss som et eksempel anta at våre data er av typen antall rom i bolighus. Dette er kvantitative data på forholdsnivået. Dataene

kan imidlertid transformeres til kategoriske data på nominalnivå ved at vi definerer at egenskapen til hus ikke er antall rom, men at hus kan ha egenskapene a,b,c osv.. En slik transformasjon gjør at analyseprogrammet vil tro at det jobber på nominelle data, fordi programmet vil ikke kjenne til forskjellen på egenskapene a,b,c osv..

Geografiske data karakteriseres også ved om deres utbredelse er kontinuerlig eller diskret. Overalt på jordoverflaten kan vi for eksempel snakke om at luften har en temperatur, men våre måledata kan være samlet inn ved målinger i spredte (diskrete) punkter. Utbredelsen av geografiske forekomster kan være knyttet til geometriske enheter som punkter, linjer, areal og volum. Vi snakker derfor om dataenes informasjonsbærende enhet. Forestillingen om at informasjonsbærende enhet er 0-, 1-, 2- eller 3-dimensjonal, er ofte knyttet til en skarp (eng. *crisp*) oppfattelse av de geografiske forekomster. Virkeligheten er imidlertid mere kompleks, fordi de fleste geografiske forekomster har flytende (eng. *fuzzy*) avgrensninger og varierende grad av konsentrasjon innenfor den informasjonsbærende enheten.

Kapittel 2

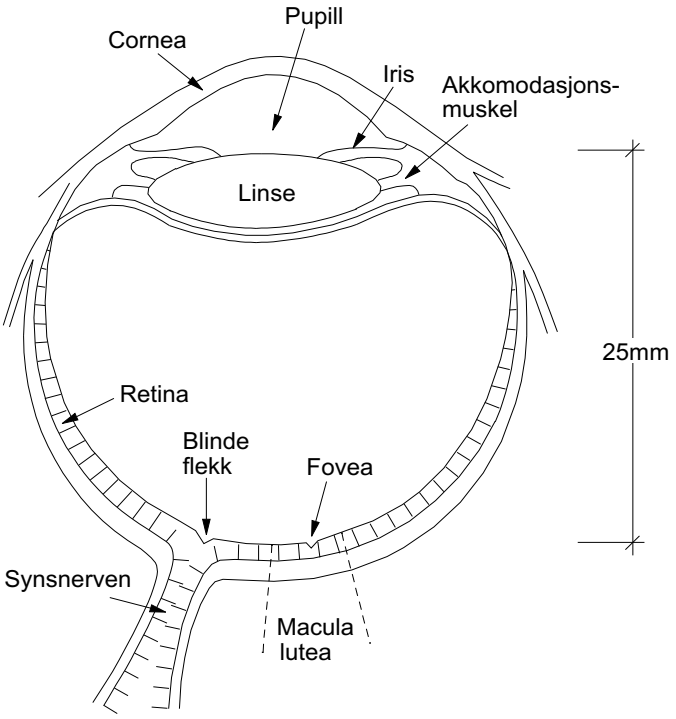
Persepsjon

De deler av dette kaptitlet som handler om fargesyn og fargelære, bygger i stor grad på [CC80], [McL86] og [RMM+95]. Som utfyllende litteratur om persepsjon vil vi anbefale [Kea82].

2.1 Menneskeøyets anatomi

Menneskeøyet, som er illustrert i figur 2.1, er et lukket, blæreformet kameraøye med tykk vegg. De to øynene er bygd som speilbilder av hverandre, men er ellers identiske. Øyet ligger i øyehulen (*orbita*) og er fortil beskyttet av øyelokkene (*palpabrae*). Disse inneholder talgkjertler og er på kanten bevokst av hår (*cilier*), som hindrer at støv og fremmedlegemer kommer inn på øyet. Øyet kan beveges ved hjelp av 6 ytre øyemuskler, slik at de to øynene alltid er rettet mot samme punkt. Derved oppnås samsyn (stereoskopisk syn).

Selve øyeeplet (*bulbus oculi*) ligner et fotografiapparat. Det er formet som en kule med diameter ca. 24 mm. Veggen består ytterst av en fast senehinne (*sklera*, det hvite i øyet") som fortil blir gjennomsiktig og får hornaktig konsistens (hornhinnen, *cornea*). Innenfor senehinnen ligger den mørke årehinnen (*chorioidea*) som fortil går over i strålelegemet (*corpus ciliare*). Denne gir feste for regnbuehinnen (*iris*). Iris er øyets blender. Hullet i iris, *pupillen*, varierer reflektorisk i størrelse alt etter belysningen. I sterkt lys blir pupillen liten, mens den i mørke er stor. Fargen på iris kan variere fra person til person. Irisfargen er arvelig betinget. Mellom hornhinnen og iris ligger det fremre øyekammeret (*camera anterior*), som er fylt med kammervann (*aqueous humor*). I synsaksen, like bak pupillen ligger linsen. Linsens krumning varieres ved hjelp av akkomodasjonsmusklene. Det at linsens krumning tilpasser seg avstanden til det objektet øyet er rettet mot, gjør at øyet kan danne skarpe bilder over et område fra de største avstander til nært hold. På grunn av den kromatiske aberasjon, har ulike bølgelengder ulik brytning i linsen. Differansen i brytningsindeks er størst for farger som ligger i hver sin ende av spekteret. Dette gjør at øyelinsen stadig må endre krumning for å danne et skarpt bilde av en flate som består av



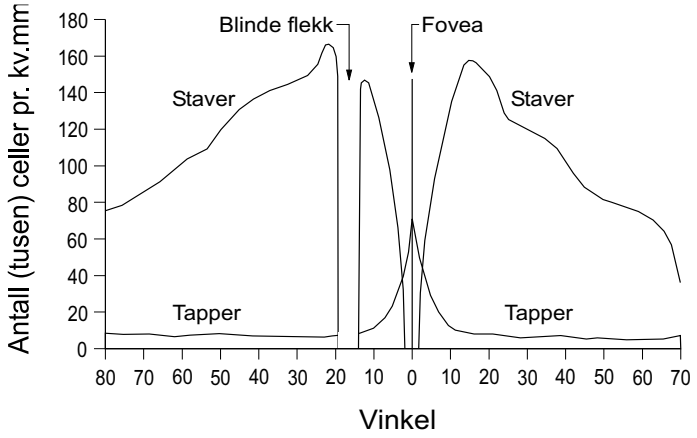
Figur 2.1: Prinsippskisse av menneskeøyet

røde og blå flekker. For en betrakter kan denne muskelaktiviteten bli slitsom. Vi bør derfor unngå fargevalg på kart som gir røde flekker på blå bunnfarge, eller omvendt.

Bak linsen ligger det geleaktige glasslegemet (*corpus vitreum*) som fyller opp det meste av øyets indre. Mellom glasslegemet og årehinnen ligger netthinnen (*retina*), hvor synspigmentene og de egentlige sanseceller (staver og tapper) befinner seg. Ved hjelp av fotokjemiske reaksjoner omdannes her lysstrålenes energi til elektriske impulser, som ledes via mellomliggende nerveceller til gangliecellelaget og videre gjennom deres nervetråder til synsnervepapillen (*papilla nervi optici*) og ut av øyet via synsnerven til hjernen, hvor tolkningen av synsimpulsene finner sted.

Bakerst i øyet har netthinnen en liten gulaktig flekk (*macula lutea*), hvor tappene er særlig tett sammenstilt. Her er derfor synet skarpest. Skades dette området, vil lesesynet gå tapt. I macula lutea finner vi en liten fordypning som kalles *fovea centralis*. Denne inneholder bare tappceller og er derfor det området som gir det skarpeste synet. Det området av retina der bunten av nerveceller føres ut, kalles den blinde flekk. Her finnes ikke lysømfintlige celler, derav navnet den blinde flekk. Figur 2.2 illustrerer fordelingen av staver og tapper på retina.

Netthinnens sanseceller har ikke all sin maksimale følsomhet over en og samme bølgelengde. Sannsynligvis finnes tre forskjellige slags tapper, hver med sitt karakteristiske pigment med tilhørende frekvensområde. Alt etter i hvilket innbyrdes forhold disse tappene stimuleres, vil fornemmelse av farger kunne oppstå. Hver netthinne har omkring 120 millioner staver og 5 millioner tapper, men det finnes mindre enn



Figur 2.2: Fordelingen av tapper og staver på retina

1 million optiske nervefibre som leder til gangliecellene. I de ytre deler av retina er så mange som 600 staver koblet til en enkelt optisk nervefiber, mens i fovea er det nesten en-til-en kobling mellom tapper og fibre. Dette er med og forklarer hvorfor det er maksimal visuell skarphet i fovea og hvorfor stavene er mere lysømfintlige enn tappene. På grunn av en-til-mange koblingen for stavene, blir lyssignalet summert over en større flate. På den annen side gjør en-til-mange koblingen at oppløsningen til stavene blir mindre enn for tappene som har nær en-til-en kobling. I svak belysning kan vi derfor ikke se så fine detaljer som i god belysning. Dessuten har stavene liten fargeseparasjon som gjør at alt ser fargeløst ut i svak belysning.

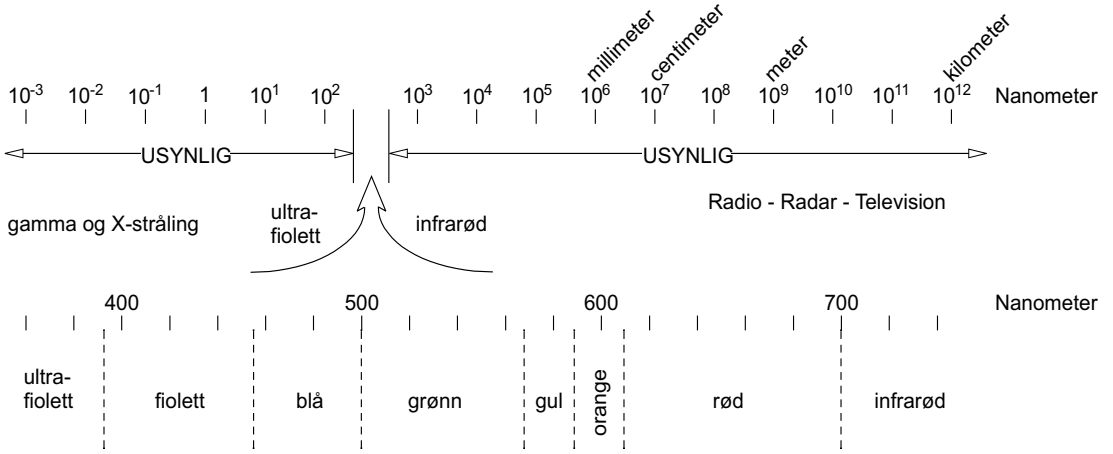
Kart som skal designes for å kunne brukes under ekstreme belysningsforhold, må i særlig grad ta øyets fargefornemmelse, fargeseparasjon og geometriske oppløsning i betraktning. Et forhold som har vært lite påaktet i kartdesign, er at en stor del av befolkningen har svekket fargesyn. For enkelte yrker der bruk av kart og lyssignaler spiller en vital rolle, stilles det krav om normalt fargesyn (flyvere, skipsførere o.l.).

2.2 Fargesyn

Øyet er følsomt for den delen av det elektromagnetiske spektrum som ligger i intervallet $[400, 700]$ nm. En namometer (nm) er 10^{-9} meter. Figur 2.3 illustrerer bølgeområder i det elektromagnetiske spektrum.

2.2.1 Trikromatisk teori

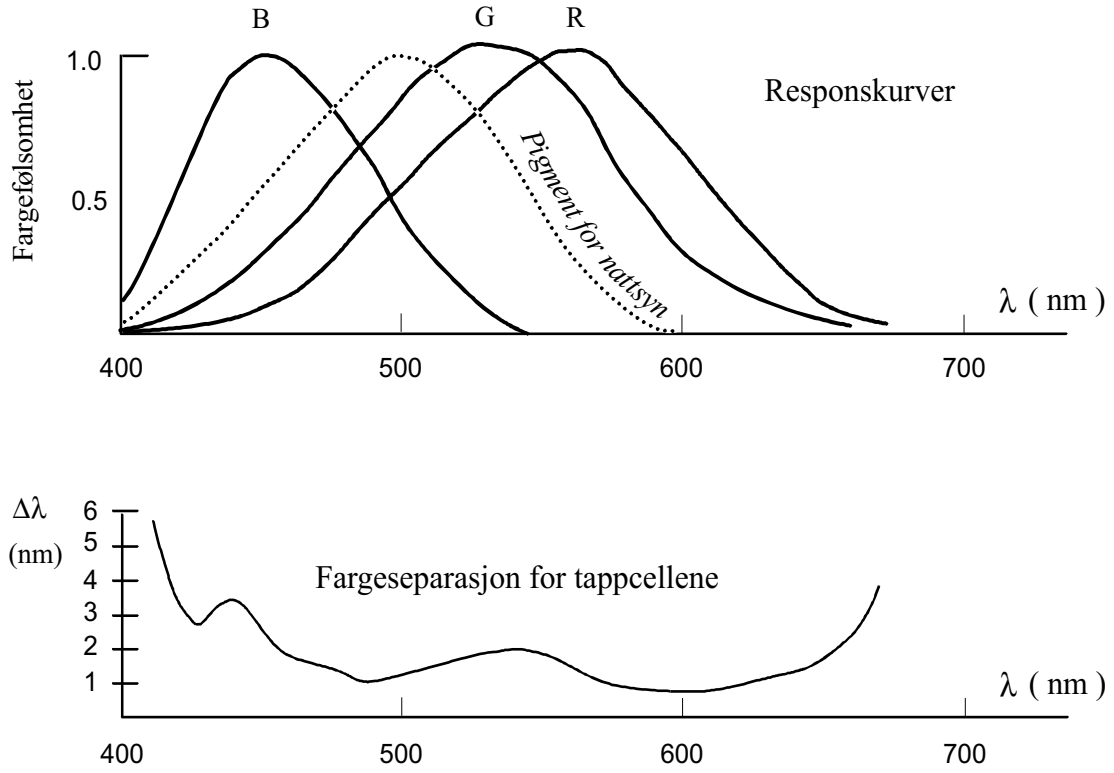
Konseptet om at det finnes tre primære farger og at alle andre farger kan framstilles ved en blanding av disse *(trikromatisk teori)*, ble allerede etablert av Robert Boyle i 1664. Siden den gang har fargesyn vært gjenstand for omfattende studier, men det er først i våre dager at den trikromatiske teori må kunne anses som bevist. En har lenge kjent til at det fantes to ulike typer lysømfintlige celler i netthinnen (staver



Figur 2.3: Bølgeområder i det elektromagnetiske spektrum. Øyet er følsomt for bare en liten del av spekteret.

og tapper), men det var først i 1964 at det ble påvist tre forskjellinge tappceller hvor hver celletype hadde sin spesielle følsomhetskurve. De tre typene celler viste seg å ha maksimal følsomhet omkring 445, 535 og 575 nm ([McL86], side 73). Denne oppdagelsen gav den trikromatiske teori et avgjørende bevis. Figur 2.4 tar sikte på å illustrere responskurvene for øyets lysømfintlige celler. I den øverste figuren er både illustrert kurvene for øyets tappceller (R,G,B) og kurven for øyets stavceller (nattsyn). Kurven nederst i figuren illustrerer hvor små intervaller av λ øyet kan skille. Vi ser at denne kurven korresponderer til (R,G,B)-kurvene. Der hvor (R,G,B)-kurvene stiger bratt og hvor de har god overlapp, finner vi områdene med best fargeseparasjon med omsyn på bølgelengde.

T. Young la i 1802 fram en teori om at det ikke kunne finnes så mange fargefølsomme sentrer i øyet som det fantes fargenyanser. Han antok at det fantes tre slags fargefølsomme partikler i øyet. James Clerk Maxwell (1831-1879) undersøkte spesielt blandingsfarger, og han viste hvorledes alle farger kan oppfattes som en blanding av tre vilkårlig valgte grunnfarger, forutsatt at ingen av disse ikke kan dannes ved en blanding av de to andre. Hans resultater bekreftet Young's teorier. En konkurrerende teori ble i 1878 fremsatt av tyskeren E. Hering. Han antok at det fantes fire grunnfarger: rødt, gult, grønt og blått og at det i øyet fantes to stoffer. Ett stoff som er følsomt for rødt og grønt og et annet stoff som er følsomt for gult og blått, men slik at hver av dem reagerte ved kjemiske prosesser som gikk i hver sin retning for de to farger stoffet var følsomt for. Teorien viste seg å beskrive de fleste fenomener like bra som Young's teori og forskerne hadde inntil 1960 ikke klart å finne avgjørende bevis for hvorvidt tre-farge- eller fire-farge-teorien var riktig. Mens Young's fargeteori synes bekreftet hva angår den måten lyset absorberes på i øyet, er det eksperimenter som tyder på at når informasjonene om farger skal føres videre som nerveimpulser til hjernen, skjer det en omforming slik at man kan klare seg med



Figur 2.4: Responskurver for øyets lysømfintlige celler. R,G,B-kurvene er generaliserte varianter av kurver presentert hos [McL86], side 70.

to nervebaner som formidler både negative og positive signaler. På den måten får man en prosess som er i samsvar med Hering's fargeteori. Det blir da forståelig at begge teoriene synes å gi en riktig beskrivelse av de fenomener som forekommer ved fargesyn.

2.2.2 Fargetonesirkel

Newton viste at hvitt lys ikke bare framkommer som en blanding av alle spektrets farger, men at man også kunne få hvitt lys ved å addere to av spektrets farger. To slike farger kalles komplementære. Newton ordnet spektralfargene etter hverandre langs periferien av en sirkel slik at komplementærfargene ble stående diametralt. I denne sirkelen mangler et område mellom rødt (700 nm) og fiolett (400 nm). Dette området fylles ut med de såkalte purpurfarger. Disse er altså ikke spektralfarger, men representerer fargetoner som framkommer ved en blanding av rødt og fiolett.

Komplementærfarger står i et gunstig forhold til hverandre både når det gjelder fargekontrast og estetikk. Det at komplementære farger framhever hverandre kan utnyttes i kartografisk sammenheng. For eksempel ved at vi velger komplementærfargen til kartets bunnfarge som fargen til kartsymboler vi ønsker å framheve. Siden

blå og orange er komplementære farger, vil for eksempel orange symboler mot en blå bunnfarge gi god fargekontrast.

2.2.3 Defekter i fargesynet

Fargeblindhet kan erverves gjennom sykdom, men skyldes som regel arvelige faktorer. Omtrent åtte prosent av den mannlige delen av befolkningen har defekter i fargesynet. For den kvinnelige delen av befolkningen er tilsvarende tall bare en halv prosent. Grunnen til denne markante forskjellen mellom kjønnene, er de genetiske mekanismer som overfører fargesynet. Det genet som produserer fargeblindhet finnes på X-kromosomet og virker derfor dominant hos menn og recessivt hos kvinner. Det finnes flere former for fargeblindhet, men to av dem omfatter nesten alle former for fargeblindhet.

Akromatopsi Total fargeblindhet forekommer sjelden og følges gjerne av sterkt nedsatt syn (akromatopsi).

Tritanomali Tritanomali eller tritanopi som det også kalles, skyldes nedsatt evne til å se blått. Denne formen for fargeblindhet er lite utbredt.

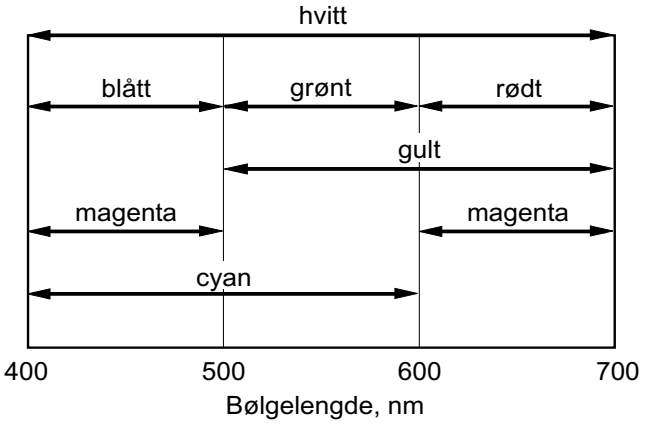
Deuteranomali Deuteranomali skyldes nedsatt evne til å se grønt og er den hyppigst forekommende formen for fargeblindhet. Den omfatter ca. 70% av alle fargeblinde. Denne formen for fargeblindhet spiller en stor praktisk rolle siden det i sikkerhetstjenesten til lands, sjøs og i luften er forutsatt normalt fargesyn for grønt og rødt.

Protananomali Denne gruppen utgjør ca. 30% av alle fargeblinde og skyldes nedsatt evne til å se rødt. Protananomali spiller i likhet med deuteranomali er stor praktisk rolle.

2.3 Fargeblanding

Fargeblandig baserer seg på to hovedprinsipper: (1) *additiv* og (2) *subtraktiv* fargeblanding. Ved additiv fargeblanding blandes lys fra ulike lyskilder, som for eksempel når vi framstiller kart på en grafisk-skjerm. Ved subtraktiv fargeblanding blandes fargestoffer, som for eksempel når vi lager kart på en fargeskriver. Subtraktiv fargeblanding baseres på at fargestoffer absorberer lys. De additive primærfarger kalles rød (R), grønn (G) og blå (B). De subtraktive primærfarger kallses cyan (C), magenta (M) og gul (eng. yellow (Y)). Vi omtaler gjerne additiv og subtraktiv fargeblanding som fargeblanding i henholdsvis (RGB)-systemet og (CMY)-systemet. Figur 2.5 tar sikte på å illustrere forholdet mellom additive og subtraktive primærfarger.

Det er en rekke praktiske problemer knyttet til fargeblanding. Dette gjelder blant annet overganger mellom (RGB)-systemt og (CMY)-systemet. Når kart designes på



Figur 2.5: Forholdet mellom additive og subtraktive primærfarger

en grafisk skjerm, er det ofte ønskelig at de fargene vi ser på skjermen kommer ut på fargeprinteren. Dette betinger en overgang mellom to ganske forskjellige fargesystemer. En utledning av transformasjonsformlene fra (RGB) til (CMY) må blant annet ta i betraktning refleksjonsegenskapene til de subtraktive primærfargene og egenskaper til papiret kartet skal tegnes på. Det finnes flere studier av denne overgangen. Vi viser til litteratur innen grafisk databehandling for nærmere detaljer.

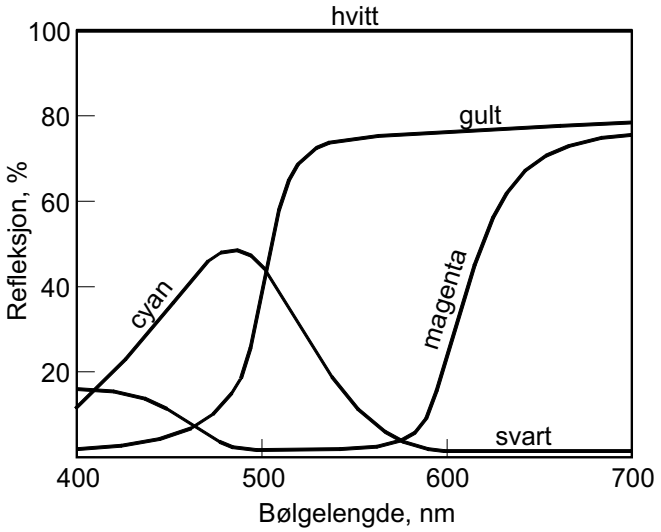
De subtraktive primærfarger er designet ut fra det ideelle mål om at de hver skal reflektere sin 2/3 av det synlige spekteret. Dersom fargepigmentene holdt disse kravene, ville de alle tre i blanding kunne absorbere alt lys. I praksis har det vist seg umulig å designe trykkfarger med de nevnte ideelle spesifikasjoner. Dette gjør blant annet at det ved fargeblanding i CMY-systemet ikke er mulig å lage svart (bare tilnærmet). Svart benyttes derfor ofte som egen trykkfarge. Figur 2.6 viser refleksjonskurvene for en gitt prøve av de subtraktive primærfarger. Dersom det er ekstremt store krav til nøyaktigheten i fargegjengivelsen, som vi for eksempel kan ha for geologiske kart, baserer man seg i mindre grad på fargeblanding og benytter heller et større antall trykkfarger.

2.4 Fargemåling

Det finnes flere systemer for å måle/klassifisere farger. Flere av disse systemene er basert på visuell sammenligning mot fargeprøver, men det er også utviklet et system for objektiv bestemmelse av farger.

2.4.1 Fargeblandingslikning

Moderne metoder for fargemåling bygger i stor grad på teorier utviklet av Maxwell, men studier av Helmholtz og Grassmann må også nevnes. Maxwell stilte opp likninger for kvantifisering av fargeblanding, og han viste at dette ville kunne benyttes til



Figur 2.6: Refleksjonskurver for en gitt prøve av trykkfargene cyan, magenta og gul

en objektiv metode for fargemåling. Et problem en objektiv metode for fargemåling støter på, er at farger er et fysiologisk- psykologisk fenomen. Farger er derfor ikke gjenstand for direkte måling som for eksempel areal, hastighet, pH osv. Den fysiologiske (menneskelige) faktor ved opplevelsen av lys (farger) må derfor innkorporeres i metoder for objektiv fargemåling.

Maxwell's metode anvender tre standardiserte monokromatiske lys som i blanding er i stand til å simulere (gi en betrakter inntrykk av) alle farger som kan lages ved bruk av fargepigmenter. Dette forhold ved metoden ivaretar den fysiologiske faktor. Det består en viss frihet i valg av de tre primære bølgelengder. Kravet er at de skal gi uavhengige fargeopplevelser. Det vil si at den ene ikke skal kunne simuleres ved blanding av de to andre. Fargeblanding kan uttrykkes ved en fargeblandingslikning

$$Q_1 \equiv R + G + B. \tag{2.1}$$

Det viser seg at ikke alle farger lar seg simulere ved likningen på denne formen. Maxwell's vitale bidrag til fargemålingen var at han viste at ved å addere en av primærfargene til venstre side av likningen, vil likevekt opprettes. En farge Q_2 kan for eksempel simuleres ved

$$Q_2 + R \equiv G + B. \tag{2.2}$$

Ved at Maxwell betraktet fargeblanding med utgangspunkt i algebraiske likninger, førte det til uttrykket

$$Q_2 \equiv G + B - R. \tag{2.3}$$

Det kan kanskje virke noe underlig at vi subtraherer lys, men likningen har en enkel tolkning. Et minustegn på høyre side av likningen betyr at vedkommende primærlys skal adderes til Q. Ved å blande en av primærfargene med den prøven som skal

undersøkes, vil en få fram en farge som også kan lages ved en blanding av de to andre primærfargene. Maxwell påpekte også mulightene for at to primærlys må adderes til Q.

2.4.2 CIE-systemet

Den første internasjonale enighet om matematisk behandling av fargedata kom under en konferanse i Cambridge i England i 1931. Konferansen ble kalt *Commission International de L'Éclairage*. Fargemålingssystemet som baserer seg på denne konferansen, kalles *CIE-systemet*. Etter den tid har det kommet visse tillegg og modifikasjoner som munnet ut i anbefalinger av *Committe 1.3, 1971*. CIE-systemet er akseptert som internasjonal standard for fargemåling.

En farge bestemmes i CIE-systemet ved tre tall, *tristimuliverdiene* (X,Y,Z). Disse fastlegger både fargetone, metningsgrad og lyshetsgrad. Ser man bort fra lyshetsgraden, er to tall nok. Man benytter da de trikromatiske koeffisienter (kromatiske koordinater som de også kalles). For alle spektralfargene er tristimuliverdiene definert og finnes tabulert som funksjon av bølgelengden. Denne tabellen benevnes gjerne som CIE-systemet's *standard observatør*.

Standard observatør

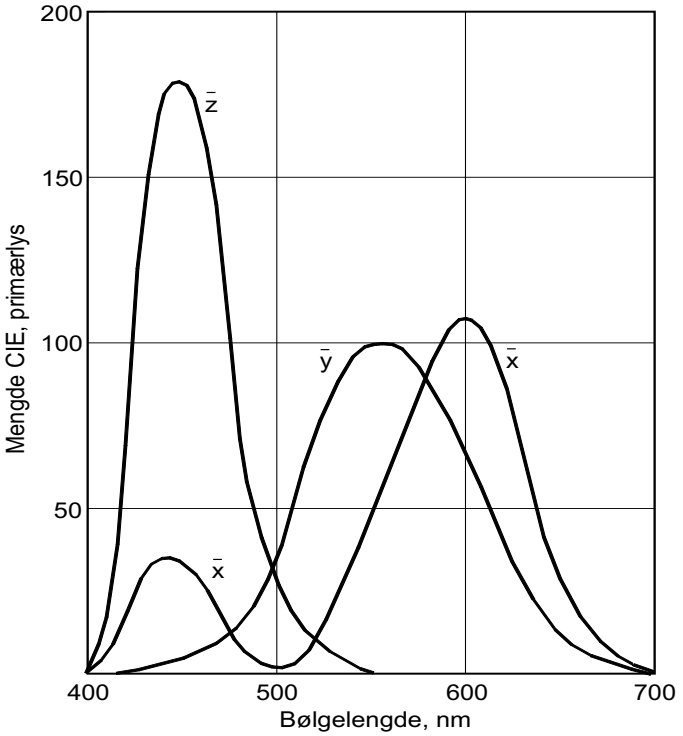
For å kunne ta hensyn til fargeoppfatningen til et normalt øye, definerer CIE-1931 en standard observatør. Denne hypotetiske person representerer et gjennomsnitt for noen titalls virkelige observatører. Testene ble utført i flere laboratorier. Forsøkspersonene ble bedt om å lage spektralfargene (innenfor meget små bølgelengdeintervall) ved bruk av tre primærfarger. De tilhørende tristimuli verdiene som symboliseres med $\overline{x}, \overline{y}, \overline{z}$ og benevnes *standard observatør*, er illustrert ved de tre kurvene i figur 2.7. Som vi ser av figuren ligner disse kurvene på responskurvene for de tre typer tappceller.

Standard belysning

Et objekts farge avhenger av lyskildens spektrum. Derfor definerer CIE standard belysninger. Figur 2.8 illustrerer den relative energi i de ulike bølgeområder for lyskildene A, B og C. Disse representerer henholdsvis lys fra en glødelampe, direkte sollys og gjennomsnittlig dagslys. CIE definerer også andre standard lyskilder enn de tre som er nevnt.

Reelle primærfarger

Normen i dag er at CIE-systemet benytter bølgelengdene $R = 700$ nm, $G = 546.1$ nm og $B = 435.8$ nm som primære bølgelengder. I CIE-systemet angis ikke (RGB)-verdiene direkte, men det foretas en transformasjon til (X, Y, Z)-koordinater (tristimuliverdiene).



Figur 2.7: CIE standard observatør

Bestemmelse av tristimuliverdiene

En objektiv bestemmelse av tristimuliverdiene for en farge foretas ved hjelp av et *spektrofotometer*. Her belyses fargeprøven med alle spektrets bølgelengder, tatt område for område, og refleksjonen for hvert av områdene bestemmes. Beregning av tristimuli verdiene baserer seg på en middeltallsberegning der standardobservatøren inngår som en vektsfunksjon, gitt ved:

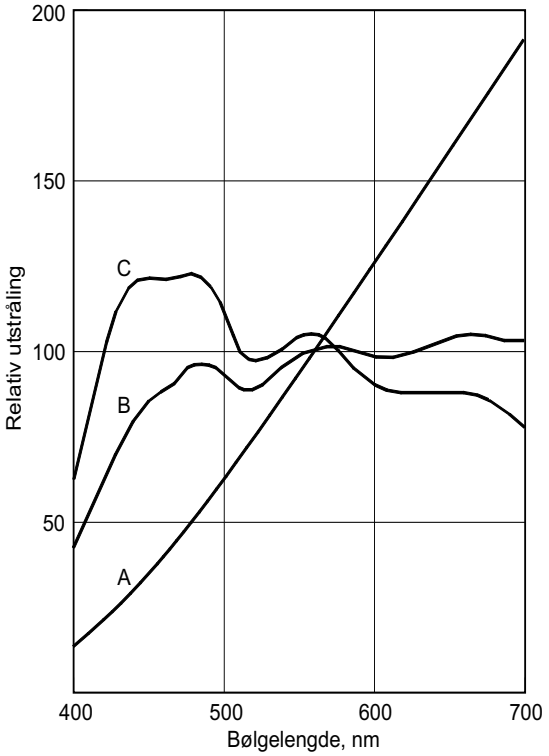
$$X = k \int \phi(\lambda) \cdot \overline{x}(\lambda) \cdot d\lambda \tag{2.4}$$

$$Y = k \int \phi(\lambda) \cdot \overline{y}(\lambda) \cdot d\lambda \tag{2.5}$$

$$Z = k \int \phi(\lambda) \cdot \overline{z}(\lambda) \cdot d\lambda \tag{2.6}$$

hvor $\phi(\lambda)$ er den spektrale fordelingen av lysenergien og k er en normaliserings konstant.

Et objekts farge avhenger av lyskildens spektrum. Derfor definerer CIE standard belysninger. Figur 2.8 illustrerer den relative energi i de ulike bølgeområder beskrevet for lyskildene A,B og C. Disse representerer henholdsvis lys fra en glødelampe, directe sollys og gjennomsnittlig dagslys. CIE definerer også andre standard lyskilder.



Figur 2.8: CIE standard belysninger

I nøyaktige spektrofotometre blir lys fra en lampe splittet i sine enkelte komponenter ved bruk av prismer. Ved hjelp av speil blir så små bølgeområder av spektret isolert og i sin tur brukt for å belyse fargeprøven. Det finnes spektrofotometre som baserer seg på bruk av fargefiltre og i noen tilfeller bare tre filtre. Når så få som tre filtre benyttes, må filtrene reflektere responskurvene for standard observatør". Å få til dette i praksis, er nesten å be om det umulige. Selv om det er gjort flere forsøk på å lage tre slike filtre, er det enda ikke lykkes å lage filtre som er særlig nøyaktige. Det finnes også såkalte visuelle kolorimetre. Disse baserer seg på at det ved hjelp av standardiserte lyskilder lages en sammenligningsfarge som gjøres mest mulig lik den fargen som skal bestemmes. Fargebestemmelsen blir med visuelle kolorimetre ikke objektiv, fordi resultatet er avhengig av at en person bedømmer fargen.

Kromatiske koordinater

CIE benytter en transformasjon av (X, Y, Z)-koordinatene til de såkalte *kromatiske* koordinater, som betegnes med (x, y, z). Transformasjonen er gitt ved

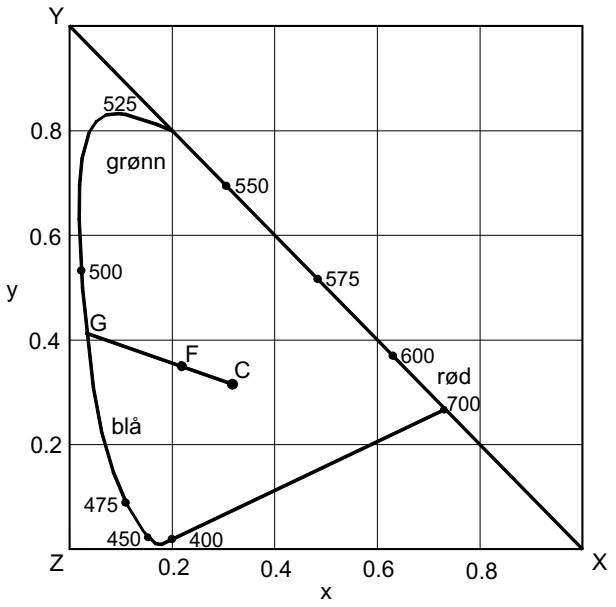
$$x = \frac{X}{X+Y+Z} \qquad y = \frac{Y}{X+Y+Z} \qquad z = \frac{Z}{X+Y+Z} \tag{2.7}$$

Vanligvis oppgis de to koordinatene (x, y) sammen med måleverdien for Y, som benevnes luminansfaktoren. De kromatiske koordinater til CIE's reelle primærfarger

er angitt i tabell 2.1.

Tabell 2.1: Kromatiske koordinater til CIE's reelle primærfarger

Reell primærfarge	Kromatiske koordinater		
	x	y	z
Rød 700.0 nm	0.73467	0.26533	0.00000
Grønn 546.1 nm	0.27376	0.71741	0.00883
Blå 435.8 nm	0.16658	0.00886	0.82456
Lyskilde B	0.34842	0.35161	0.29997



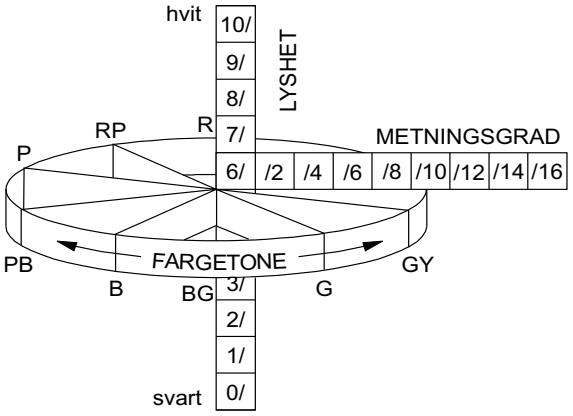
Figur 2.9: Spektralfargene plottet i xy-systemet (kromatiske koordinater)

Figur 2.9 viser spektralfargene plottet i xy-systemet. Speltralfargene ligger på en hesteskoformet kurve med hvitt (lyskilden) i sentrum C av hesteskoen. La oss anta at vi har en farge F med de kromatiske koordinater (0.22, 0.36). En rett linje fra C gjennom F skjærer linjen for spekteret i punktet G. Skjæringspunktet kalles for F sin dominerende bølgelengde og forholdet CF/CG definerer F sin metningsgrad. For å angi det absolutte nivå på enegien i lyset, angis måleverdien Y.

2.4.3 Munsell fargesystem

Det finnes flere systemer for klassifikasjon av farger som baserer seg på en visuell sammenlikning av farger. Det mest kjente av disse systemene er Munsell fargesystem som har sitt navn etter den amerikanske maleren A. H. Munsell. Fargene klassifiseres her på grunnlag av fargetone, lyshet og metningsgrad. Fargetone svarer til

bølgelengde og lyshet svarer til energien i det reflekterte lyshet. Metningsgrad har ikke noen fysikalsk tolkning, men er et psykologisk/fysiologisk fenomen. Fargemetningen er et uttrykk for hvor mye fargen skiller seg fra grå. Munsell fargesystem skiller mellom mer enn 1500 farger. Fargene er ordnet i et tredimensjonalt aksesystem, gjerne organisert i en ringperm med et ark for hver fargetone. Hvert ark inneholder en rekke fargede plater som er organisert slik at gråaksen ligger parallelt med permens rygg med aksen for metningsgrad vinkelrett på gråaksen. Figur 2.10 illustrerer Munsell fargesystem. Det er utarbeidet kurver som viser sammenhengen mellom Munsell fargesystem og CIE-systemet [McL86].



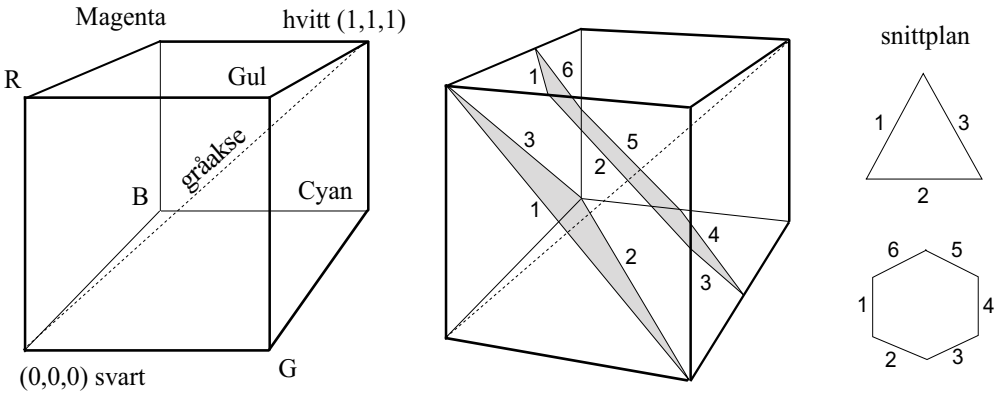
Figur 2.10: Munsell fargesystem

2.4.4 Fargeterning

En farge beskrives ved sine tre komponenter (1) fargetone (F), (2) metningsgrad (M) og (3) lyshet (L). Vi vil utdype disse begrepene med utgangspunkt i en fargeterning og vise hvordan vi kan etablere en transformasjon fra (R,G,B)-koordinater til (F,M,L)-koordinater. Det gjøres oppmerksom på at det finnes flere alternative måter for å modellere overgangen fra (RGB)-systemet til system som opererer med de tre perseptuelle komponenter til en farge, se for eksempel [FvDFH93]. Vårt kommende valg av modell er derfor bare én blant flere alternative betraktningsmåter.

La oss anta de tre additive primærfarger og at vi velger dem som akser i et rettvinklet 3D-aksesystem. Vi antar videre at de tre primærfargene kan anta intensitetsverdier i intervallet $[0, 1]$. En terning som avgrenses av det angitte intervallet, utspenner nå alle de farger vi kan framstille ved en blanding av de tre primærfargene. Siden hvitt framkommer ved lik blanding av de tre primærfargene, finner vi gråaksen langs diagonalen mellom hjørnene $(0, 0, 0)$ og $(1, 1, 1)$ som illustrert i figur 2.11.

Definisjon 1 *Alle farger som ligger i et plan vinkelrett på gråaksen har samme lyshet.*



Figur 2.11: Fargeterning

Definisjon 2 *En farges metningsgrad kan uttrykkes ved dens avstand fra gråaksen.*

Definisjon 3 *En farges fargetone kan uttrykkes ved en rotasjon om gråaksen i et plan vinkelrett på gråaksen .*

De tre definisjonene ovenfor kan benyttes til å finne en transformasjon fra (R,G,B)-koordinater til (F,M,L)-koordinater. Hvor godt de (F,M,L)-verdier som på denne måten framkommer svarer til komponentene i Munsell fargesystem, er ikke uten videre opplagt. Her vil perseptuelle faktorer og valg av bølgeområder for de tre primærfargene spille en avgjørende rolle.

La oss med bakgrunn i våre definisjoner på en farges tre komponenter etablere et koordinatsystem der vi legger den ene aksen langs gråaksen. Vi setter oss nå til oppgave å finne en transformasjonsmatrise **A** som gir oss

$$\begin{pmatrix} I \\ X \\ Y \end{pmatrix} = \mathbf{A} \cdot \begin{pmatrix} R \\ G \\ B \end{pmatrix} \tag{2.8}$$

hvor I er målt mot en akse som ligger langs gråaksen. I likning 2.8 har koordinatene X og Y ikke noen direkte tolkning i forhold til fargetone og metningsgrad, men vi skal senere vise hvordan M kan uttrykkes som en funksjon av $\sqrt{X^2 + Y^2}$ og hvordan F kan uttrykkes ved en rotasjon i XY-planet.

Vi skal nå utlede et uttrykk for **A**. Det kan stilles opp følgende betingelser for basisvektorene til aksene i det nye koordinatsystemet:

$$(\overrightarrow{X} \cdot \overrightarrow{I}) = 0; \quad (\overrightarrow{Y} \cdot \overrightarrow{I}) = 0; \quad (\overrightarrow{X} \cdot \overrightarrow{Y}) = 0 \tag{2.9}$$

hvor

$$\overrightarrow{X} = [X_R, X_G, X_B]; \quad \overrightarrow{Y} = [Y_R, Y_G, Y_B]; \quad \overrightarrow{I} = [I_R, I_G, I_B].$$

Siden $\overrightarrow{I}$ peker langs gråaksen, vil vektoren $\overrightarrow{I} = [1, 1, 1]$ kunne velges som basisvektor. Dette innført i likningssettet 2.9 gir oss følgende tre likninger til bestemmelse

av de seks resterende vektorkomponenter:

$$X_R + X_G + X_B = 0, \tag{2.10}$$
$$Y_R + Y_G + Y_B = 0, \tag{2.11}$$
$$X_R Y_R + X_G Y_G + X_B Y_B = 0. \tag{2.12}$$

Grunnen til at vi foreløpig har flere ukjente enn likninger, er at vi ikke har formulert betingelser som entydig fastlegger vektorparet $(\vec{X}, \vec{Y})$ sin posisjon i rommet. De manglende betingelser gjør at vektorparet kan rotere fritt i et plan vinkelrett på gråaksen. Vi må derfor gjøre et valg som kan sammenlignes med det å velge nullmeridian i et geografisk koordinatsystem. For å skaffe de nødvendige tilleggsvilkår, velger vi derfor en vilkårlig orientering av vektorparet. Ved å velge

$$X_R = 1; \qquad X_G = 1; \qquad Y_R = 1$$

må

$$X_B = -2; \qquad Y_B = 0; \qquad Y_G = -1$$

for at likningene 2.10-2.12 skal tilfredsstilles. Vi kan nå stille opp basistransformasjonen

$$I = (1,1,1) \cdot \begin{pmatrix} R \\ G \\ B \end{pmatrix} \tag{2.13}$$
$$X = (1,1,-2) \cdot \begin{pmatrix} R \\ G \\ B \end{pmatrix} \tag{2.14}$$
$$Y = (1,-1,0) \cdot \begin{pmatrix} R \\ G \\ B \end{pmatrix} \tag{2.15}$$
$$\tag{2.16}$$

Ved å normalisere basisvektorene får vi

$$\begin{pmatrix} I \\ X \\ Y \end{pmatrix} = \mathbf{A} \cdot \begin{pmatrix} R \\ G \\ B \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & \frac{-2}{\sqrt{6}} \\ \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{2}} & 0 \end{pmatrix} \cdot \begin{pmatrix} R \\ G \\ B \end{pmatrix} \tag{2.17}$$

Fra transformasjonen i likning 2.17 finner vi lett en farges avstand S fra gråaksen ved $\sqrt{X^2 + Y^2}$. Selv om S er et uttrykk for en farges metningsgrad M, er det likevel bedre å angi en normalisert verdi for M. Normaliseringen kan gjøres i forhold til den

farge som har størst avstand S_0 fra gråaksen. De seks hjørner i fargeterningen som ikke ligger på gråaksen, gir alle $S_0 = \sqrt{2}/\sqrt{3}$. Ved å beregne metningsgraden ved

$$M = \sqrt{\frac{3}{2}(X^2 + Y^2)}, \tag{2.18}$$

vil M ta verdier i intervallet $[0, 1]$.

La oss anta at vi for vår applikasjon har funnet en farge som vi mener har ideell metningsgrad og lyshet, men at fargetonen burde justeres noe. Denne justeringen kan modelleres ved en rotasjon om gråaksen gitt ved

$$\begin{pmatrix} R \\ G \\ B \end{pmatrix} = \mathbf{A}^{-1} \cdot \mathbf{B} \cdot \begin{pmatrix} I \\ X \\ Y \end{pmatrix} = \mathbf{A}^{-1} \cdot \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\phi & -\sin\phi \\ 0 & \sin\phi & \cos\phi \end{pmatrix} \cdot \begin{pmatrix} I \\ X \\ Y \end{pmatrix} \tag{2.19}$$

hvor ϕ er en rotasjonsvinkel om gråaksen. Siden A er en ortogonal matrise, har vi at $\mathbf{A}^{-1} = \mathbf{A}^{tr}$. Dette gir oss likningen

$$\begin{pmatrix} R \\ \\ G \\ \\ B \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} \\ \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} & \frac{-1}{\sqrt{2}} \\ \\ \frac{1}{\sqrt{3}} & \frac{-2}{\sqrt{6}} & 0 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 & 0 \\ \\ 0 & \cos\phi & -\sin\phi \\ \\ 0 & \sin\phi & \cos\phi \end{pmatrix} \cdot \begin{pmatrix} I \\ \\ X \\ \\ Y \end{pmatrix} \tag{2.20}$$

Planet vinkelrett på gråaksen danner et snittplan med fargeterningen som enten har en trekantet eller en sekskantet form, se illustrasjon i figur 2.11. Det er åpenbart at planet som går gjennom de tre (R, G, B)-punktene $(1, 0, 0), (0, 1, 0)$ og $(0, 0, 1)$, danner et trekantet snittplan med fargeterningen. Vi kan lett vise at dette planet står vinkelrett på gråaksen.

Teorem 1 *Planet gjennom (R, G, B)-punktene $(1, 0, 0), (0, 1, 0)$ og $(0, 0, 1)$ står vinkelrett på gråaksen*

Bevis: Vi benevner posisjonsvektorene som svarer til de tre angitte punktene for henholdsvis $\overrightarrow{R}, \overrightarrow{G}$ og $\overrightarrow{B}$. Vektoren som definerer gråaksen er pr. definisjon gitt ved $\overrightarrow{I} = (1, 1, 1)$. Av forutsetningene følger at det aktuelle trekantplanet avgrenses av de tre vektorene

$$\overrightarrow{RB} = \overrightarrow{B} - \overrightarrow{R} = (-1, 0, 1), \tag{2.21}$$

$$\overrightarrow{RG} = \overrightarrow{G} - \overrightarrow{R} = (-1, 1, 0), \tag{2.22}$$

$$\overrightarrow{BG} = \overrightarrow{G} - \overrightarrow{B} = (0, 1, -1). \tag{2.23}$$

Herav ser vi lett at

$$(\overrightarrow{RB} \cdot \overrightarrow{I}) = (\overrightarrow{RG} \cdot \overrightarrow{I}) = (\overrightarrow{BG} \cdot \overrightarrow{I}) = 0,$$

hvilket skulle vises. □

På tilsvarende måte kan vi vise at planet gjennom (R, G, B)-punktene $(0, 1, 1)$, $(1, 0, 1)$ og $(1, 1, 0)$ $(Cyan, Magenta, Gul)$ danner et trekantplan vinkelrett på gråaksen. De to trekantplan gitt ved posisjonsvektorene $(\overrightarrow{R}, \overrightarrow{G}, \overrightarrow{B})$ og $(\overrightarrow{Cyan}, \overrightarrow{Magenta}, \overrightarrow{Gul})$ representerer overgangene mellom trekantform og sekskantform til snittplanet vinkelrett på gråaksen. De korresponderende skjæringspunktene med gråaksen er gitt ved

$$I = (1/3)I_{\max} \quad \text{og} \quad I = (2/3)I_{\max}$$

hvor I_{max} er I-verdien til hvitt. Med de dimensjoner vi har valgt for fargeterningen, er maksimal I-verdi $\sqrt{3}$. De aktuelle skjæringspunktene lar seg lett utlede med utgangspunkt i enkel vektorbetraktning. For å kontrollere at ovennevnte resultat er riktig, kan vi for eksempel sjekke om vektoren fra punktet (1/3,1/3,1/3) til punktet (1,0,0) står vinkelrett på gråaksen. Vi får $\overrightarrow{[1,0,0]} - \overrightarrow{[1/3,1/3,1/3]} = \overrightarrow{[2/3,-1/3,-1/3]}$. Herav får vi $(\overrightarrow{[2/3,-1/3,-1/3]} \cdot \overrightarrow{[1,1,1]}) = 0$, hvilket viser at vektoren står vinkelrett på gråaksen. Den snittflaten med konstant lyshet som har størst areal, finner vi ved $(1/2)I_{\max}$. Dette snittplanet står i en gunstig stilling siden vi her har et stort antall fargenyanser.

2.5 Bruk av farger på kart

Bruk av farger på kart har to aspekter: (1) nytteaspekt og (2) estetisk aspekt. I en balansert design er begge aspektene ivaretatt.

2.5.1 Fargekomposisjon

En enkelt farge kan ikke sies å være hverken stygg eller pen, men det er i forhold til omgivelsene at vi kan danne oss en mening om fargens estetiske verdi. Vår oppfatning av harmoni, akkord og melodi er knyttet til en komposisjon, d.v.s. hvordan de ulike elementer i komposisjonen står i forhold til hverandre. Det er derfor vanskelig å gi regler for hvordan lage en god fargekomposisjon på kart, men noen aksepterte retningslinger skal likevel angis.

Komplementære farger står i et estetisk gunstig forhold til hverandre. Dette gir grunnlag for å sette opp toergrupper av farger. Treergrupper kan også settes opp med utgangspunkt i at dette er farger som i blanding kan gi hvitt. Basert på [Imh65] vil vi angi toer- og treer-gruppene i tabell 2.2. Enda bedre virker slike to- eller treklanger når fargene blandes med like deler hvitt eller svart. En tredje gruppe farger som harmoniserer, er fargerekker som varierer i lyshet. Dette er fargerekker som blant annet benyttes på koropletkart. Størrelse, form og plassering til de enkelte flater har stor betydning for kartet's estetiske nivå. Dette gjør det derfor vanskelig å velge farger for landsdekkende kartserier, fordi de underliggende geografiske forekomster kan selvsagt variere sterkt fra landsdel til landsdel.

Gul regnes for å være en vanskelig farge. Gul kan virke bra i moderate mengder som bunntone, men gir et uheldig inntrykk i kombinasjon med hvitt. Grå er forøvrig

Tabell 2.2: Estetisk gunstige fargekombinasjoner

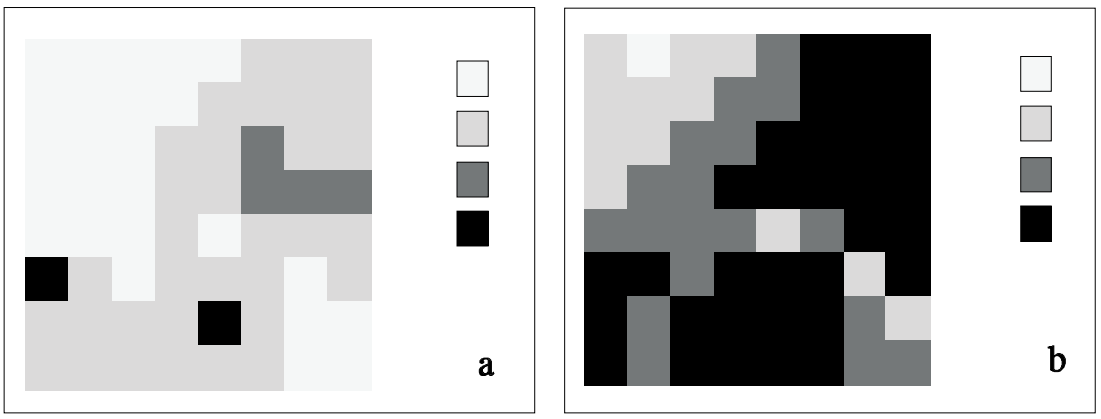
Toergruppe	Treergruppe
gul - fiolett	gul - rød - blå
gulorange - blåfiolett	gulorange -rødfiolett - blågrønn
orange - blå	orange - fiolett - grønn
rødorange - blågrønn	rødorange - blåfiolett -gulgrønn
rød - grønn	
rødfiolett -gulgrønn	

en utmerket bunnfarge for det fargede tema. Et misforhold mellom farger på et kart kan dempes ved å tynne fargene ut med grå.

Rød er en sterk farge og må kun brukes i små porsjoner. Generelt kan vi si at rene, skinnende eller svært kraftige farger ikke må brukes på store flater. I små mengder på en nøytral bunnfarge kan de komme til sin rett. Et fargerikt tema kan bare bygges på en rolig bakgrunn. Sterke farger må derfor bare brukes på viktige tema som dekker små flater. I spørsmål om form og fargekomposisjon dreier det seg om å søke etter en enkel, klar, sterk og godt avbalansert uttrykksform. Viktige og særegne trekk må komme klart fram, det generelle eller mindre viktige skal spille med lett og svakt.

2.5.2 Fargekontrast

Forskjellen mellom fargene på et kart må være så stor at fargene lett lar seg identifisere i alle situasjoner. Velges for eksempel grønn som farge for skog, kan vi ikke plassere grønne symboler på denne bunnfargen. Selv om dette er en selvfølge, forekommer likevel slik feilaktig bruk av farger på kart. Grunnen er gjerne at kartografen overser alle de kombinasjoner av kartelementer som kan forekomme. På grunn av

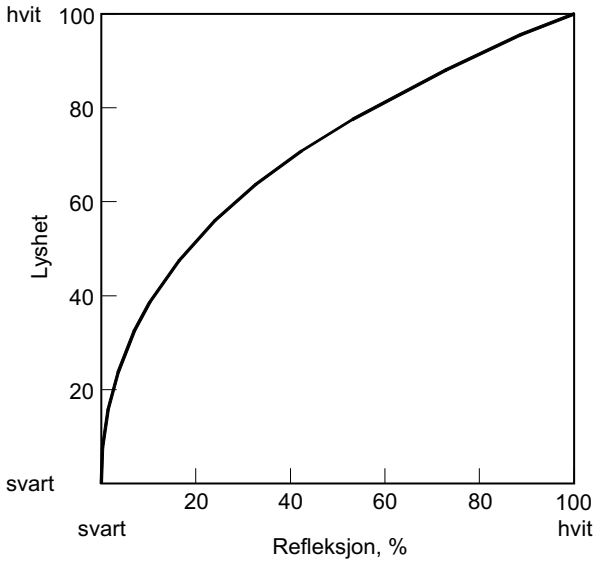


Figur 2.12: Opplevelsen av en farge er avhengig av dens omkringliggende farger

overstråling og nabovirkninger vil fargen til en flate oppleves forskjellig avhengig av flatens størrelse og fargen til naboflater. Dette problemet er knyttet til små flater. For eksempel i koropletkartene i figur 2.12 oppleves gråfargene annerledes i de små rutene i tegnforklaringen enn på selve kartet.

2.5.3 Ekvidistant gråtoneskala

I noen sammenhenger ønsker vi å lage gråtoneskalaer der vår fargeopplevelse av skalaen er at den har ekvidistante trinn. På grunn av at øyets respons er logaritmisk i forhold til endringer i energien i lyset, trenger en lys flate langt større endring i refleksjonen enn en mørk flate for å gi inntrykk av samme endring i gråtone. Sammenhengen mellom refleksjon og vår opplevlese av en flates lyshet er illustrert i figur 2.13.



Figur 2.13: Sammenhengen mellom refleksjon og lyshet i Munsell gråtoneskala

2.5.4 Normer for fargevalg

For landsdekkende kartserier vedtas normer for fargevalget. Dette har gjort at enkelte fargevalg har festet seg i folks besvissthet. Eksempler på normer som har bred utbredelse:

1. blå - vann, kulde, positive tall;
2. grønn - vegetasjon, lavland, skog;
3. gul - tørre områder, sparsom vegetasjon;

4. brun høydekurver, landformer;

5. rød - negative tall, viktige detaljer som veier og byer.

Av hensyn til effektiv kommunikasjon er det viktig at fargevalget er i tråd med tradisjonelle fargeassosiasjoner. Ved avvikende fargevalg bør en vurdere om det er berettiget og vellykket. Farger kan ha forskjellig betydning i de ulike kulturer. For eksempel symboliserer hvitt sorg i enkelte land mens i vestlige land forbinder vi hvitt med renhet og svart med sorg.

Kapittel 3

Grafisk semiologi

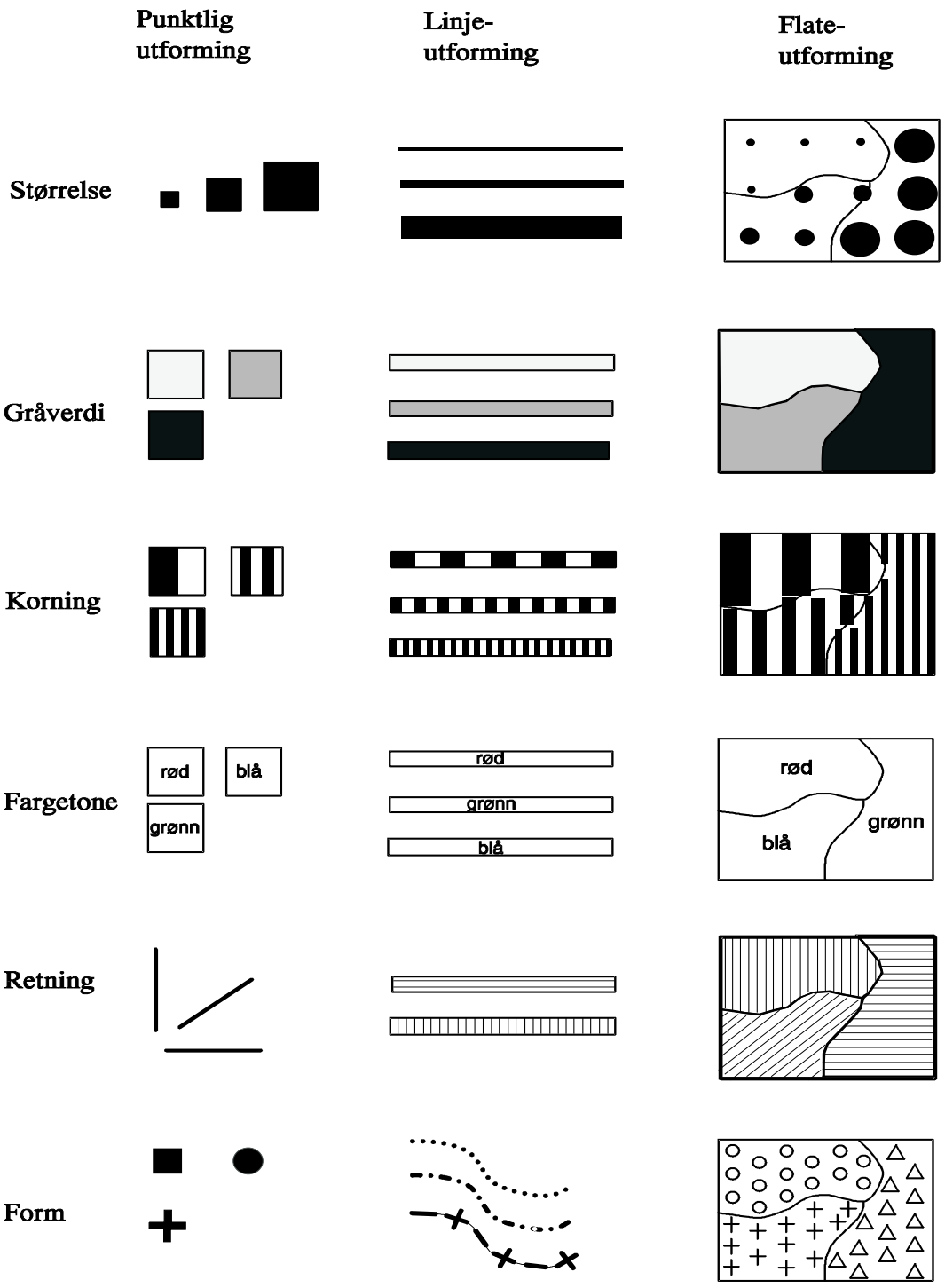
Begrepet *semiologi* brukes innen språkvitenskapen om læren om tegn. Tegn kan være alle former for symboler som benyttes i kommunikasjon (semiologi kommer av gresk, semion = tegn og logos = lære). Med *grafikk* mener vi her et system med regler og redskaper for bildeframstilling (grafikk kommer av det greske ordet grafein = tegne). Redskapene er de *visuelle variable* og reglene bruken av de visuelle variable. Selv om grafisk semiologi ikke begrenser seg til kart, vil vi i dette kapitlet med grafisk semiologi forstå læren om karttegn. Den franske kartograf Jaques Bertin har høstet store fortjenester i utviklingen av den grafiske semiologi. Bertin [Ber81] var den første som forklarte visuelle variable og de tre persepsjonsnivåene: (1) lesekart, (2) sebare kart og (3) kommuniserbare kart. Bertins bok [Ber81] ble opprinnelig skrevet på fransk, men ble i 1981 oversatt til engelsk. Bertin ble i 1993 tildelt en pris av International Cartographic Association (ICA) for denne boken. Bertins system av grafiske variable ble opprinnelig knyttet til statiske kart, men det har siden vist seg at en utvidelse til dynamiske kart er holdbar [Mac94].

Det er først mot slutten av 1980-tallet at teoriene til Bertin ble allminnelig kjent utenfor Frankrike. Geograf Axel Baudouin introduserte Bertins teorier i kartografiundervisningen i Norge fra begynnelsen av 1980 tallet og utga blant annet rapporten [BA84] i samarbeid med P. Anker.

3.1 Visuelle variable

Bertin [Ber81] beskriver åtte visuelle variable:

1. planets x-koordinat,
2. planets y-koordinat,
3. størrelse,
4. gråverdi,



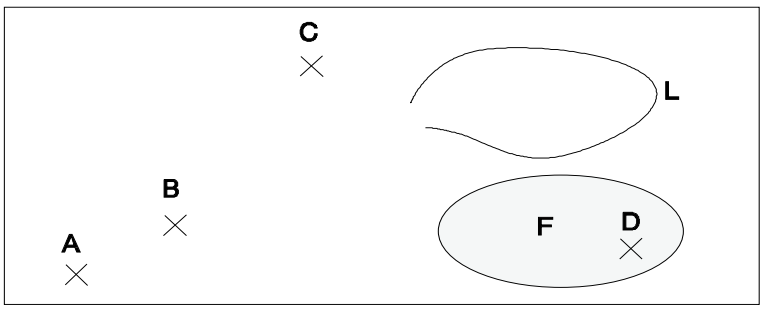
Figur 3.1: Visuelle variable og deres utformingsmuligheter.

5. korning,
6. fargetone,
7. retning,
8. form.

De to første visuelle variable definerer posisjonen i planet. De øvrige seks variable, som er illustrert i figur 3.1, kan best forklares ved å hevde at de visuelle variable er ortogonale. Det betyr at vi skal kunne endre en av de visuelle variable uten at endringen forplanter seg til de andre visuelle variable. Dette gjelder innenfor visse rammer. For eksempel kan vi anta en rett linje og at vi endrer formen til linjen slik at den får flere og flere buktninger. Formendringen fører til en gradvis endring av linjens fraktale dimensjonen (fraktal dimensjon er 2 når en linje er så buktet at den fyller ut hele planet). Endring av fraktal dimensjon til linja vil opplagt føre til at gråverdien til det kartplanet linjet er avbildet i, vil endre seg.

Planets visuelle egenskaper

Bertin viser at planet har tre visuelle egenskaper: (1) selektiv egenskap ($\neq$), (2) ordnende egenskap (O) og (3) kvanititativ egenskap (Q). Disse tre egenskapene er illustrert i figur 3.2. Av figuren ser vi at punktene A og B har forskjellig beliggenhet (selektiv egenskap), at punkt B ligger mellom punktene A og C (ordnende egenskap) og at avtsanden fra B til C er omtrent dobbelt så stor som avstanden mellom A og B (kvantitativ egenskap). Planet har også en fjerde egenskap som ikke direkte nevnes av Bertin, nemlig planets topologiske egenskap. I figur 3.2 ser vi at punkt D ligger innenfor areal F og at linje L ikke danner en lukket kurve.



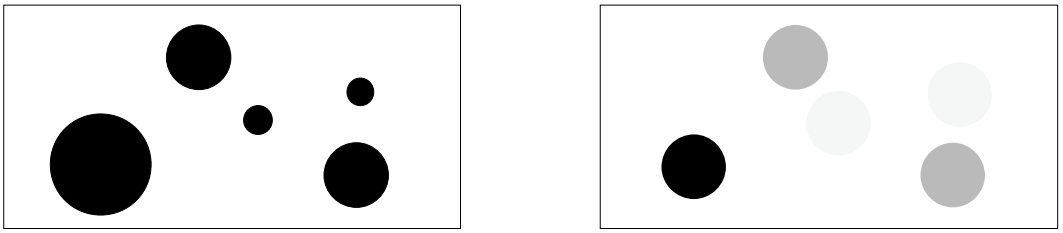
Figur 3.2: Planets visuelle egenskaper.

Størrelse og gråverdi

Størrelse er kvantitativ (Q). Vi kan for eksempel bedømme størrelsesforholdet mellom to sirkler og si at den ene sirkelen er dobbelt så stor som den andre. Siden

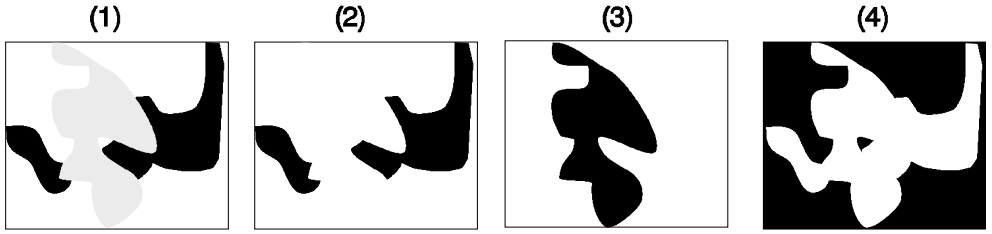
størrelse er kvantitativ, er den også ordnende. Gråverdi, derimot, er ikke kvantitativ, men er kun ordnende (O), fordi hvit kan ikke danne basis for å måle svart. Vi innser det meningsløse i å spørre om hvor mye mørkere svart er enn hvitt. Derimot kan vi lett oppfatte at grå ligger mellom hvit og svart. Se figur 3.3.

Størrelse og gråverdi har variabel synlighet ($\not\equiv$). En stor sirkel har større synlighet enn en liten sirkel og lys grå synes dårligere mot hvit bakgrunn enn svart. Se figur 3.3 der det er klart at den store sirkelen i det venstre vinduet oppfattes før de to små sirklene. Tilsvarende for det høyre vinduet der den svarte sirkelen oppfattes raskere enn de lyse sirklene.



Figur 3.3: De visuelle variable størrelse og gråverdi induserer et visuelt hierarki.

Størrelse og gråverdi har sterkt selektive egenskaper ($\neq$), noe som gjør at disse variable er egnet til å danne delbilder og visuelle mønstre. Siden størrelse og gråverdi ikke har egenskapen lik synlighet ($\equiv$), vil de indusere et visuelt hierarki som favoriserer visse elementer. Dette er illustrert i figur 3.4 der karakteristikk (2) er favorisert i bilde (1). De svarte flekkene har maksimal kontrast til den hvite bakgrunnen og framtrer derfor i bilde (1). Bildene (2), (3) og (4) inneholder hver én av karakteristikkene fra bilde (1).



Figur 3.4: Gråverdi er sterkt selektiv og vil danne delbilder. Gråverdi induserer et visuelt hierarki.

Korning

Korning representerer en fotografisk forminskning av et gitt mønster. Dette gjør at gråverdien kan holdes konstant mens korningen varierer. Korning er selektiv ($\neq$) og har lik synlighet ($\equiv$). Korning er også ordnende (O). I motsetning til gråverdi kan korning vise en ordning uten at synligheten til de skraverte feltene varierer.

Selv om korning har ordnende egenskaper, anses likevel dens optimale egenskap å være dens selektive egenskap. Korning brukes derfor helst til å vise kvalitative data. Dersom vi fokuserer på størrelsen til kornene, vil vi kunne bedømme ordningen mellom objektene. Om vi derimot fokuserer på hele flatene, oppfatter vi ikke så lett ordningen, men det at flatene har ulike egenskaper (seleksjon).

Fargetone (≈ Bertins visuelle variable farge)

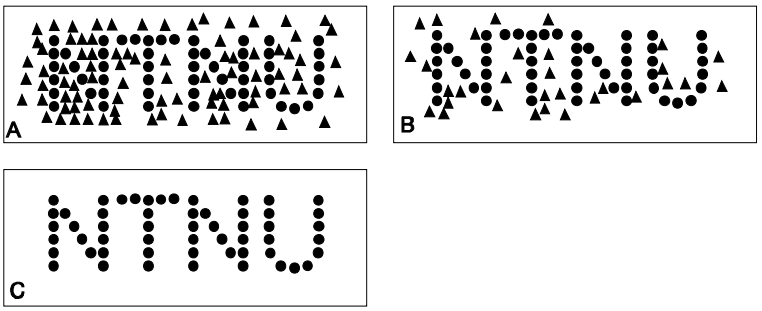
Farge kan splittes i de tre komponentene: (1) fargetone, (2) metningsgrad og (3) gråverdi. Siden det er vanskelig å oppfatte endringer i fargetone for farger som ligger i nærheten av gråaksen, vil fargetonens selektive egenskaper kun utnyttes når vi utelukker farger som ligger nære gråaksen. Når Bertin snakker om farge, mener han fargetonen til farger som ligger langt fra gråaksen. Bertin utelukker metningsgrad som visuell variabel. Andre forfattere har senere definert de tre komponentene til farger som separate visuelle variable. Farge, i Bertins betydning, er en meget selektiv visuell variabel ($\neq$). Den har ikke ordnende egenskaper og må derfor kun benyttes til å vise kvalitative data. Selv om enkelte farger kan oppfattes som blandingsfargen av to farger, som for eksempel orange, bør ikke fargeskalaer som { gult, orange, rødt } benyttes til å vise ordninger. I stedet bør det velges en skala der bare gråverdien varierer.

Retning

Visuell variabel retning er selektiv ($\neq$) og har lik synlighet ($\equiv$). Retning vil i punktutforming ha like god selektivitet som farge og er alltid å foretrekke framfor form. I arealutforming er selektiviteten dårligere enn for farge og forutsetter at mønsteret er glissent. I linjeutforming bør det bare benyttes to klasser. Retning kan vise dynamiske objekter som strømretninger og bevegelsesretninger.

Form

Symboler som bare varierer i form, gir ikke noe helhetsbilde av variasjonen, fordi form er svært lite selektiv. Hvor lite selektiv form er, illustres i figur 3.5. Ordet NTNU lar seg ikke oppfatte i bilde A. I bilde B, der noen av trekantene i bilde A er fjernet, kan vi begynne å ane NTNU. Først i bilde C der alle trekantene er fjernet, kommer NTNU tydelig fram. Form som visuell variabel har egenskapen *lik synlighet* ($\equiv$). Det er denne egenskapen til form som er dens optimale egenskap i kartografiske anvendelser. Desverre finnes det en rekke eksempler på at form benyttes på kart som om den skulle ha gode selektive egenskaper. Siden formvariasjon lett kan benyttes til å lage et stort antall ulike symboler, kan vi lett fristes til å legge alle informasjonsvariable inn på ett enkelt kartblad. Det kan ikke sterkt nok understrekes at form er lite selektiv og derfor har liten evne til å danne delbilder. Form er kun egnet for kart der det er tilstrekkelig at hvert enkelt symbol leses, altså der vi ikke forlanger at grupper av symboler skal oppfattes.



Figur 3.5: Form er en lite selelektiv visuell variabel. NTNU oppfattes ikke i bilde A.

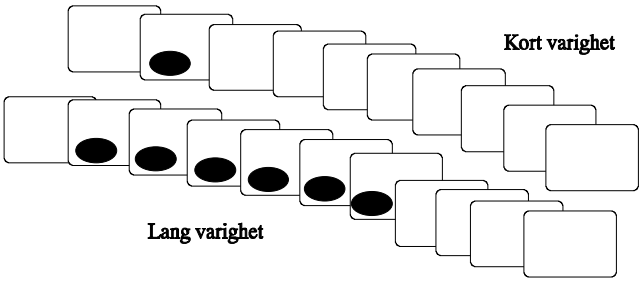
På grunn av at øyet kan skille mellom uendelig mange former, har enkelte organisasjoner blitt fristet til å finne standardiserte symboler for ulike temaområder. For eksempel utarbeidet Statens Kartverk i samarbeid med Pumatec AS i begynnelsen av 1990-årene 194 symboler for friluftsliv og sport. Det å huske så mange symboler er nesten umulig. Derfor må kun et lite utvalg av disse symbolene benyttes på et enkelt kart. Likevel finnes det eksempler på kart der flere titalls av symbolene er benyttet. Bare det å finne igjen et symbol i tegnforklaringen, er anstrengende. Det hjelper noe at en del av symbolene er bildesymboler og at de derfor til en viss grad gir assosiasjoner til den informasjon de skal formidle. Vi må iakta at hensikten med kart er å formidle geografisk informasjon på en rask måte. Problemet med standardisering av symboler er ikke knyttet til det å utforme symbolene, men å avgrense temaområdet.

Se forøvrig figur 3.12 som illustrer den geografiske fordeling av visse militærstrategiske mål i krigen mellom Irak og Kuwait i 1992. Selv om det i kartet bare er benyttet sju typer bildesymboler, er vi likevel ikke i stand til å danne delbilder. For eksempel så er delbildet til utbredelsen av kjemiske våpen umulig å oppdage. Kartet er kun egnet til å svare på spørsmål av typen: "hva har vi av mål i nærheten av Bagdad?"

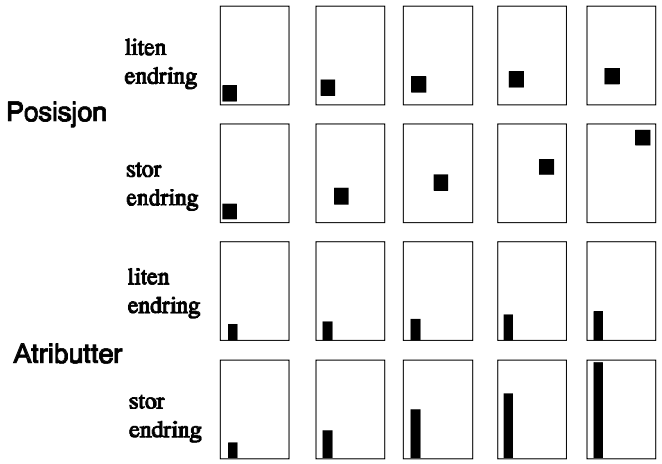
Tid som kartografisk variabel

Øyet er meget følsomt for bevegelse. Derfor er bevegelse en meget sterk visuell variabel. Dette gjør at vi raskt blir slitne av å se på et kart med blinkende symboler. Blinkende symboler må derfor kun benyttes til å framheve spesielt viktige elementer i kartet og må benyttes med stor forsiktighet. MacEachren [Mac94] beskriver følgende dynamiske visuelle variable:

1. varighet (eng. duration);
2. endringsgrad (eng. rate of change);
3. rekkefølge (eng. order), ordning etter tid eller etter attributtverdi;
4. fase (eng. phase), rytmisk repetisjon av visse hendelser.



Figur 3.6: Tid som visuell variable. Temaets varighet er illustrert ved kort og lang varighet.



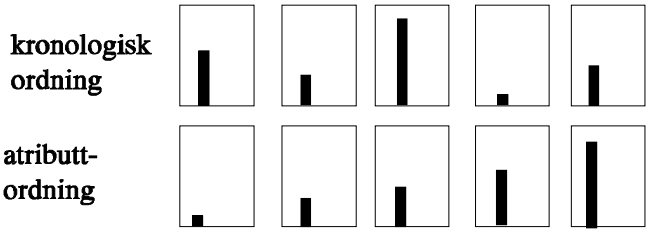
Figur 3.7: Tid som visuell variable. Temaets endring er illustrert ved lav og stor endringsgrad i både posisjonsdomenet og i atributtdomenet.

De visuelle variable varighet, endringsgrad og rekkefølge er illustrert i fiurene 3.6, 3.7 og 3.8. Varighet benyttes til å bestemme hvor lang tidsperiode objektene skal være synlige. Endringsgrad fastlegger graden av endring i et objekt fra det ene bildet til det neste bildet. Endringsgrad gjelder både det posisjonelle domenet (X, Y) og atributtdomenet.

Dersom vi fastholder som utgangspunkt at kart skal gi effektiv kommunikasjon av geografiske mønstre, vil det være rett å hevde at bruk av tid som kartografisk variabel har en rekke fallgruver. Bevegelse i et bilde vil tiltrekke seg oppmerksomhet. Dersom hensikten er å framheve visse objekter, kan bevegelse være en god løsning, men det finnes flere eksempler på fjernsynskart der blinkende objekter virker forstyrrende på oppfattelsen av mønstre (regionaliseringer) i kartet.

Andre perseptuelle variable

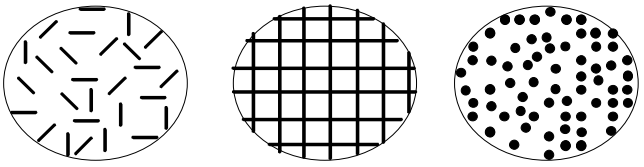
Enkelte forfattere angir *metningsgrad*, fargetone og gråverdi som visuelle variable. Bertin opererer som kjent med bare to fargekomponenter som visuelle variable (farge



Figur 3.8: Tid som visuell variable. Temaets ordning er illustrert ved kronologisk ordning og ordning etter atributtet.

og gråverdi). Logisk sett skulle det være mulig å benytte metningsgrad til å illustrere ordninger. Dette er likevel ikke å anbefale, fordi endring i metningsgrad kan forveksles med endring i gråverdi. Dessuten er det liten grunn til å benytte metningsgrad, siden gråtone er bedre egnet.

Anson og Ormelig [AO93] benytter *struktur* som visuell variabel. Innenfor et symbol kan enkle grafiske elementer fordeles regulært slik at de danner en spesiell struktur som vist i figur 3.9.



Figur 3.9: Struktur som visuell variabel

Bruk av *lydeffekter* kan kombineres med kart. Dette tilfører kartet en ny dimensjon. Det har vært utført noe forskning omkring lyd som kartografisk variabel, men feltet er foreløpig nytt. Ved State University of New York at Buffalo ble det i 1993 avlagt en dr.grad på dette emnet.

I *taktile kart* (*lat.* tactilis = som gjelder berøring) benyttes ruheten til overflater til å markere geografiske fordelinger. Oppvarming av deler av en overflate kan også utnyttes til å lage taktile kart. Taktile kart benyttes som kart for blinde.

Optimale egenskaper til Bertins visuelle variable

Tabell 3.1 viser optimale kartografiske egenskaper til Bertins visuelle variable. Størrelse og gråverdi har sterkt selektive egenskaper ($\neq$), noe som gjør at disse variable er godt egnet til å danne delbilder og visuelle mønstre. Siden størrelse og gråverdi ikke har egenskapen lik synlighet ($\equiv$), vil de indusere et visuelt hierarki som favoriserer visse elementer. Størrelse og gråverdi er begge ordnende (O) mens det bare er størrelse som er kvantitativ (Q) (kan vise relative dimensjoner). De optimale egenskaper til korning, fargetone og retning er seleksjon ($\neq$) og lik synlighet ($\equiv$). Selv om korning har ordnende egenskaper, regnes ikke ordning for å være en optimal

egenskap til korning. Visuelle variable som ikke har ordnende egenskaper, er kun egnet til å vise nominelle (kvalitative) data.

Form har bare én optimal egenskap, nemlig lik synlighet (≡). Det kan ikke sterkt nok understrekes at form er lite selektiv. Likevel finner vi ofte kart der dette ikke er tatt hensyn til. Resultatet blir gjerne kart som ikke formidler informasjonen på en effektiv måte.

Tabell 3.1: Optimale kartografiske egenskaper til Bertins visuelle variable

Visuell variabel	*Egenskap*			
X Y Planets 2 dimensjoner	Q	O	$\not=$	$\equiv$
Størrelse	Q	O	$\neq$	$\not\equiv$
Gråverdi		O	$\neq$	$\not\equiv$
Korning		o	$\neq$	$\equiv$
Fargetone			$\neq$	$\equiv$
Retning			$\neq$	$\equiv$
Form				$\equiv$

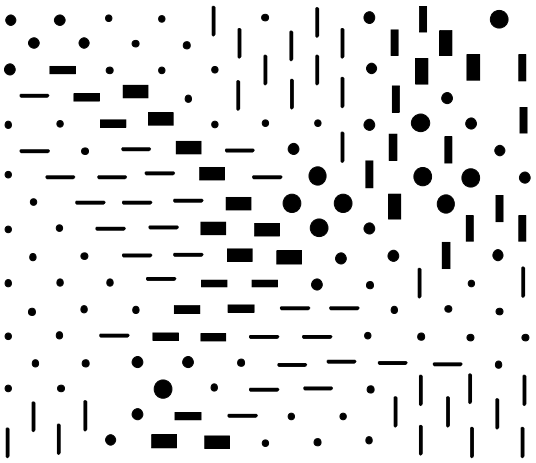
	Tegnforklaring
Q	kvantitativ
O	ordnende
o	svakt ordnende
$\neq$	selektiv
$\not\equiv$	variabel synlighet
$\equiv$	lik synlighet

3.2 Visuelt hierarki

Som vi allerede har diskutert, har de visuelle variable ulik grad av selektivitet (≠) og ulik evne til lik synlighet (≡). Av de visuelle variable er det bare form som har egenskapene ikke selektiv (=) og lik synlighet (≡). Det betyr at formsymboler satt sammen i et bilde ikke gjør oss i stand til å skille ut delbilder. De andre visuelle variable er selektive og kan skape delbilder. Størrelse er mest selektiv, så kommer gråtone, fargetone, korning, retning og form. I et kart laget med flere visuelle variable, vil mønsteret laget med størrelse og gråtone bli mere synlig enn mønsteret laget med retning og form. Forskjeller i egenskaper til de visuelle variable kan brukes bevisst til å forsterke eller dempe ned variasjoner i et bilde. Figur 3.10 illustrerer dette. I bildet framtrer mønsteret av store og små symboler umiddelbart, deretter oppfatter vi mønsterert av vertikale og horisontale streker. Til slutt oppdager vi ganske svakt mønsteret av sirkler og firkanter.

I figur 3.11 har vi to identiske datasett som er illustret med henholdsvis størrelse og korning. Dataene er delt inn i tre klasser. Det er uten videre klart at den geografiske utbredelsen til de tre klassene kommer tydeligst fram i bildet der størrelse er valgt. Noe som illustrerer at størrelse er mere selektiv enn korning.

Figure 3.12 illustrerer hvor lite selektiv form er. For eksempel klarer vi ikke i kartet å se mønsteret av forekomster av biologiske våpen eller hvor vi har flybaser. Kartet må leses og kan ikke gi svar på spørsmål av typen: *hvor har vi?* I figur 3.13 har kartet fått en annen design. Her oppdager vi lett mønsteret av store svarte



Figur 3.10: Visuelt hierarki. Størrelse er mere selektiv enn retning og form. Form er minst selektiv. Størrelse vil derfor lettere enn retning og form danne visuelle mønstre.

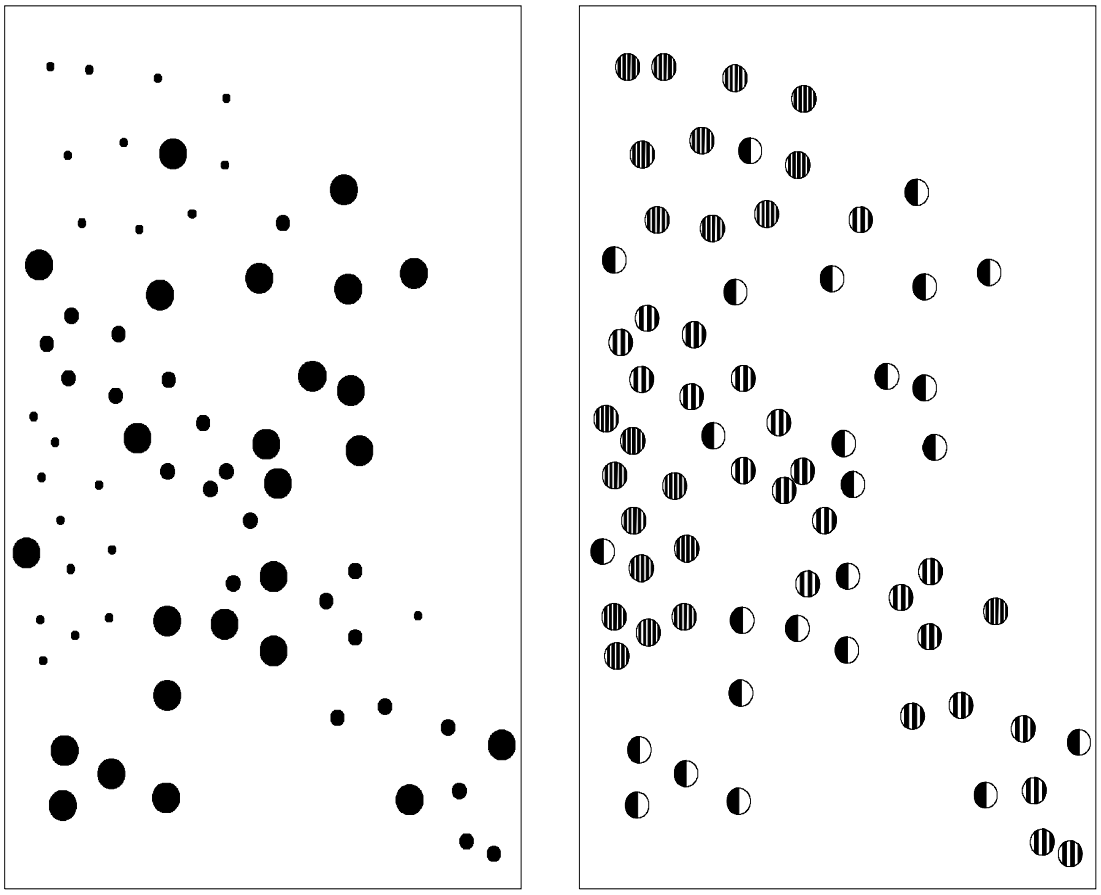
sirkler og mønstrene dannet av orienterte rektangler. Rakettbaser og flåtebaser kommer mindre tydelig fram på grunn av at deres tilhørende symboler har en lys grå farge. Det visuelle hierarkiet er i dette tilfellet utnyttet på en strategisk måte slik at mindre viktig informasjon ligger i bakgrunnen mens viktig informasjon er løftet inn i forgrunnen. Hva som er viktig informasjon er selvsagt et spørsmål om brukers behov for informasjon.

3.3 Lesekart, sebare kart og kommuniserbare kart

Det er definert tre persepsjonsnivåer for kart: (1) *elementnivå*, (2) *gruppenivå* og (3) *globalt nivå*. På elementnivå er det mulig å oppfatte kun ett og ett symbol av gangen. På gruppenivå lar flere symboler seg oppfatte av gangen mens på globalt nivå får vi en umiddelbar forståelse for det mønsterert kartsymbolene danner. Jo flere informasjonsvariable vi har på et kart, jo lengere dras kartets persepsjonsnivå mot elementnivå.

Kart som inneholder så mange informasjonsvariable at de ikke gir et globalt persepsjonsnivå, kalles for *lesekart*. Typisk for lesekart er at et lite antall symboler oppfattes av gangen. Lesekart kalles også for *referansekart* eller *generelle kart*.

Kart som har et globalt persepsjonsnivå, kalles for *sebare kart*. Det er kun de sebare kart som er egnet til å svare på spørsmål om hvor ulike forekomster er lokalisert. For eksempel: hvor har vi badeplasser, hvor har vi hoteller osv.. Sebare kart kalles også for *prosesskart* eller *tematiske kart*. Begrepet tematiske kart brukes i praksis ofte i en noe feilaktig betydning. Enkelte tror at kart som ikke er topografiske kart, er tematiske kart. Alle kart har selvsagt et tema, slik at begrepet tematisk kart er ikke egnet til å klassifisere kart på dette grunnlaget. Begrepet tematisk kart har en

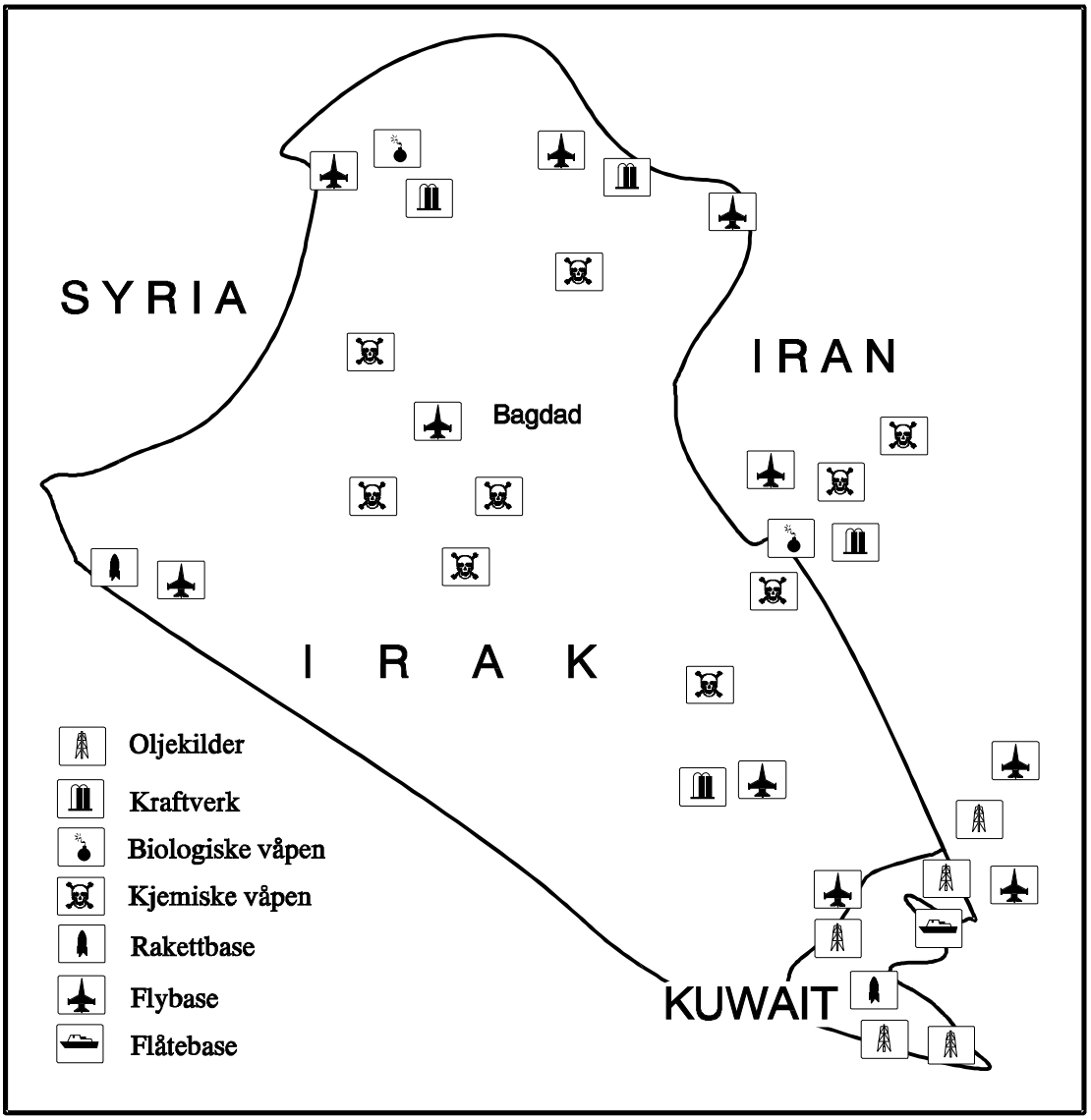


Figur 3.11: Visuelt hierarki. Størrelse er mere selektiv enn korning. Størrelse vil derfor lettere enn korning danne visuelle mønstre.

helt spesiell betydning, nemlig kart med et globalt persepsjonsnivå. Et tematisk kart vil alltid inneholde noe bakgrunnsinformasjon for å kunne gi kartinformasjonen en geografisk referanse. Bakgrunnskartet må ikke dominere over elementene i det som skal være det tematiske kartet.

Selv om topografiske kart vanligvis regnes for å tilhører gruppen lesekart, kan det finnes informasjonsvariable i topografiske kart som formidles på et globalt persepsjonsnivå. For eksempel i fylkeskartet over Oslo og Akershus i målestokk 1:150 000 oppdager vi lett hvor vi har store tettsteder. For dette temaet har kartet et persepsjonsnivå i nærheten av globalt nivå. Selv om gruppene lesekart og sebare kart ikke er skarpt definerte, har begrepene likevel den viktige funksjon at de gjør oss bevisst på noen fundamentale valg vi må gjøre ved utforming av kart.

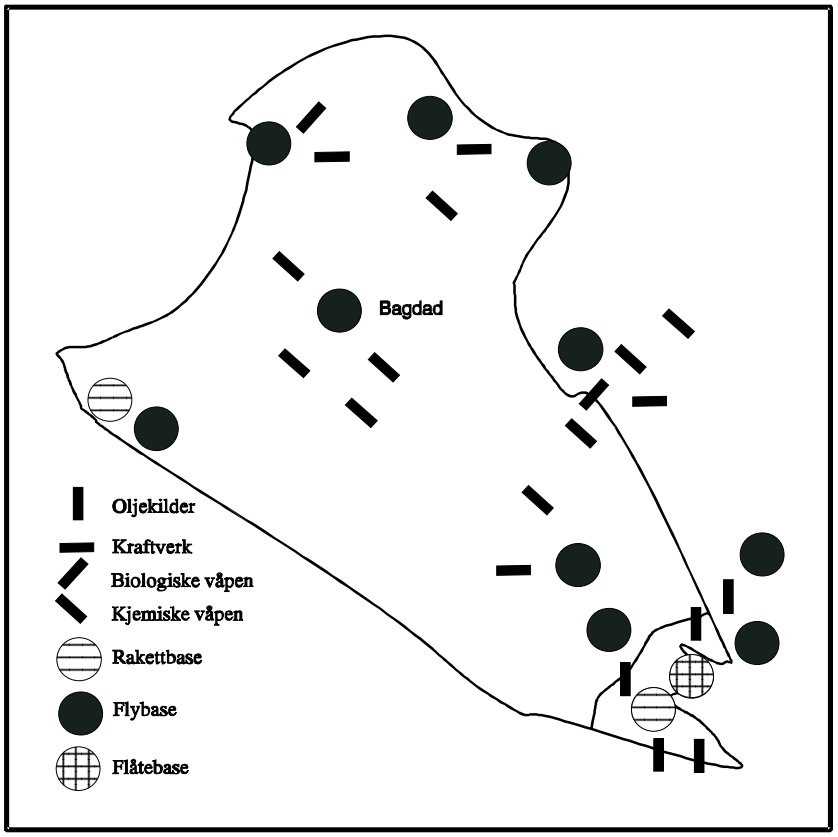
Kartet i figur 3.12 har et persepsjonsnivå på element/gruppenivå mens kartet i figur 3.13 har et persepsjonsnivå på gruppe/globalt nivå. Legg merke til at i figur 3.13 oppfatter vi mønsteret av symbolene for kjemiske våpen, mens dette mønsteret ikke kommer fram i figur 3.12. Dette bekrefter at retning er mere selektiv enn form.



Figur 3.12: Form er lite selektiv og vil ikke danne visuelle mønstre. Kartet er et lesekart og har et persepsjonsnivå på element-/gruppenivå.

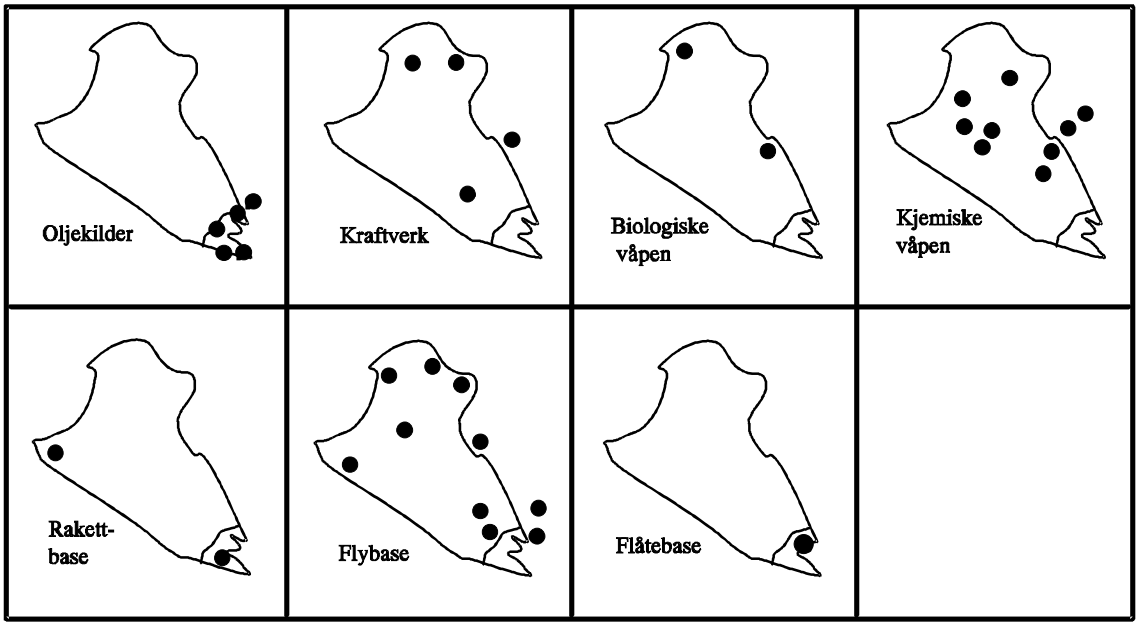
Siden de svarte sirklene på kartet vist i figur 3.13 er framtredende, er det riktig å definere persepsjonsnivået for disse symbolene til globalt nivå. De øvrige symbolene i dette kartet er på gruppenivå. Det er først i figur 3.14 at vi kan snakke om kart med et globalt persepsjonsnivå. Prisen vi har betalt for de sebare kartene, er at vi har spredd informasjonen på mange kart.

Selv om kartet i figur 3.13 lettere danner visuelle mønstre enn kartet i figur 3.12, er det først når vi reduserer antall informasjonsvariable til én (1) variabel pr. kart, at vi definitivt har et kart med et globalt persepsjonsnivå. I stedet for å legge mange informasjonsvariable inn på ett enkelt kart, kan vi altså lage flere kart. Ulempen er at kartene ikke lenger er godt egnet til å svare på spørsmål av typen *hva vi har av ulike forekomster på ett gitt sted.* Fordelen er at sebare kart lett gir svar på spørsmål om *hvor vi har en gitt type forekomster.*



Figur 3.13: Størrelse, gråverdi, retning og form. Symbolene danner delbilder. Kartets persepsjonsnivå er på gruppe/globalt nivå.

Figur 3.15 illustrerer et sterkt generalisert kart. Generalisering er ikke noe mål i seg selv, men et middel til å lage lett forståelige kart. Målet med det foreliggende kartet er å lage et kart som er lett å huske og som er i stand til å svare på både *hvor har vi?* og *hva har vi?* Vi har to informasjonsvariable i kartet: (1) (flybaser, rakettbaser og flåtebaser) og (2) (oljekilder, kraftverk, kjemiske- og biologiske våpen).



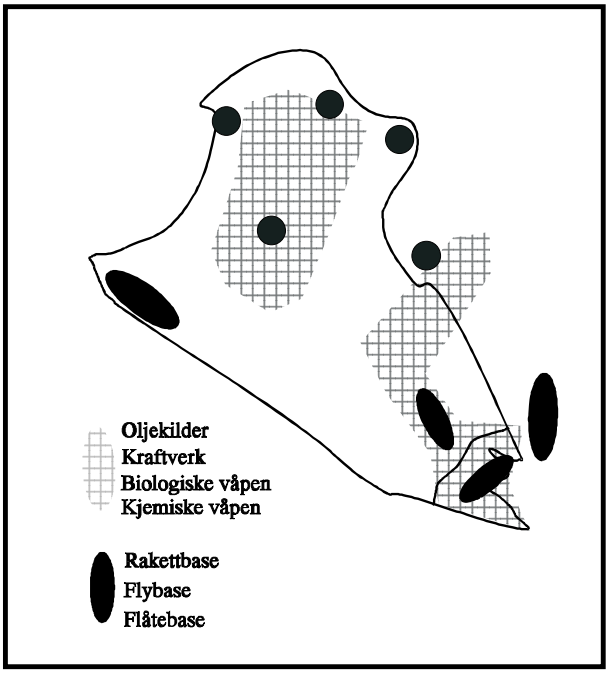
Figur 3.14: Sebare kart, én informasjonsvariabel pr. kart. Persepsjonsnivået er på globalt nivå.

Begge informasjonsvariable formidles på et globalt nivå. Følgelig kan kartet svare på spørsmål om hvor vi har. Kartet kan også svare på spørsmål om hva vi har. Et kart av denne typen kaller vi for et *kommuniserbart kart*. Generaliseringsnivået er alltid åpent for diskusjon og må rette seg etter den aktuelle bruken av kartet. Kommuniserbare kart nyttes til å formidle hovedresultatene av en geografisk analyse. For eksempel kan vi tenke oss at vi lager kommuniserbare kart for å vise konsekvensene av en reguleringsplan.

3.4 De ti kartografiske bud

Noen sentrale kartografiske regler vil bli summert opp i det vi har kalt De ti kartografiske bud.

1. Du må alltid sørge for at kartet får en logisk forbindelse mellom informasjonsvariabel og visuell variabel.
2. Du må aldri fylle opp kartet med selvfølgeligheter eller trivialiteter.
3. Du må aldri overfylle kartet med informasjon. Du skal i stedet lage flere kart.
4. Du skal utheve viktig informasjon.
5. Du skal ta i betraktning brukerens forhåndskunnskaper og hva brukeren ønsker å vite noe om.



Figur 3.15: Kommuniserbart kart. Kartet formidler svar på begge type spørsmål: hvor? og hva? Kartet er sterkt generalisert, men er lett å huske.

6. Du skal frykte lesekartene (kart med et persepsjonsnivå på element-gruppenivå).
7. Du skal ære de sebare kart (kart med et globalt persepsjonsnivå).
8. Du skal elske de kommuniserbare kart (kart som er lette å huske og som formidler sentrale resultater av din geografiske analyse)
9. Du må ikke begjære de unyttige kart.
10. Du må huske at en datamaskin aldri kan tenke eller bli stilt til ansvar for sine handlinger. Du er alltid ansvarlig for de kartografiske produkter som blir laget på ditt datasystem.

Kapittel 4

Noen utvalgte karttyper

Vi vil i dette kapitlet beskrive noen standardmetoder for å visualisere geografiske data. Behandlingen av emnet er ikke uttømmende, men noen sentrale hovedtrekk med tanke på dagens kartprodukter er likevel behandlet. Som utfyllende litteratur anbefales bøkene [RMM+95] og [Imh65]. Boken til Robinson og medforfattere dekker feltet kartografi på en utmerket måte. Den andre boken er skrevet av Imhof og behandler kun topografiske kart. Siden Imhof's bok er fra 1965, er naturlig nok ingen EDB-baserte teknikker bekrevet. Mange av prinsippene Imhof beskriver, er imidlertid allmenngyldige og bør også ha mye å gi til dagens systemutviklere og brukere av GIS.

4.1 Prikkekart

Prikkekart (eng. dot map) er en populær framstillingsmåte for kart i små målestokker. Blant annet er denne karttypen benyttet i Nasjonalatlas for Norge. I [RMM+95] finner vi flere karteksempler som illustrerer teknikken.

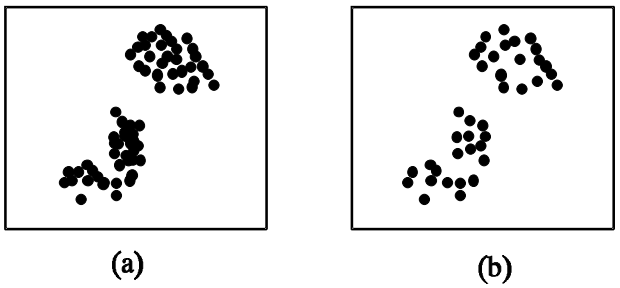
Definisjon 4 *Et prikkekart er et kart som viser den geografiske fordeling av forekomster ved hjelp av gjentatte prikkesymboler av gitt verdi.*

Konstruksjonsreglene for prikkekart er:

1. velg prikkeverdi,
2. velg prikkestørrelse,
3. velg prinsipp for å lokalisere prikkene.

Prikkeverdi og prikkestørrelse bør velges slik at det skal være mulig å oppfatte hver enkelt prikk selv i områder med stor prikketetthet. Hvis prikkene er for store eller prikkeverdien for liten, smelter prikkene sammen i tette områder. På den annen side får vi ikke fram den geografiske fordelingens karakter dersom prikkene er for små eller prikkeverdien er for stor. Problemet er defor å finne en passe balanse

mellom prikkestørrelse og prikkeverdi. Figur 4.1 viser to prikkekart og illustrerer avhengigheten mellom prikkenes enhetsverdi og prikketettheten. Kartet i bilde (a) har lavere prikkeverdi enn kartet i bilde (b).



Figur 4.1: Prikkekart med forskjellig prikkeverdi

Det er noen problemer knyttet til plasseringen av prikkene. Prikkene kan plasseres i tyngdepunktet for de enheter de representerer. I noen tilfeller kan dette gi en uheldig plassering av prikkene. La oss for eksempel anta at vi skal lage et prikkekart som viser saueholdet i Norge og at vi velger enhetsverdi for prikkene til 500 sauer. Tyngdepunket vi skal plassere en prikk i, må i dette tilfellet regnes over flere bruksenheter. Dersom gårdene ligger spredt, kan det hende at tyngdepunktet blir lokalisert i et område hvor det umulig kan ligge noen gårder (for eksempel et sjøområde). Basert på en analyse mot veinettverk, vannsystem o.l., er det mulig å foreta en etterkorrigering av prikkenes plassering.

I noen tilfeller har vi statistikk som refererer seg til administrative enheter (skolekrets, tellekrets, kommune osv.). Vi vil i slike tilfeller ikke vite hvor prikkene skal plasseres innenfor den administrative enheten. Det vi kan gjøre i dette tilfellet, er å fordele prikkene jevnt ut over deres respektive administrative enheter. Metoden kan forbedres dersom vi masker ut de områder vi vet forekomsten ikke kan være lokalisert i, og fordeler prikkene jevnt ut over de resterende områdene.

En stor fordel med prikkekart er at de gir en umiddelbar oppfattelse av hvor det er små og hvor det er store konsentrasjoner av den aktuelle forekomsten. På minussiden kan vi anføre at det er vanskelig å estimere hvor store mengder det dreier seg om innenfor et gitt område. Vi kan selvsagt telle antall prikker, men dette er en omstendelig prosess.

4.2 Kart med skalerte sirkler

Mengdeproporsjonale symboler er mye anvendt for å vise kvantitative data. En uheldig framstilling som av og til forekommer, er at symbolene utformes som bildesymboler. Dersom temaet for eksempel er antall skutte elg, kan det kanskje være fristende å benytte et bildesymbol av en elg, men sirkler er som regel å foretrekke. Sirkler har

en enkel geometrisk form som ikke ødelegger den globale oppfattelsen vi er ute etter for kartet (kompliserte former til symbolene øker entropien i kartet).

Definisjon 5 *Et kart med skalerte sirkler viser kvantitative data ved hjelp av mengdeproporsjonale sirkler.*

Sirklenes areal gjøres proporsjonal med mengden, eventuelt at det innføres en viss korreksjon i forhold til ren proporsjonalitet. Det viser seg nemlig at øyet har en tendens til å undervurdere størrelsen til store sirkler. Derfor benyttes av og til en psykologisk skalering av sirklene [RMM+95]. Følgende regel kan benyttes for psykologisk skalering av sirklene: (1) finn logaritmen til mengdene som skal framstilles, (2) multipliser disse verdiene med 0.57, (3) beregn antilogaritmen k, (4) gjør sirklenes radius proporsjonal med k. La oss betegne sirklenes areal med A, deres radius r og de foreliggende mengdeverdier M. La oss sette diameteren r_0 til enhetssirkelen lik l. Fra relasjonen

$$\frac{A}{A_0} = \frac{M}{M_0} \text{ hvor } \frac{A}{A_0} = \frac{\pi r^2}{\pi r_0^2} = \frac{r^2}{r_0^2}$$

får vi

$$r = \sqrt{\frac{M}{M_0}} \text{ for } r_0 = 1. \tag{4.1}$$

Herav har vi at $\log r = 0.5(\log M - \log M_0)$. Ved å innføre en liten korreksjon til logaritmene, får vi så den psykologisk skalerte sirkelradiusen r':

$$\log r' = 0.57(\log M - \log M_0). \tag{4.2}$$

Fra likningene 4.1 og 4.2 får vi eksemplet i tabell 4.1 som sammenligner verdier for r og r'.

Tabell 4.1: Sammenligning av sirkler uten og med psykologisk skalering

M	Sirkelradius	
	ikke psykologisk skalering	psykologisk skalering
10	1.0	1.0
100	3.2	3.7
1000	10.0	13.8

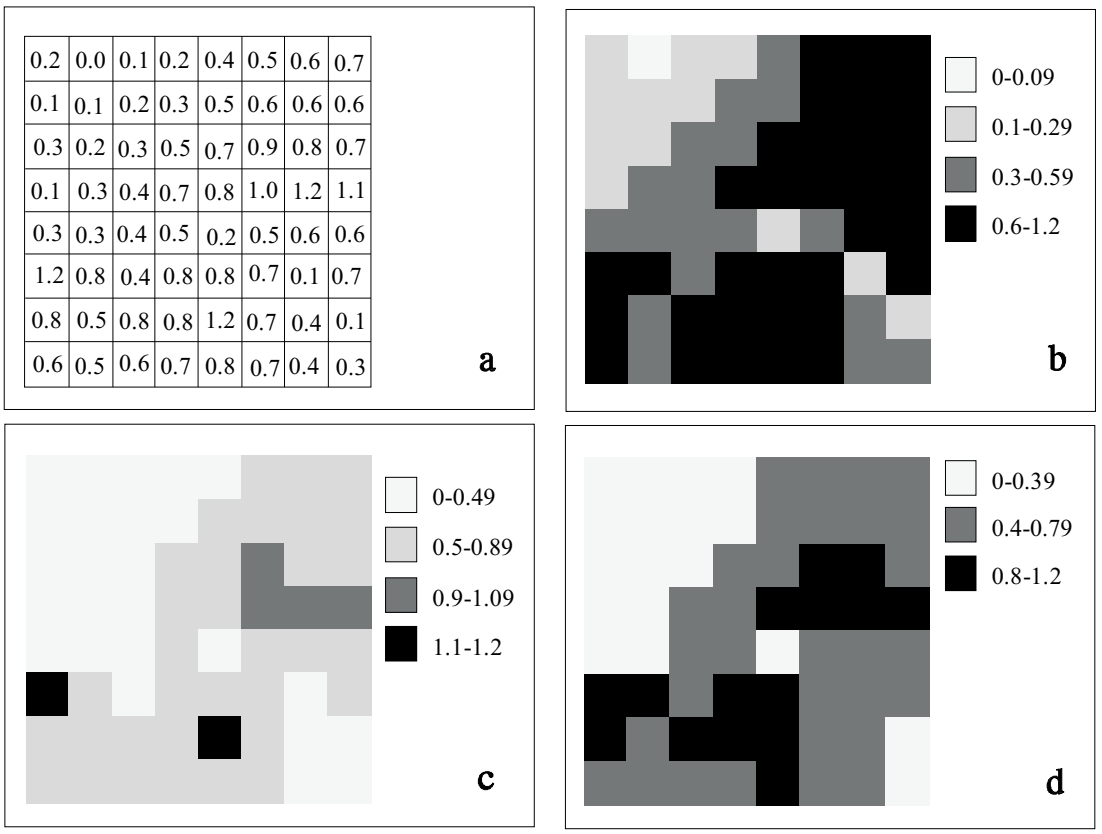
Et problem er ofte å finne en høvelig størrelse for den største sirkelen. For eksempel i befolkningskart over Norge blir problemet framtredende, fordi i Oslo-området har vi store befolkningskonsentrasjoner kombinert med at det geografiske området er lite i forhold til resten av Norge. Følgen er at vi ikke unngår overlappende sirkler i dette området. Dette kan til en viss grad løses ved å gjøre de overdekkende sirkler gjennomskinnelige. På den måten blir det mulig å se også de sirkler som er dekket over.

Som en løsning på problemer knyttet til data som har store spenn i verdiområdet, er foreslått å anvende kuler i stedet for sirkler. Kuler projisert inn i et 2D kart er vanskelige å estimere størrelsen av. Metoden kan neppe anbefales.

Det finnes en rekke undersøkelser om hvor nøyaktig vi klarer å estimere størrelsen til sirkler. Konklusjonene går i hovedsak ut på at vi ikke klarer å estimere størrelsen så altfor nøyaktig. Blant annet har vi naboeffekter som at en sirkel oppfattes som mindre når den er omgitt av store sirkler enn når den er omgitt av små sirkler. Metoden med flateproporsjonale sirkler har sin styrke i at vi raskt får et inntrykk av hovedtendensen i variasjonen.

4.3 Koropletkart

Figur 4.2 illustrerer prinsippet for koropletkart (eng. choropleth map). Ordet kkoropletstammer fra de to greske ordene choros (areal) og plethos (mengde).



Figur 4.2: Tre koropletkart. Kartene ser svært forskjellige ut selv om de er basert på det samme grunnlagsmaterialet.

Definisjon 6 *Et koropletkart viser den geografiske fordelingen av kvantitative forekomster ved hjelp av den visuelle variable gråtone i arealutforming.*

I definisjon 6 sies det at visuell variabel gråtone skal benyttes for koropletkart. Dette er hovedregelen, men det kan gjøres unntak fra denne regelen ved koropletkartframstilling av en terrengoverflates høyde over havet. Her benyttes ofte et fargevalg som gjenspeiler hvor rik vegetasjonen er (fargeskala fra mørk grønn via brunt til hvitt i høyfjells- og breområder).

Konstruksjonen av koropletkart gjøres etter følgende regler:

1. gjør om til relative verdier (mengde pr. arealenhet),
2. velg antall klasser,
3. velg klassegrenser,
4. velg gråtoner.

En feil som ofte gjøres på koropletkart, er at framstillingen baserer seg direkte på absolutte verdier. I de tilfeller at arealenhetene er tilnærmet like store, kan dette gå bra, men i motsatt fall frarådes å benytte absolutte verdier. Problemet stikker i at vi intuitivt har lett for å multiplisere symbolets areal med dets verdi. Vi skal ta et eksempel. Produksjonen av elektrisk kraft i Japan og i tidligere Sovjetunionen var omtrent like store målt i totalt volum. Dersom vi lager et koropletkart basert på det totale volumet, vil de to landene bli fargelagt med samme gråtone, men det er uten videre klart at mange lett vil overvurdere produksjonen i Sovjetunionen i forhold til Japan på grunn av den store forskjellen i de to landenes geografisk utstrekning. Vi skal ta et annet eksempel. La oss anta at tallene i bilde a) i figur 4.2 representerer gjennomsnittlig snødybde innenfor de respektive rutene. Siden vi forutsetter at gjennomsnittshøyden er regnet relativt i forhold til arealet, kan tallene benyttes direkte i koropletkartet. Et problem er knyttet til valg av antall klasser. Som hovedregel kan det sies at antall klasser ikke bør være stort. Fire klasser er ofte nok. Et høvelig antall klasser kan sies å ligge i intervallet 3 til 6 klasser. Med dagens teknologi er det lett å lage klasseløse koropletkart, men dette anses vanligvis for å være en uheldig framstillingsmåte. Kartleseren vil likevel være på jakt etter differanser og klasser. For effektiv kommunikasjon av de geografiske mønstre, bør derfor denne analysen være gjort på forhånd og ikke overlates til kartleseren. I figur 4.2 har kartene b) og c) fire klasser mens kart d) har bare tre klasser.

Problemet knyttet til valg av klassegrenser er ikke uttømmende behandlet i den kartografiske litteraturen. Vi kan likevel formulere en hovedregel med utgangspunkt i at vi oppfatter antall klasser som en begrenset ressurs. Denne ressursen må vi derfor utnytte på en best mulig måte i det foreliggende kartet. Dette krever at vi gjør oss opp en mening om hvilke områder av variasjonsintervallet vi er mest interessert i å vite noe om. Dersom vi for eksempel er interessert i de snødekte områder som har liten snødybde, er det fornuftig å gjøre et valg der vi har små klasseintervaller i den nedre delen av variasjonsområdet og heller bruke noe større klasseintervaller i den øvre delen av variasjonsområdet. Se kartene b) og c) i figur 4.2 der kart b) vektlegger områder med liten snødybde mens kart c) vektlegger områder med stor snødybde.

Før klassegrensene fastlegges, er det viktig å danne seg et bilde av hvordan dataene er fordelt. Frekvenshistogram er i denne forbindelsen et nyttig verktøy. Ut av frekvenshistogrammet kan vi raskt avsløre om de valgte klassegrenser vil gi en eller flere tomme klasser. Tomme klasser bør ikke forekomme. I praksis benyttes gjerne følgende prinsipper for å fastlegge klasseintervallene:

Ekvidistant inndeling Klassene gjøres like brede, for eksempel 0-9.9, 10-19.9, 20-29.9.

Aritmetisk inndeling Klasseintervallene endrer bredde i henhold til en aritmetisk rekke, for eksempel 0-9.9, 10-29.9, 30-59.9, 60-99.9.

Geometrisk inndeling Klasseintervallene endrer bredde i henhold til en geometrisk rekke, for eksempel 0-9.9, 10-29.9, 30-69.9, 70-149.9.

Logaritmisk inndeling Klasseintervallene endrer seg i henhold til en logaritmeberegning, for eksempel 0.1-1.0, 1.1-10.0, 10.1-100.0. Dette er nyttig for tallserier med veldig stor sprednig.

Uregelmessig varierende klassebredder Dette er nyttig dersom vi ønsker å dele datamaterialet inn i naturlige grupper, d.v.s. grupper som framkommer ved markante sprang i frekvenshistogrammet.

Klasseinndeling med likt antall observasjoner i hver klasse

Klasseinndeling etter middelverdi og standardavvik I de tilfeller at datane er normalfordelte, kan en inndeling basert på middelverdi μ og standardavvik σ være hensiktsmessig. Vi kan for eksempel få klassegrenser i henhold til $\mu \pm 1\sigma, \mu \pm 2\sigma, \mu \pm 3\sigma$.

En svakhet ved koropletkart er at de gir inntrykk av homogen fordeling innenfor arealenheten. For eksempel i en koroplekartframstilling av saueholdet i Norge basert på fylker som geografiske enheter, vil vi få inntrykk av at saueholdet har en jevn utbredelse innenfor hvert fylke. Vi får dessuten inntrykk av at det er skarpe overganger mellom fylkene, men dette er ikke tilfellet. Overganger mellom områder med stort sauehold og områder med lite sauehold er glidene og ikke sprangvise.

4.4 Topografiske kart

Topografiske kart er kanskje den kartypen som er best kjent av folk flest.

Definisjon 7 *Et topografisk kart viser formen til en terrengoverflate samt et utvalg av de objekter på terrengoverflaten som er lett synlige (hus, veier, vann osv.).*

Den person som i nyere tid har bidratt mest til fremme av utformingen av topografiske kart, er avdøde professor Dr. Eduard Imhof, Sveits. Hans mange glimrende illustrasjoner, kart, artikler og bøker bør være en kilde til inspirasjon og kunnskap også for dagens kartografer. Vi vil her på det sterkeste anbefale hans bok [Imh65] (boken finnes også i en engelsk oversettelse).

4.4.1 Høydekurver og dybdekurver

Høydekurver har sin anvendelse for kart i målestokker større enn 1:250 000. Dersom målestokken blir mindre, blir ekvidistansen (den vertikale avstand mellom høydekurvene) så stor at kurvene uttrykker lite om terrengoverflatens topografi. Dette forhold lar seg lett demonstrere med utgangspunkt i et resonnement om grafisk minste mulige ekvidistanse.

Optimal ekvidistanse

La oss anta terrenghelningen α, en strektykkelse på 0.1mm og at linjer som ligger nærmere hverandre enn 0.4mm ikke er separerbare (den tilhørende avstand mellom kurvenes senterlinjer blir i dette tilfellet 0.5mm). Den grafisk minste mulige ekvidistanse e er under de foregående forutsetninger gitt ved

$$e = \frac{0.5 \cdot M \cdot \tan \alpha}{1000} \text{meter} \tag{4.3}$$

hvor M er målestokkstallet. Basert på likning 4.3 angir [Imh65] verdier for ekvidistansen gitt i tabell 4.2. Tallene er rundet av til tallverdier som er lette å huske. På grunn av at det er ønskelig med så mange høydekurver som mulig for å få en god beskrivelse av terrengoverflatens topografi, baserer tabellen seg på minste mulige ekvidistanse. Imidlertid er for de små målestokker angitt en mindre ekvidistanse enn den som kan utledes fra likning 4.3. Grunnen til dette ligger i en optimaliseringstanke om å maksimere antall kurver i forhold til kravet om at kurver ikke skal flyte sammen.

Mellomkurver

Et problem med bruk av høydekurver er at det er de bratteste terrengområder som blir dimensjonerende for ekvidistansen. I praksis er det ofte slik at det er de minst bratte partier som er av størst økonomisk interesse, og følgelig burde disse områdene også vært med å bestemme ekvidistansen. Det endelige valget av ekvidistanse vil måtte bero på en samlet vurdering av de topografiske forhold i det landområdet som skal kartlegges.

I enkelte tilfeller kan bruk av mellomkurver løse problemet for de flate terrengområder, men mellomkurver har den ulempen at de kan gi inntrykk av brattere terreng enn det som er tilfellet. Avvik fra den normale ekvidistansen bør derfor bare

Tabell 4.2: Anbefalt ekvidistanse for kart i ulike målestokker (fra [Imh65]). Tallene i parantes angir alternative valg.

Målestokk	Ekvidistanse i meter		
	Fjelland $\alpha = 45°$	Middels bratt $\alpha = 26°$	Flatt land $\alpha = 9°$
1:1000	1	0.5	0.25
1:2000	2	1.0	0.5
1:5000	5	2	1
1:10 000	10	5	2
1:20 000	20	10	2.5
1:25 000	20	10	2.5
1:50 000	(20) 30	(10) 20	5
1:100 000	50	25	(5) 10
1:200 000	100	50	10
1:250 000	100	50	10 (20)
1:500 000	200	100	20
1:1 000 000	200	100	20 (50)

forekomme i spesielle situasjoner. Under ingen omstendigheter bør mellomkurvene trekkes lengere enn det som er nødvendig for å tilføre informasjon som ikke kan leses ut fra de normale kurvene. Hvis for eksempel en mellomkurve i et parti går midt mellom to normale kurver, bør denne delen av mellomkurven fjernes.

Generalisering av høydekurvene

For generalisering av høydekurvene kan følgende regler anføres:

1. isolerte småbevegelser fjernes dersom de virker forstyrrende,
2. for kurver som berører hverandre foretas en forsiktig adskillelse,
3. ved stor ekvidistanse fjernes flere småbevegelser enn ved liten ekvidistanse,
4. graden av generalisering økes med avtagende datanøyaktighet,
5. overalt i kartet benyttes de samme generaliseringsregler.

Tegnetekniske normer

Vanligvis tegnes hver femte kurve med tykkere strek enn de øvrige kurvene. Disse kurvene kalles for tellekurver og påføres høydetall. Høydetallene plasseres der det passer best av hensyn til annet kartinnhold og skrives opprett med midtlinjen i

Tabell 4.3: Anbefalt strektykkelse til høydekurver (fra [Imh65]). Verdier er angitt i mm.

Målestokk	Tellekurver heltrukne			Normalkurver heltrukne			Mellomkurver stiplet		
	svart	blå	brun	svart	blå	brun	svart	blå	brun
1:1000	0.25	0.30	0.35	0.15	0.18	0.20	0.10	0.13	0.15
1:10 00	0.20	0.25	0.30	0.10	0.15	0.18	0.07	0.09	0.10
1:25 000	0.18	0.20	0.24	0.10	0.12	0.15	0.05	0.06	0.08
1:50 000	0.15	0.18	0.20	0.08	0.10	0.12	0.03	0.04	0.05
1:100 000	0.10	0.13	0.15	0.08	0.10	0.10	0.03	0.04	0.05

kurvens retning. Mellomkurver skal være underordnet normalkurvene. De tegnes derfor med tynnere strek eller med stiplet linje.

Av og til kan det være vanskelig å skille mellom lukkede toppkurver og lukkede bunnkurver. Det kan brukes tilleggstegn for å skille mellom de to tilfellene, for eksempel ved korte streker som viser fallretningen eller ved + og - tegn. En annen løsning ligger i å skrive på høydetall der hvor det forekommer lukkede kurver. Det er særlig for sjøkart at problemet med bunnkurver og toppkurver er størst. På land vil andre objekter som vannsystem være med å understreke fallretninger og gjøre kurvebildet lettere å forstå.

Kurvenes farger kan rette seg etter det materialet terrengoverflaten er dekket av. Vanlige kurvefarger er blå (sjø), brun (land) og svart. Tabell 4.3 gir anbefalte linjetykkelser for høydekurver ved trykking av kart. På grunn av at vår oppfatning av bredden til en linje er avhengig av dens farge, er de anbefalte linjetykkelser korrigert for denne psykologiske faktoren. For eksempel vil en svart linje på 0.25mm synes like bred som en blå linje på 0.30mm eller en brun linje på 0.35mm.

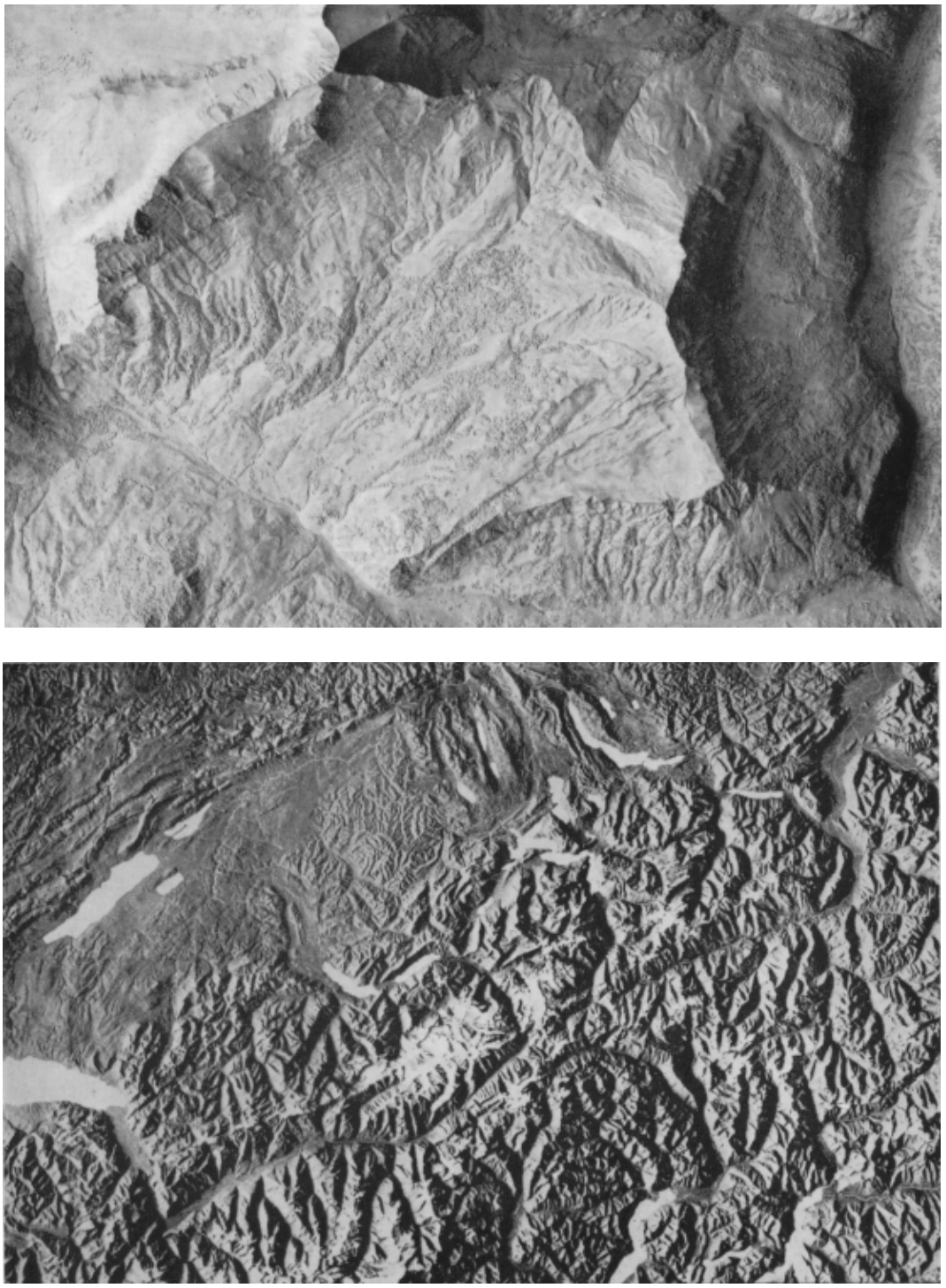
Synbarheten til en stiplet linje er avhengig av tre faktorer: linjetykkelsen t og forholdet mellom lengden på strekene s_s og lengden på gapet s_g i stiplingen. Følgende formel angir et gunstig forhold mellom de tre nevnte størrelser:

$$t : s_s : s_g = 1 : (8 \text{ til } 12) : (4 \text{ til } 6) \tag{4.4}$$

hvor t er angitt i 1/10mm og s_s og s_g er gitt i mm. For punkterte linjer bør avstanden være 1-1.5 ganger prikkestørrelsen.

4.4.2 Skyggeleggingsteknikker

Ved ulike skyggeleggingsteknikker kan vi gi inntrykk av en tredimensjonal modell selv om tegningen bare er 2D.



Figur 4.3: Eksempler på kartografisk skrålysskygge, fra [Imh65] side 198-199

Skyggelegging

Skyggelegging er ikke noen eksakt metode, men metoden har den fordel at vi kan gjengi små variasjoner i terrengoverflaten, se bildene i figur 4.3. Skyggelegging benyttes derfor gjerne på kart i små målestokker eller i kombinasjon med høydekurver i kart i de målestokker som egner seg for høydekurver. I dag kan vi gjøre beregninger av gråtonen ut fra en terrengmodell. For å oppnå maksimal 3D-effekt er det for enkelt å basere seg på kun en enkelt lyskilde og regne gråtonen som en funksjon av cos til vinkelen mellom solvektoren og flatenormalen. Forsøk har vist at 3D-effekten kan forbedres ved å regne gråtonen som et gjennomsnitt av flere lyskilder. Skyggelegging i kombinasjon med fargelagte høydesjikt, kan gi gode 3D-effekter, se eksempelkart hos [Imh65].

Skravur

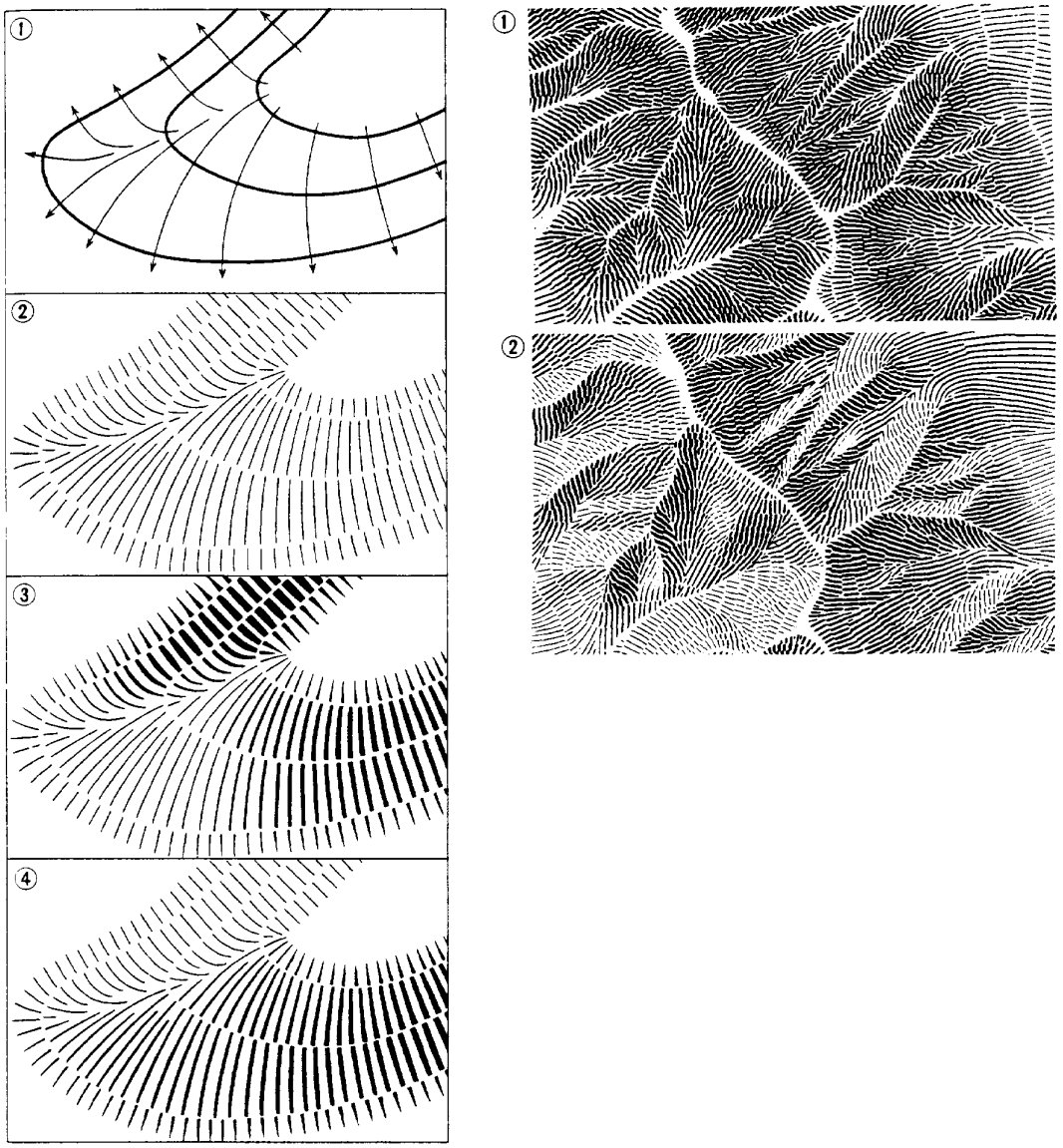
Tidligere ble skravur benyttet for å lage 3D-effekter. Teknikken brukes lite i dag. Dette skyldes muligens at det er noe vanskelig å automatisere prosessen. Dessuten vil skyggelegging stort sett dekke det samme bruksområdet som skravur. Skravur har den fordel i forhold til skyggelegging at skravur over alt viser fallretningen og at streklengden er en funksjon av helningen. Skravur fråmstår derfor som en mere eksakt metode enn skyggelegging. En ulempe med skravur i forhold til skyggelegging er at skravur ikke kan gjengi så små variasjoner i terrengformen som skyggelegging kan gjøre. Prinsippene for skravur er illustrert i figur 4.4.

Bergsymboler og symboler for skrent

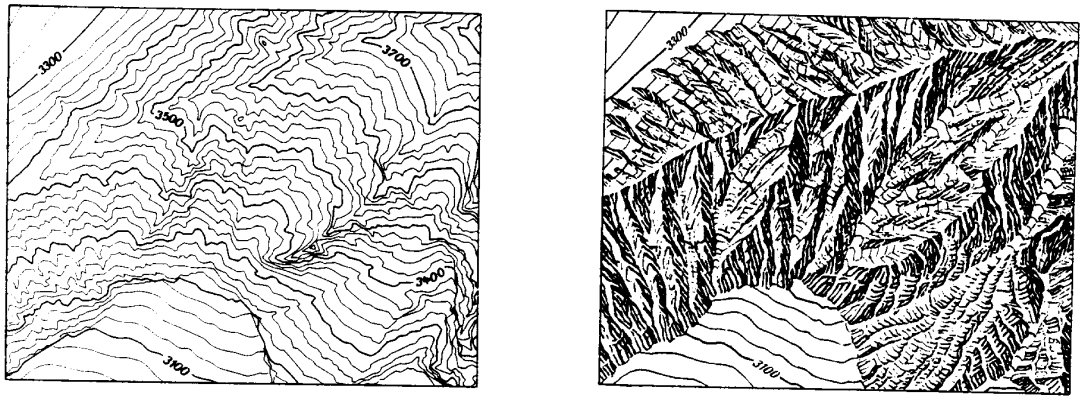
I bratte partier kan benyttes spesielle symboler. På sveitsiske kart er en slik teknikk utviklet til et høyt kunstnerisk nivå. Se figur 4.5 som illustrerer teknikken. Se forøvrig [Imh65] for en nærmere beskrivelse av teknikken. Metoden brukes lite i dag, noe som antakelig skyldes at metoden er vanskelig å automatisere.

Høydekurver med forsterket 3D-effekt

Det har vært gjort flere forsøk med å utforme høydekurvene slik at skaren av kurver gir en tredimensjonal virkning. En slik teknikk går ut på å variere tykkelsen til høydekurvene. Tykkelsen kan varieres etter to prinsipper: (1) strektykkelsen øker med høyden over havet og (2) strektykkelsen øker i skyggepartier i forhold til en tenkt skrå belysning av terrengmodellen. En noe annen teknikk går ut på å la strekfargen variere mellom hvit og svart mot en mørk grå bakgrunn. Denne metoden kan i helt spesielle tilfeller gir gode 3D-effekter.



Figur 4.4: Prinsippene for skravur, fra [Imh65] side 241 og side 246-247. Rammene 1-4 til venstre i figuren viser konstruksjonsprinsippene for skravur. Her viser ramme 1 og 2 at skravuren ordnes i bånd som følger høydekurvene og at skravuren står vinkelrett på disse. I ramme 3 er tenkt en vertikal belysning av terrengoverflaten (skråningsskravur) mens i ramme 4 er belysningen skrå (skyggeskravur). Gråtonen til skravuren framkommer ved at bredden på de svarte strekene varierer i henhold til det valgte belysningsprinsippet. Summen av den svarte strekens bredde og avstanden til nabostreken holdes konstant. Rammene 1 og 2 til høyre i figuren gir eksempler på henholdsvis skråningsskravur og skyggeskravur.



Figur 4.5: Eksempel på bruk av bergsymbol, fra [Imh65] side 296

4.4.3 Fargelagte høydesjikt

På topografiske kart i små målestokker anvendes ofte en koropletmetode for å vise variasjonen i terrengets høydeforhold. Problemet er som for koropletkart knyttet til det å finne et passende antall klasser og bestemme klassegrensene. Dette problemet er diskutert hos [Imh65] og han viser at klassegrensene vil avhenge av om det er terreng på land eller sjøbunnen som skal kartlegges. Ved å basere seg på en vurdering om at informasjon om grunne sjøområder og lavereliggende områder på land er av større interesse for kartbrukeren enn sjøbunnen på store havdyp og høyereliggende områder på land, stilles opp tabell 4.4. I tabell 4.4 er det benyttet en geometrisk rekke for

Tabell 4.4: Eksempel på klasseinndling i kart med høydelagte fargesjikt

Klassegrenser i meter	
Land	Hav
0	5
50	10
100	20
200	50
500	100
1000	200
2000	1000
4000	2000
	3000
	4000
	5000

klassegrensene på land mens det for havområdene er benyttet en kombinasjon av en

geometrisk rekke og en ekvidistant rekke (skillet mellom de to sistevnte prinsippene er i tabellen markert med en horisontal strek).

Vi må også velge prinsipp for fargelegging av de ulike fagesjiktene. For sjøområder er det vanlig å benytte en skala i blått der fargens lyshet endres. For land kan vi gjerne bruke et prinsipp om at fargen skal være naturtro (slik vi opplever fargen til terrengoverflaten en sommerdag). Det siste prinsippet gir den anbefalte skalaen: blågrønn, grønn via gult til brun og rødbrun. Skalaen kan avsluttes med hvitt dersom vi har breområder over en viss høyde. Vi anbefaler også her illustrasjonene hos [Imh65].

4.4.4 Andre teknikker for å beskrive terrengets topografi

Vannsystem

Nettverket av elver, bekker, kanaler og sjøer gir informasjon om terrengets høydeforhold. I områder med rikt utviklet vannsystem kan vannsystemet derfor benyttes som selvstendig element for å vise hovedtrekkene i terrengets topografi. Metoden kan for eksempel anvendes for veikart i kombinasjon med metoder for å visualisere fjelltopper og markante fjellrygger.

Høyde og dybdetall

Dybdetall benyttes i stror utsrekning på sjøkart. Her gjøres et lite utvalg av alle de dybdemålinger som ligger til grunn for kartleggingen. Prinsippet for utvelgelsen er at det det skal være tilnærmet jevnt fall mellom to nærliggende dybdeangivelser. Når høydetallet skrives med desimalpunktum, kan det i enkelte tilfeller være hensiktsmessig å la desimalpunktumet komme på det sted høyden referer til. Det er også mulig å la tyngdepunktet for tallet angi det sted høyden referer til (vanlig på sjøkart). Se [Imh65] for nærmere angivelse av tegnetekniske normer.

Kapittel 5

Modeller for kartografisk kommunikasjon

I dette kapitlet vil vi presentere overordnede modeller av prosessen kartografisk kommunikasjon. De ulike modellene framhever hver for seg spesielle sider ved kartografisk kommunikasjon. Tilsammen utfyller modellene hverandre.

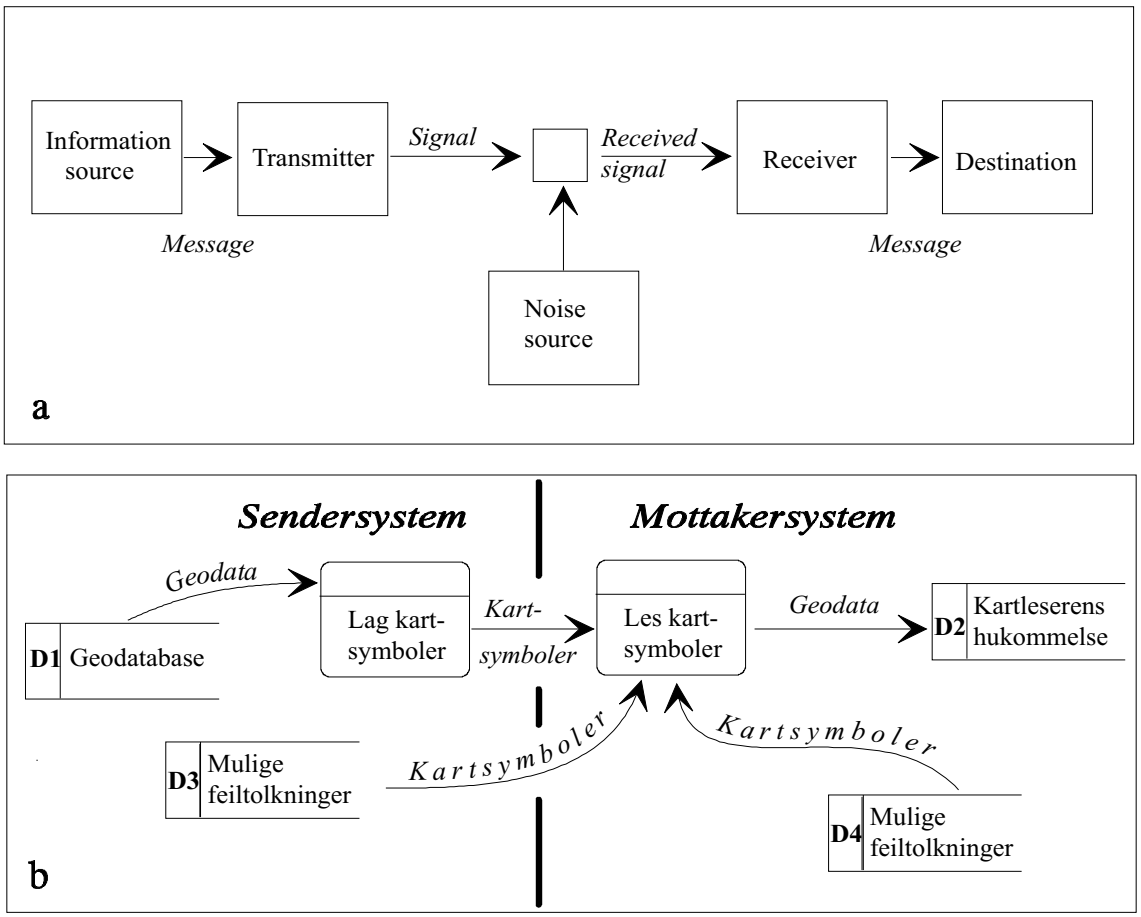
5.1 Shannon og Weavers kommunikasjonssystem

Kommunikasjonssystemet etter Shannon og Weaver [SW64] er gjengitt i bilderamme a) i figur 5.1. I bilderamme b) er kommunikasjonssystemet modifisert ved at vi har benyttet dataflytdiagram og foretatt en tilpasning til den kartografiske problemstillingen. Shannon og Weavers system er generelt og gjelder i prinsippet alle former for kommunikasjon. Systemet består av tre delsystemer: (1) sendersystem, (2) mottakersystem og (3) et støysystem. Informasjonskilden velger en bestemt melding fra en mengde av mulige meldinger og omformer denne til et signal som sendes over kommunikasjonskanalen. Støykilden modifiserer signalene i en uønsket retning. På 1970 tallet ble det gjort flere forsøk på å tilpasse Shannon og Weavers kommunikasjonsteorier til kartografiske problemstillinger uten at man lyktes i særlig grad. Årsakene til dette er flere, men en viktig årsak er at man ikke var tilstrekkelig nøye med å definere hvilket kommunikasjonsnivå Shannon og Weavers teorier skulle anvendes på. Vi kommer tilbake til dette i kapitlet om informasjonsteori. Shannon og Weaver presiserer at kommunikasjon kan sies å bestå av tre nivåer.

Syntaktisk nivå Hvor nøyaktig lar symboler seg overføre.

Semantisk nivå Hvor presist er symbolene i stand til å overføre meningsinnholdet i meldingen.

Pragmatisk nivå Hvor effektivt lar den mottatte mening seg utnytte til å gjøre beslutninger om det foreliggende problem eller i den foreliggende situasjon.



Figur 5.1: Bilde a) viser Shannon og Weavers kommunikasjonssystem. I bilde b) er Shannon og Weavers system beskrevet ved hjelp av dataflytdiagram og tilpasset en kartografisk problemstilling.

Begrepene data og informasjon er velkjente. Data blir betraktet som bærere (bitmønstre) av informasjon, mens det meningsinnhold som tillegges dataene kalles for informasjon. Det syntaktiske nivå dreier seg derfor om kommunikasjonsproblemer på datanivået mens det semantiske nivå dreier seg om kommunikasjonsproblemer på informasjonsnivået. Her er vi ved en kjerne i forståelsen av Shannons informasjonsteori. Shannons informasjonsteori ble i utgangspunktet formulert med tanke på å modellere kommunikasjon på det syntaktiske nivået, men begrepet informasjonsteori leder oss lett til å tenke på kommunikasjon på det semantiske nivået. Selv om teorien i prinsippet kan anvendes på alle de tre kommunikasjonsnivåene, må vi for kartografiske anvendelser innse at teorien er lite egnet for det semantiske og det pragmatiske nivået. Teorien lar seg imidlertid anvende på det syntaktiske kommunikasjonsnivået. Ved hjelp av Shannons formel

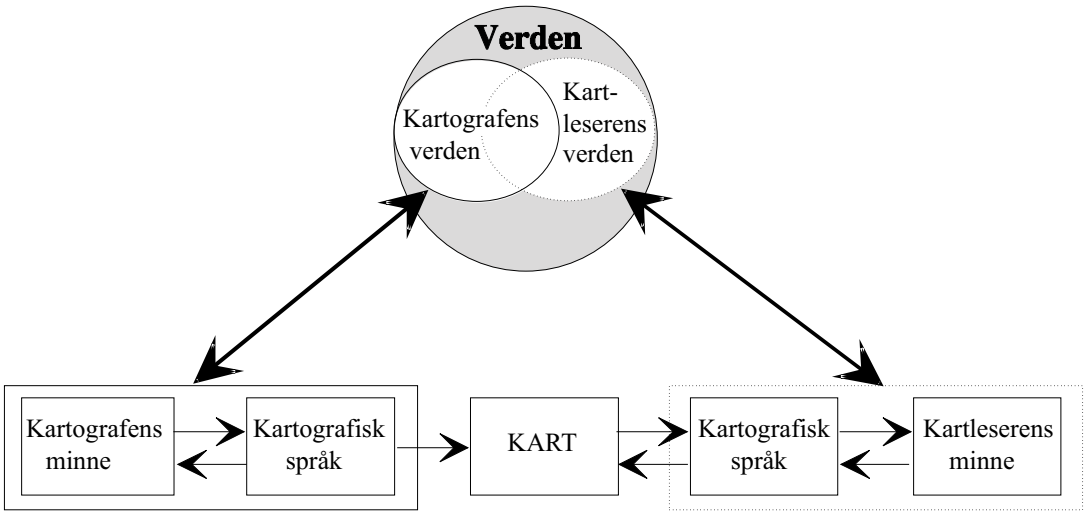
$$H = \sum_i p_i \log_2 \frac{1}{p_i} = -\sum_i p_i \log_2 p_i$$

beregnes en indeks for variasjonen i en datamengde. Shannon og Weaver kaller målenheten for *bits of information*.

Støykilden er et sentralt element i Shannon og Weavers modell. I bilderamme b) i figur 5.1 er støykilden representert ved de to datalagrene D3 og D4. En vertikal linje deler systemet i et sendersystem og et mottakersystem. De to datalagrene D3 og D4, er plassert i hvert sitt delsystem. Dette er gjort for å markere at grunner til feiltolkninger av kartsymboler kan ha sin rot i mangler ved selve kartsymbolene (sendersystemet) og visuelle begrensninger hos kartleseren (mottakersystemet). Mangler ved sendersystemet kan eksemplifiseres ved at en gitt fargetone ikke er korrekt gjengitt på kartet. Feiltolkninger som følge av begrensninger til mottakersystemet kan for eksempel være at øyet ukorrekt gjenkjenner en farge i kartet.

5.2 Koláčnys diagram

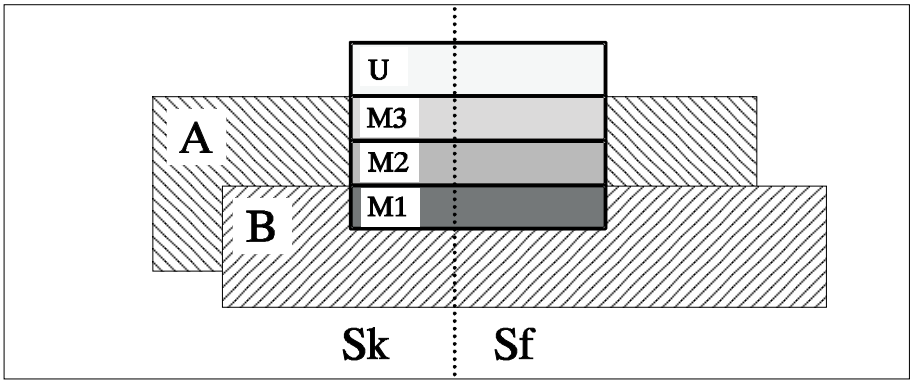
En forenklet versjon av Koláčnys diagram [Kol69] er vist i figur 5.2. Koláčnys diagram har en del til felles med Shannon og Weavers modell. En forskjell er at Koláčnys diagram ikke modellerer at kartleseren kan feiltolke deler av kartinnholdet. På en annen side beskriver diagrammet at det er forskjell mellom kartografens og kartleserens kunnskap om virkeligheten. Noe som ikke direkte framgår av diagrammet, men som understrekes av Koláčny, er at design og bruk av kart må betraktes som et hele. Derfor må kartografen være opptatt av såvel bruken av kart som dets konstruksjon.



Figur 5.2: Forenklet versjon av Koláčnys diagram over kartografisk kommunikasjon

5.3 Robinson og Petcheniks modell

Robinson og Petchenik [RP76] anvender Venndiagrammet i figur 5.3 for å beskrive relasjoner mellom kognitive komponenter i en kartografisk kommunikasjon. De deler geografisk kunnskap i de to mengder Sk og Sf, korrekt og feilaktig kunnskap om objekter i det geografiske rom. Mengdene A og B representerer den kunnskap som



Figur 5.3: Venndiagram etter Robinson og Petchenik

innehas av henholdsvis kartografen og kartleseren. Et rektangel M, som betegner et gitt kart, er delt inn i delmengdene M_1, M_2, M_3 og U. M_1 er den delen av M kartleseren visste fra før. Denne delen av kartet representerer derfor ikke ny kunnskap for kartleseren. M_2, som oppfattes av kartleseren, er den delen av M kartleseren ikke visste fra før. M_2 representerer derfor ny kunnskap for kartleseren. M_3 er den andel av M kartleseren ikke oppfatter. Mengden U ligger utenfor både A og B og representerer derfor en indirekte økning i geografisk kunnskap. I logikken dekkes dette forholdet av begrepet induksjon. Med induksjon menes å etablere ny kunnskap på grunnlag av allerede eksisterende kunnskap. Siden U gjør at kartet kan inneholde mere informasjon enn kartografen er klar over, vil kartografen ved å ta kartleserens plass, kunne utvide sin kunnskapsmengde.

Robinson og Petcheniks modell har likhetspunkter med de foregående modellene, men det er også ulikheter vi skal merke oss. Selv om Robinson og Petcheniks modell operer med mengdene Sk og Sf, ligger ikke her konseptet om en støykilde slik som i Shannon og Weavers modell. Robinson og Petcheniks modell fokuserer på kvaliteten til den informasjonen som skal overføres mens Shannon og Weaver antar at informasjonskilden er feilfri, men at feil kommer inn gjennom selve overføringen av informasjonen. I kartografisk sammenheng må vi ta hensyn til begge konseptene:

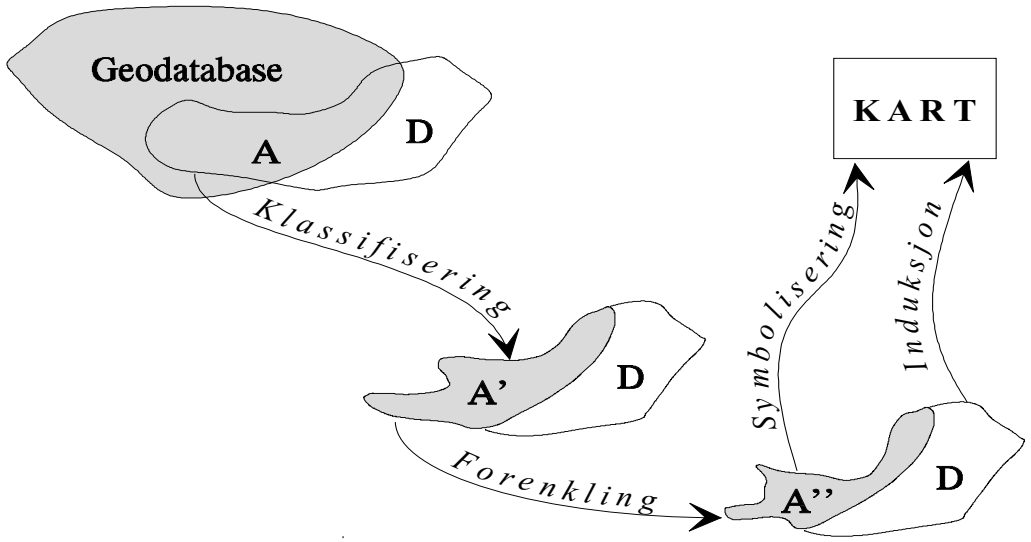
1. kvaliteten til datagrunnlaget (nøyaktighetskriterier, fullstendighet m.m.);

2. feil som skyldes ufullkommenheter til kommunikasjonskanalen.

De fire komponentene M er delt inn i, er sentrale i Robinson og Petcheniks modell, men modellen sier ikke noe om optimale størrelser til disse komponentene. Vi vil for

eksempel lett tenke at M_1 bør være lik $\emptyset$, men dette vil ikke alltid være ønskelig. La oss anta et befolknigskart over Norge. For å ha en geografisk referanse for temaet, lager vi et kart over Norge der kystlinjen og riksgrensen er tegnet inn. De fleste vil si at mye av denne informasjonen tilhører mengden M_1, men informasjonen er likevel en nødvendig del av vårt temakart. På den annen side må vi ikke fylle opp kartet med selvfølgeligheter eller trivialiteter.

5.4 Morrisons modell



Figur 5.4: Overføring av informasjon fra geodatabase (kartografens erkjennelsesområde) til kart, basert på Morrison (1976).

Morrison [Mor76] dekomponerer kartografisk kommunikasjon i følgende delprosesser:

1. velge ut data fra kartografens erkjennelsesområde,
2. klassifikasjon og forenkling av dataene,
3. symbolisering av dataene for å danne det fysiske kartet,
4. induksjon eller induktiv generalisering i samspill med symbolisering,
5. oppdage symboler på kartet,
6. bedømme ulikheter mellom kartsymbolene,
7. gjenkjenne og vurdere det aktuelle symbolet.

Prosessene 1) til 4) tilhører sendersiden i kommunikasjonssystemet og er illustrert i figur 5.4. Prosessene 5) til 7) tilhører mottakersiden. Morrison [Mor74, Mor76] definerer kartografisk generalisering som unionen av klassifisering, forenkling, symbolisering og induktiv generalisering (induksjon). I figur 5.4 velges datamengden A fra geodatabasen. Vi antar at det finnes en del informasjon om de valgte geografiske forekomster som ikke er beskrevet i databasen. Denne mengden er representert ved mengden D. Etter at utvalget er gjort, følger så klassifikasjon. Klassifikasjon kan bare gjøres på mengden A og influerer ikke på D. Etter klassifikasjonen er noe informasjon gått tapt. Dette er illustrert ved at $A' \subset A$. Deretter forenkles A'. Dette gir oss mengden A'' som er en delmengde av A', d.v.s. noe informasjon er gått tapt gjennom forenklingen. Til slutt følger symbolisering og induktiv generalisering (induksjon). Induksjon i Morrisons modell svarer til induksjon i Robinson og Petcheniks modell i figur 5.3, men Morrison benytter induksjon på en noe mere aktiv måte enn Robinson og Petchenik. Morrisons modell sier at informasjon kartleseren allikevel vil resonnere seg fram til, ikke behøver å bli representert eksplisitt i kartet (induktiv generalisering). Anta for eksempel et veikart over Norge. I områder som er markert som byer og tettsteder, vil vi av erfaring slutte oss til at her har vi bensinstasjoner, spisesteder osv.. De tilhørende karttegn kan derfor utelates fra kartet, siden informasjonen likevel ligger implisitt i kartet. Derimot på øde veistrekninger er det nyttig med karttegn som informerer om bensinstasjoner og spisesteder. Her kan vi ikke si at den aktuelle informasjonen ligger implisitt i kartet. Ved induksjon kan kartleseren også ekstrahere informasjon som ikke finnes i geodatabasen, altså vil det ved induksjon være mulig å gjøre slutninger om mengden D. Induksjon kan følgelig brukes på to måter:

1. induktiv generalisering for å representere informasjon implisitt i kartet (vi kan redusere antall kartsymboler);

2. induksjon for å etablere ny kunnskap om de geografiske forekomster som ligger lagret i geodatabasen.

Disse to aspektene ved induksjon kommer ikke klart fram av Morrisons modell i figur 5.4. Et problem knyttet til induksjon er selvsagt at kartografen ikke med sikkerhet kan vite hvilke slutninger kartleseren vil trekke. La oss anta et veikart og at vi ikke symboliserer spisesteder i byer, fordi vi mener denne informasjonen ligger implisitt i symbolet for by. Kan det likevel tenkes at noen vil kjøre forbi Lillehammer fordi de vil tro at det ikke er spisesteder i byen og heller tar matpausen på Dovrefjell, fordi her har kartet et tegn for spisested?

Morrison [Mor74] anvender egenskaper som surjektive og injektive funksjoner for å klassifisere prosesser innenfor kartografisk kommunikasjon.

Surjektiv (på) Dersom f er en funksjon fra A til B og billedområdet til A er B, er f surjektiv og vi sier at f avbilder A *på* B. Sagt på en annen måte $f : A \rightarrow B$ er surjektiv dersom alle elementer i B kan forklares ved en avbildning fra A, altså dersom $f(A) = B$.

Injektiv (en-entydig) Dersom $f(x) = f(y) \Rightarrow x = y$, er f injektiv (en-entydig).

Bijektiv En funksjon som er både surjektiv (på) og injektiv (en-entydig), sies å være bijektiv.

Tabell 5.1: De 16 prosesser i Morrisons klassifikasjonssystem. Prosesser med uthevet tekst er tolket av Morrison.

	f		f^{-1}		
nr.	(1:1)	(på)	(1:1)	(på)	tolkning
1	+	+	+	+	*Utvalg*
2	+	+	+	-	*Forenkling*
3	+	+	-	+	
4	+	+	-	-	Støyprosess på mottakersiden
5	-	+	+	+	
6	-	+	+	-	
7	-	+	-	+	*Klassifikasjon*
8	-	+	-	-	Klassifikasjon med informasjonstap
9	+	-	+	+	*Symbolisering*
10	+	-	+	-	
11	+	-	-	+	
12	+	-	-	-	Fra kontinuerlig til diskret modell
13	-	-	+	+	
14	-	-	+	-	Støyprosess på sendersiden
15	-	-	-	+	
16	-	-	-	-	*Induksjon*

Anta en funksjon f og dens inverse funksjon f^{-1}. Den inverse funksjon kan vi tolke som en prosess som skal gjøres av kartleseren. En funksjon er enten injektiv eller så er den ikke injektiv. Tilsvarende for surjektiv egenskap. Dette gir oss 16 kombinasjonmuligheter for f og f^{-1} som vist i tabell 5.1. Morrison tolker fem av disse kombinasjonene. De fem prosessene, utvalg, klassifikasjon, forenkling, symbolisering og induksjon, er grunnlaget for modellen i figur 5.4.

Utvalg er en prosess som er både injektiv og surjektiv (den er bijektiv). Det at den inverse funksjon også er bijektiv, tolkes slik at ikke noe informasjon har gått tapt gjennom utvalget. Morrison definerer klassifikasjon som $(-+-+)$, men dersom vi skal oppfatte f^{-1} fra kartleserens side, må det sies at det i praksis ikke vil være mulig for kartleseren å rekonstruere det opprinnelige datasettet ut fra det klassifiserte datasettet. Noe informasjon har gått tapt gjennom klassifikasjonen. Derfor er prosess nr. 8 $(-+--)$ mere beskrivende for klassifikasjon som kartografisk prosess enn prosess nr. 7 $(-+-+)$. Prosess nr. 2 $(+++-)$ kaller Morrison forenkling. Det at den inverse funksjon ikke er surjektiv, betyr at kartleseren ikke fullt ut klarer å rekonstruere det opprinnelige datasettet fra det forenklede datasettet. Noe

informasjon har gått tapt i forenklingen, men dette er også noe av hensikten med kartografisk generalisering, nemlig å redusere informasjonsmengden. Forenkling defineres som å bestemme de viktige egenskaper til dataene, elimineringen av uønskede detaljer, bibehold og eventuelt overdrivelse av de karakteristiske egenskaper. Prosess nr. 9 $(+ - ++)$ har kjennetegnene til symbolisering. Det at funksjonen ikke er surjektiv, betyr at symbolisering alene ikke kan forklare all informasjon som finnes i kartet. Induksjon, prosess nr. 16 $(- - --)$, vil sammen med symbolisering forklare informasjonsinnholdet i kartet. Symbolisereing alene representerer informasjon som er eksplisitt uttrykt i kartet mens induksjon ekstraherer informasjon som ligger implisitt i kartet.

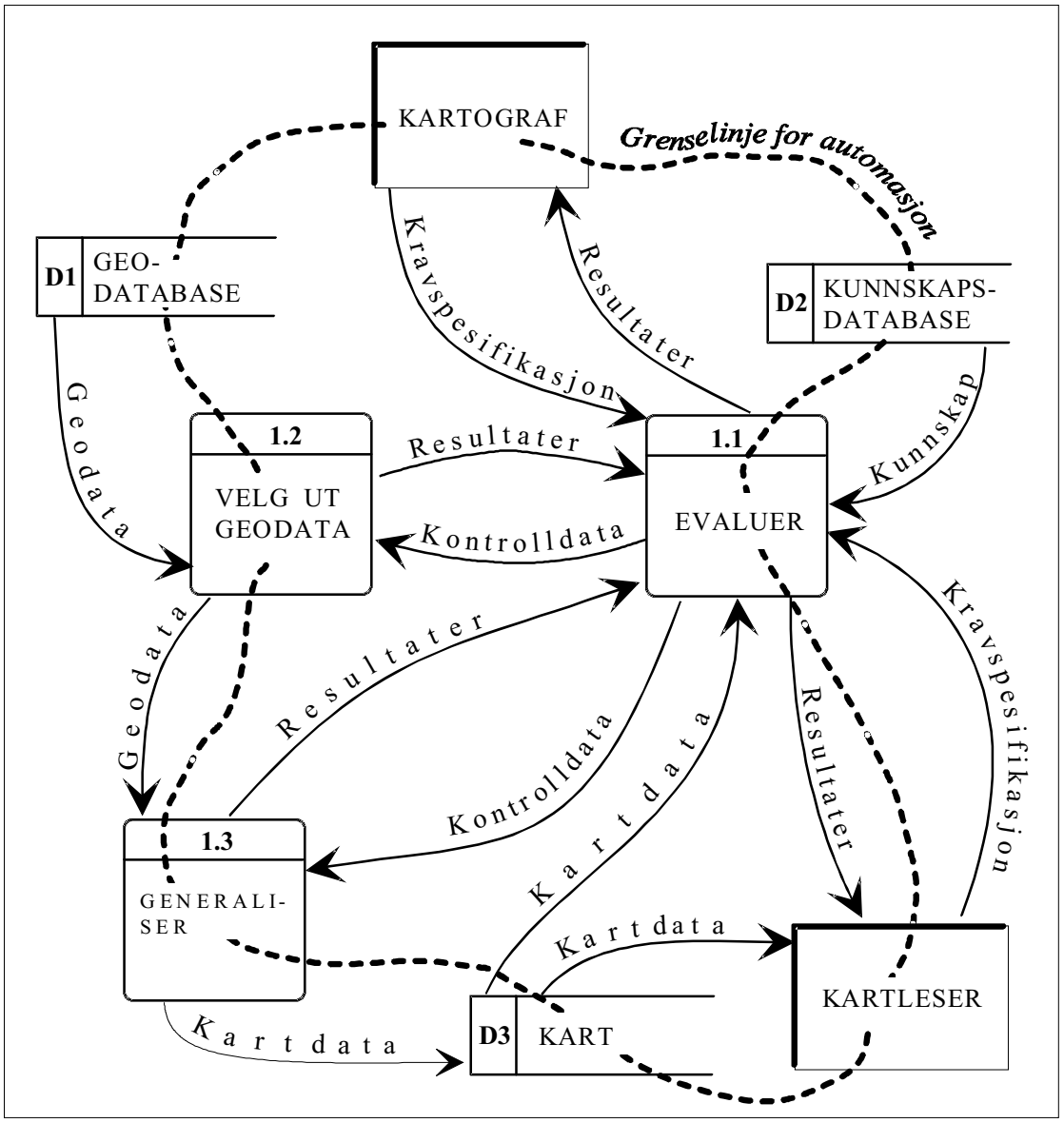
Vi kan spørre oss om Morrison har funnet alle de delprosesser det er naturlig å dele kartografisk kommunikasjon inn i. Jeg tror det må være riktig å betrakte Morrisons klassifikasjonssystem kun som et hjelpemiddel til å identifisere delprosesser og at vi ikke har noen garanti for at Morrison har funnet alle delprosesser innenfor kartografisk kommunikasjon. Siden Morrison ikke har tolket alle prosessene i tabell 5.1, vil vi forsøke å tolke de resterende prosessene for å se om det kan sette oss på sporet av andre viktige egenskaper til kartografisk kommunikasjon.

Prosess nr. 4 $(+ + --)$ har kjennetegnene til en støyprosess på mottakersiden. Dersom vi tenker oss at f gir signalet i Shannon og Weavers modell, vil f^{-1} svare til mottatt signal. Det at den inverse funksjon er $1 : N$ tolker vi slik at mottaker blander sammen kartsymboler. Det at den inverse funksjon ikke er surjektiv betyr at noe informasjon har går tapt som følge av sammenblandingen. Informasjonstapet og sammenblandingen tolker vi som et resultat av en støyprosess på mottakersiden. Prosess nr. 14 $(--+-)$ tolker vi også som en støyprosess, men denne gangen ligger støyprosessen på sendersiden. La oss for eksempel anta at under symboliseringen blir det ved en feiltagelse valgt samme linjefarge og linjetype for informasjonsvariablene skiløype og kommunegrense. Prosessen, som er 1:N, vil føre til tap av informasjon. La oss videre anta at funksjonen ikke er surjektiv. Dette kan begrunnes ut fra at en støyprosess aldri vil opptre som en selvstendig prosess, men at den inngår i andre prosesser. Slik sett kan vi si at det finnes noe i det transformerte datasettet støyprosessen alene ikke kan forklare.

Prosess nr. 5 $(- + ++)$ kjennetegner en prosess som etablerer en kontinuerlig modell fra diskrete elementer. For eksempel det å lage en digital terrengmodell fra en punktsky.

Ved at vi har forsøkt å tolke alle prosesser i Morrisons klassifikasjonssystem, har vi kommet opp med det viktige bidraget at vi har identifisert støyprosesser både på sendersiden og mottakersiden i kommunikasjonssystemet. Disse støyprosessene er representert ved datalagrene D3 og D4 i bilde b) i Shannon og Weavers kommunikasjonssystem i figur 5.1 En svakhet med Morrisons klassifikasjonssystem er at vi har en rekke delprosesser innenfor kartografisk generalisering som ikke er blitt identifisert. Dette skyldes ikke at det er noe galt med klassifikasjonssystemet, men at det er for grovmasket. Vi trenger med andre ord ytterligere klassifikasjonskriterier for å kunne identifisere alle delprosesser innenfor kartografisk kommunikasjon. Dessuten

er det slik at Morrisons klassifikasjonssystem ikke entydig identifiserer prosesser innenfor kartografisk kommunikasjon. For eksempel dersom vi har støyprosesser både på sendersiden og mottakersiden, vil disse prosessene tilsammen ha de samme injektive og surjektive egenskaper som prosess 16. Vi skal i et eget kapittel utdype delprosesser innenfor kartografisk generalisering.

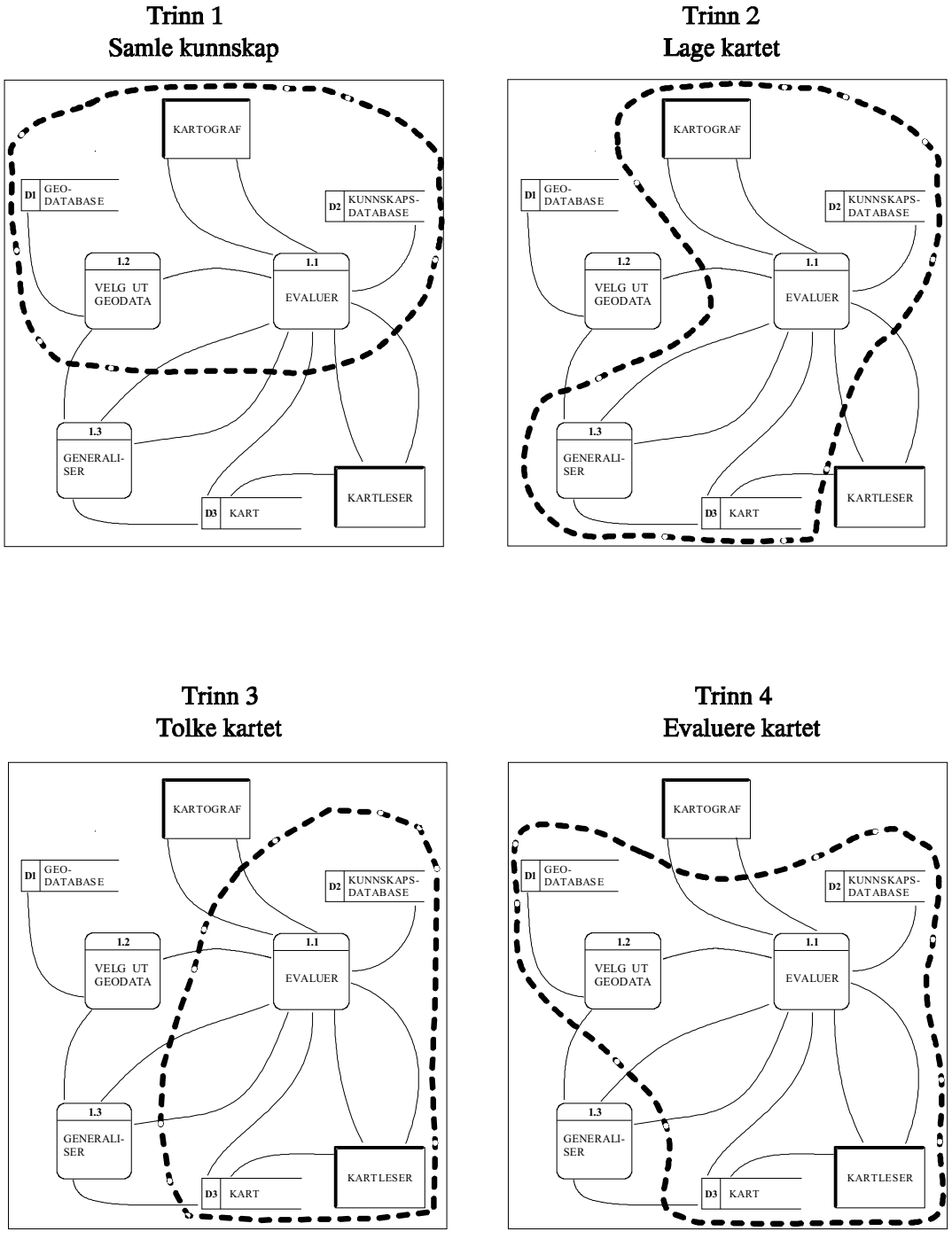


Figur 5.5: Kartografisk kommunikasjon modellert ved hjelp av dataflytdiagram. Etter Bjørke 1987.

5.5 Bjørkes modell

De modeller vi hittil har presentert, har ikke sagt noe om automatiserte prosesser i et kartografisk kommunikasjonssystem. Bjørke [Bjø87] presenterer en modell, som vist i figur 5.5, der dataflytdiagram er benyttet for å modellere kartografisk kommunikasjon. Modellen bygger i vesentlig grad på Morrisons modell i figur 5.4. Prosessene 1.2 og 1.3 har samme betydning som utvalg og generalisering i Morrisons modell. En stiplet linje viser grenselinjen mellom automatiserte og ikke automatiserte prosesser. Som vi ser går denne linjen gjennom samtlige prosesser og datalagre. Det betyr at modellen beskriver grader av automasjon fra det fullstendig automatiserte systemet til det helt manuelle systemet. I et automatisert system kan mottaker (kartleser) være en datamaskin. Kravene til generaliseringsprosessen blir selvsagt helt andre i et system der målet for generaliseringen er et datalesbart datalager enn om målet er et kart. Begrepet kartografisk generalisering får en utvidet betydning. Terminologien på området er noe uklar og det er ikke etablert en omforent terminologi. Begrepet modellgeneralisering er introdusert i litteraturen. Jeg vil anbefale en terminologi der vi begrenser bruken av kartografisk til kommunikasjonssystemer der informasjonen kodes i form av kartsymboler som presenteres på kart.

I figur 5.6 er illustrert fire hovedfaser i en kartografisk kommunikasjon: (1) samle kunnskap, (2) lage kartet, (3) tolke kartet og (4) evaluere kartet. Dataflytdiagrammet i figur 5.5 tjener som grunnlag for figur 5.6. Den uthevede stiplede linjen ringer inn de prosesser og datalagre som hører til de respektive faser. Fase (4) består i å sammenligne informasjonen fra kartet med virkeligheten. Den vanlige kartleser stoler nok vanligvis på kartet og foretar ikke systematiske evalueringer av kartet. Prosessen hører klart til katprodusentens ansvarsområde.



Figur 5.6: Ulike faser i en kartografisk kommunikasjon. Etter Bjørke 1987.

Kapittel 6

Informasjonsteori

I dette kapitlet vil vi presentere det matematiske grunnlaget for informasjosteorien etter Shannon og Weaver [SW64] og gi ideer om hvordan teorien kan anvendes på kartografiske problemstillinger.

6.1 Det matematiske grunnlaget for Shannon entropy

Det matematiske grunnlaget for Shannon entropi er beskrevet i flere lærebøker. Shannon og Weaver presenterer selv teorien sin i boken [SW64]. En utmerket presentasjon finner vi også hos Klir og Folger [KF88]. Med begrepet *entropi* (eng. entropy) vil vi i våre kartografiske anvendelser forstå variasjon. Entropibegrepet stammer fra termodynamikken, men begrepet har fått en selvstendig betydning innen informasjonsteorien. Variasjon kan måles på flere måter som for eksempel varians, konfidensintervall, maksimal feil osv.. Shannon entropi baserer seg på sannsynlighetsregning og anvender sannsynligheter (eventuelt sannsynlighetstetthetsfunksjoner) direkte i beregningene. Vi skal først presentere formelverket for Shannon entropi og senere diskutere formlenes kartografiske relevans.

La oss anta to mengder X og Y. For disse to mengdene kan vi definere følgende tre entropimål:

1. To *enkle entropier* basert på marginal sannsynlighetsfordeling:

$$H(X) = \sum_{x \in X} p(x) \log_2 \frac{1}{p(x)} = -\sum_{x \in X} p(x) \log_2 p(x), \tag{6.1}$$

$$H(Y) = \sum_{y \in Y} p(y) \log_2 \frac{1}{p(y)} = -\sum_{y \in Y} p(y) \log_2 p(y). \tag{6.2}$$

Dersom informasjonskilden er kontinuerlig, baseres entropiberegningene på X og Y sine sannsynlighetstetthetsfunksjoner.

$$H(X) = -\int_{-\infty}^{+\infty} p(x) \log_2 p(x) dx \tag{6.3}$$

$$H(Y) = -\int_{-\infty}^{+\infty} p(y) \log_2 p(y) dy \tag{6.4}$$

2. *Simultan entropi* (simultan = samtidig) defineres på grunnlag av sannsynligheten $p(x \cap y)$ for parvise hendelser definert over $X \times Y$,

$$H(X,Y) = -\sum_{(x,y)\in X\times Y} p(x,y) \log_2 p(x,y). \tag{6.5}$$

3. To *betingede entropier* definert som vektet middel av lokale betingede entropier:

$$H(X \mid Y) = -\sum_{y\in Y} p(y) \sum_{x\in X} p(x \mid y) \log_2 p(x \mid y), \tag{6.6}$$

$$H(Y \mid X) = -\sum_{x\in X} p(x) \sum_{y\in Y} p(y \mid x) \log_2 p(y \mid x). \tag{6.7}$$

Det består en del relasjoner mellom de tre entropiene som nylig er presentert. Disse relasjonene vil bli utdypet i de følgende teoremer.

Teorem 2 *Maksimal entropi etter likningen 6.1 eller 6.2 oppnås når alle forekomster er like sannsynlige:*

Bevis: Beviset består i å finne ekstremalpunktet for funksjonen

$$H = -\sum_{i=1}^{n} p(x_i) \ln p(x_i) + k(1 - \sum_{i=1}^{n} p(x_i))$$

hvor det første leddet i funksjonen er den enkle entropien uttrykt ved naturlige logaritmer og det andre leddet er en betingelse om at sannsynlighetene skal summeres til 1. Vi finner de partielle deriverte ved:

$$\frac{\partial H}{\partial p(x_i)} = -(1 \cdot \ln(p(x_i) + p(x_i)\frac{1}{p(x_i)}) - k = -\ln p(x_i) - 1 - k.$$

Ekstremalpunktet er gitt ved:

$$\frac{\partial H}{\partial p(x_i)} = 0 \Longrightarrow \ln p(x_i) = -1 - k$$

som igjen impliserer at ekstremalverdien oppnås når $p(x_1) = p(x_2) = \cdots p(x_n)$, hvilket skulle vises. □

Teorem 3

$$H(X,Y) = H(Y) + H(X \mid Y) = H(X) + H(Y \mid X) \tag{6.8}$$

Bevis: Beviset følger av relasjonen $p(x,y) = p(y)p(x \mid y) = p(x)p(y \mid x)$ og en utvikling av likningene 6.6 og 6.7. Dette gir

$$\begin{aligned} H(X \mid Y) &= -\sum_{y\in Y} p(y) \sum_{x\in X} p(x \mid y) \log_2 p(x \mid y) \\ &= -\sum_{y\in Y} p(y) \sum_{x\in X} \frac{p(x,y)}{p(y)} \log_2 \frac{p(x,y)}{p(y)} \\ &= -\sum_{y\in Y} \sum_{x\in X} p(x,y) \log_2 \frac{p(x,y)}{p(y)} \\ &= H(X,Y) + \sum_{y\in Y} \sum_{x\in X} p(x,y) \log_2 p(y) \\ &= H(X,Y) + \sum_{y\in Y} p(y) \log_2 p(y) \\ &= H(X,Y) - H(Y). \end{aligned}$$

Tilsvarende kan vi føre bevis for $H(Y \mid X)$. □

Teorem 3 lar seg generalisere til

$$\begin{aligned} &H(X_1, X_2, X_3, \ldots X_n) = \\ &\quad H(X_1) + H(X_2 \mid X_1) + H(X_3 \mid X_1, X_2) \\ &\quad + \cdots + H(X_n \mid X_1, X_2, \ldots, X_{n-1},). \end{aligned} \tag{6.9}$$

Teorem 4 *(Subadditiv egenskap)*

$$H(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i)$$

Bevis: Beviset finnes hos [KF88]. □

Egenskapen til Shannon entropi som følger av teorem 4, kalles entropiens *subadditive egenskap*. Likheten i teorem 4 holder hvis og bare hvis elementene fra de n mengdene er uavhengige i statistisk betydning. Denne antagelsen gir grunnlag for å formulere entropiens *additive egenskap* i det neste teoremet.

Teorem 5 *(Additiv egenskap)*
Dersom X og Y er statistisk uavhengige, er deres simultane entropi gitt ved :

$$H(X,Y) = H(X) + H(Y). \tag{6.10}$$

Bevis: To mengder X og Y defineres som statistisk uavhengige dersom $p(x, y) = p(x) \cdot p(y)$ for alle $x \in X$ og alle $y \in Y$. Dette innført i likning 6.5 for den simultane entropi gir:

$$\begin{aligned} H(X,Y) &= H(\, p(x_1)p(y_1), p(x_1)p(y_2), \ldots, p(x_1)p(y_s), \\ &\quad p(x_2)p(y_1), p(x_2)p(y_2), \ldots, p(x_2)p(y_s), \ldots \\ &\quad \ldots, p(x_n)p(y_1), p(x_n)p(y_2), \ldots, p(x_n)p(y_s)\,) \\ &= H(\, p(x_1), p(x_2), \ldots p(x_n)\,) + H(\, p(y_1), p(y_2), \ldots p(y_s)\,) \\ &= H(X) + H(Y), \end{aligned}$$

hvilket skulle vises. □

Anta at vi har en melding X som sendes over en kommunikasjonskanal som introduserer støy i meldingen. Vi kan i denne forbindelsen tenke på støy som det forhold at vi blander sammen karttegn. Det at vi for eksempel tolker en sirkel som en trekant, er et uttrykk for støy. Det at et symbol tolkes på ulike måter, er et uttrykk for variasjon og kan måles som entropi. På grunn av sammenblandingen må vi anta at den mottatte meldingen Y må være forksjellig fra X. *Informasjonstapet* som av Shannon og Weaver ble kalt *equivocation*, kan beregnes av formlene for betinget entropi. Den *nyttige informasjonen* (eng. useful information) R beregnes som en differans mellom entropien til informasjonskilden og usikkerheten i den meldigen som er sendt, gitt ved

$$R = H(X) - H(X \mid Y) \tag{6.11}$$

hvor $H(X)$ er entropien til informasjonskilden og $H(X \mid Y)$ er informasjonstapet uttrykt som betinget entropi. Nyttig informasjon kan også beregnes av differansen

$$R = H(Y) - H(Y \mid X) \tag{6.12}$$

hvor $H(Y)$ er entropien til de mottatte signalene og $H(Y \mid X)$ er et uttrykk for usikkerheten i de mottatte signalene. Vi kan følgelig velge om vi vil beregne nyttig informasjon enten på sendersiden eller på mottakersiden. Verdien vi beregner for R, blir lik i de to beregningstilfellene, noe det neste teoremet viser.

Teorem 6

$$R = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$$

Bevis:

$$\begin{aligned} H(X) - H(X \mid Y) &= H(X) \quad + \sum_{y \in Y} p(y) \sum_{x \in X} p(x \mid y) \log_2 p(x \mid y) \\ &= H(X) \quad + \sum_{y \in Y} p(y) \sum_{x \in X} \frac{p(x)p(y \mid x)}{p(y)} \log_2 \frac{p(x)p(y \mid x)}{p(y)} \end{aligned}$$

$$
\begin{aligned}
= H(X) \quad & + \sum_{y\in Y}\sum_{x\in X} p(x)p(y \mid x)\log_2 \frac{p(x)p(y \mid x)}{p(y)} \\
= H(X) \quad & + \sum_{y\in Y}\sum_{x\in X} p(x)p(y \mid x)\log_2 p(x) \\
& + \sum_{y\in Y}\sum_{x\in X} p(x)p(y \mid x)\log_2 p(y \mid x) \\
& - \sum_{y\in Y}\sum_{x\in X} p(x)p(y \mid x)\log_2 p(y) \\
= H(X) \quad & + \sum_{x\in X} p(x)\log_2 p(x) \\
& + \sum_{x\in X} p(x)\sum_{y\in Y} p(y \mid x)\log_2 p(y \mid x) \\
& - \sum_{y\in Y} p(y)\log_2 p(y) \\
= H(X) \quad & -H(X) - H(Y \mid X) + H(Y) \\
= H(Y) \quad & -H(Y \mid X),
\end{aligned}
$$

hvilket skulle vises. □

Korollar 1 *Nyttig informasjon kan også beregnes som summen av entropien til informasjonskilden og entropien til de mottatte signal minus den simultane entropien, gitt ved*

$$R = H(X) + H(Y) - H(X,Y). \tag{6.13}$$

Bevis: Av likning 6.8 følger at $H(Y \mid X) = H(X,Y) - H(X)$, som innsatt i likning 6.12 gir $R = H(Y) - [H(X,Y) - H(X)] = H(X) + H(Y) - H(X,Y)$, hvilket skulle vises. □

Kapasiteten C til en kommunikasjonskanal (et kart) defineres som

$$C = \max(R). \tag{6.14}$$

6.2 Eksempler på entropiberegninger

Vi vil her gi noen eksempler som har til hensikt å belyse formlene i Shannon's informasjonsteori. Eksemplene er enkle og trekker ikke inn komplekse kartografiske problemstillinger. I et senere kapittel vil den kartografiske anvendelsen av teorien bli utdypet. I entropiformlene benyttes toer-logaritmer. I prinsippet kunne naturlige logaritmer eller andre logaritmer vært valgt. Overgangen fra naturlige logaritmer til toer-logaritmer kan gjøres med utgangspunkt i $\log_2 x = \frac{\ln x}{ln2}$.

6.2.1 Enkel entropiberegning

Vi antar et værkart som viser skyforholdene ved hjelp av punktsymboler. Informasjonsvariablen kan anta verdier fra mengden $\{solskinn, lettskyet, delvisskyet, tettskydekke\}$. Vi teller opp antall punktsymboler på kartet og regner relative frekvenser for de ulike tilstandene. Vi antar følgende sannsynligheter: $p(x_1) = 0.1, p(x_2) = 0.3, p(x_3) = 0.4, p(x_4) = 0.2$. Entropien til kartet kan beregnes av den enkle entropiformelen i likning 6.1:

$$\begin{aligned} H(X) &= -\sum_{x \in X} p(x) \log_2 p(x) \\ &= -0.1 \log_2 0.1 - 0.3 \log_2 0.3 - 0.4 \log_2 0.4 - 0.2 \log_2 0.2 \\ &= 1.8464. \end{aligned}$$

6.2.2 Informasjonskilden består av flere informasjonsvariable

Gitt et punktsymbolkart der form og farge til symbolene varierer. Formvariasjonen (sirkel, trekant, firkant) beskriver typer av bensinstasjoner mens fargen beskriver nattåpen (rød) og dagåpen (blå). En opptelling innenfor kartet gir:

Type bensinstasjon	Nattåpen	Dagåpen
Esso	2	8
Texaco	2	3
Shell	1	2

Vi setter oss til oppgave å finne et uttrykk for entropien i kartet. I den foreliggende oppgaven har vi to informasjonskilder X og Y, representert ved de visuelle variable form og farge. Dette gir oss benevningen som vist i følgende tabell:

	y_1	y_2	$\sum$
x_1	2	8	10
x_2	2	3	5
x_3	1	2	3
$\sum$	5	13	18

Vi finner de marginale sannsynligheter:

$$p(x_1) = \frac{10}{18}, \quad p(x_2) = \frac{5}{18}, \quad p(x_3) = \frac{3}{18},$$

$$p(y_1) = \frac{5}{18}, \quad p(y_2) = \frac{13}{18}.$$

Vi antar at X og Y er statistisk uavhengige og benytter formelen for simultan entropi gitt i likning 6.10:

$$\begin{aligned} H(X,Y) &= H(X) + H(Y) \\ &= -10/18 \log_2 10/18 - 5/18 \log_2 5/18 - 3/18 \log_2 3/18 + \\ &\quad -5/18 \log_2 5/18 - 13/18 \log_2 13/18 \\ &= 1.4153 + 0.8524 = 2.2677. \end{aligned}$$

Dersom antagelsen om statistisk uavhengighet ikke er gyldig, er den beregnede entropien på 2.2677 for høy. Vi vil nå beregne kartet's entropi uten å gjøre forutsetningen om statistisk uavhengighet og velger å basere oss på formel 6.8:

$$H(X,Y) = H(X) + H(Y \mid X).$$

Vi må finne de betingede sannsynligheter $p(y \mid x)$ for å kunne beregne $H(Y \mid X)$.

$$p(y_1 \mid x_1) = \frac{2}{10} = 0.2, \qquad p(y_2 \mid x_1) = \frac{8}{10} = 0.8,$$

$$p(y_1 \mid x_2) = \frac{2}{5} = 0.4, \qquad p(y_2 \mid x_2) = \frac{3}{5} = 0.6,$$

$$p(y_1 \mid x_3) = \frac{1}{3} = 0.33, \qquad p(y_2 \mid x_3) = \frac{2}{3} = 0.67.$$

Ved å sette inn i likning 6.7 for den betingede entropi får vi:

$$\begin{aligned} H(Y \mid X) &= -\frac{10}{18}[0.2 \log_2 0.2 - 0.8 \log_2 0.8] - \frac{5}{18}[0.4 \log_2 0.4 - 0.6 \log_2 0.6] \\ &\quad -\frac{3}{18}[1/3 \log_2 1/3 - 2/3 \log_2 2/3] \\ &= 0.40107 + 0.26971 + 0.15305 \\ &= 0.8238. \end{aligned}$$

Dette gir oss den simultane entropien

$$H(X,Y) = 1.4153 + 0.8238 = 2.2391.$$

Vi kan også beregne den simultane entropien ved å anvende likning 6.5, men først må vi utlede sannsynlighetene $p(x \cap y)$.

$$p(x_1, y_1) = \frac{2}{18}, \qquad p(x_1, y_2) = \frac{8}{18},$$

$$p(x_2, y_1) = \frac{2}{18}, \qquad p(x_2, y_2) = \frac{3}{18},$$

$$p(x_3, y_1) = \frac{1}{18}, \qquad p(x_3, y_2) = \frac{2}{18},$$

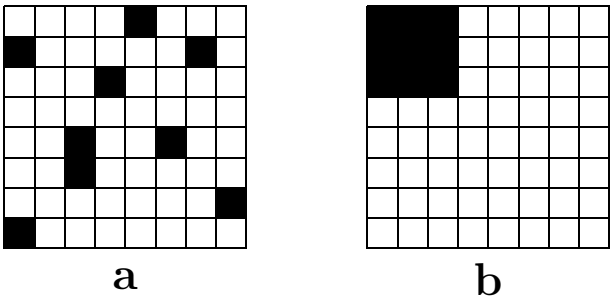
som insatt gir:

$$\begin{aligned} H(X,Y) &= -\sum_{(x,y)\in X\times Y} p(x,y)\log_2 p(x,y) \\ &= -2/18\log_2 2/18 - 8/18\log_2 8/18 - 2/18\log_2 2/18 - 3/18\log_2 3/18 \\ &\quad -1/18\log_2 1/18 - 2/18\log_2 2/18 \\ &= 2.2391. \end{aligned}$$

Hvilket stemmer overens med den foregående beregningen som ble basert på $H(X,Y) = H(X)+H(Y \mid X)$. Dersom vi sammenligner mot beregningen av $H(X,Y)$ basert på antagelsen om statistisk uavhengighet, ser vi at denne antagelsen ikke helt holder, fordi $2.2391 \neq 2.2677$. Dersom det er samvariasjon mellom X og Y, vil dette alltid føre til redusert simultan entropi, se teorem 4.

6.2.3 Romlig samvariasjon

Shannon entropi har ikke noe konsept for hvordan romlig samvariasjon skal modelleres. Det neste eksemplet belyser hvilke problem dette kan føre til. Figur 6.1 viser et eksempel på romlig samvariasjon. Selv om begge bildene i figuren hver består av



Figur 6.1: Forskjellig romlig fordeling av svarte og hvite bildelementer gir ulik grad av romlig samvariasjon. Bilde b) er lettere å huske enn bilde a) på grunn av sin store romlige samvariasjon.

9 svarte og 55 hvite bildeelementer, er mønstrene i de to bildene svært ulike. La oss først beregne den enkle entropien ved å telle antall svarte og hvite bildeelementer. Dette gir oss

$$H_a(X) = H_b(X) = -\frac{9}{64}\cdot\log_2\frac{9}{64} \;-\; \frac{55}{64}\cdot\log_2\frac{55}{64} \;=\; 0.586$$

hvor 9/64 og 55/64 er sannsynligheten for henholdsvis svarte og hvite bildeelementer. Denne modellen har gitt oss lik entropi i de to bildene, men resultatet er ikke i samsvar med vår oppfatning av bildene. Vår entropimodell må derfor på en eller annen måte tilføres informasjon om den romlige samvariasjonen. En teknikk vi

kan benytte, er å se på differanser i stedet for på absoluttverdier. I stedet for å velge bildeelementene som hendelser velger vi grenselinjen mellom to og to bildeelementer som hendelse. Vi tilordner så grenselinjene verdier fra mengden $\{+, -\}$, hvor $+$ betyr at de tilgrensende bildeelementene har like farger $(svart, svart)$ eller $(hvit, hvit)$. Tilsvarende betyr $-$ at bildeelementene har ulike farger $(svart, hvit)$ eller $(hvit, svart)$. Ut fra denne strategien beregner vi entropien i bildet ved

$$H(X) = -p^+ \cdot \log_2 p^+ - p^- \cdot \log_2 p^- \tag{6.15}$$

hvor p^+ og p^- korresponderer til vår modell om like/ulike farger til naboruter. En opptelling i bilde a) gir $p^- = 83/112$ og $p^+ = 29/112$, som gir entropien

$$H_a(X) = -83/112 \log_2 83/112 - 29/112 \log_2 29/112 = 0.825.$$

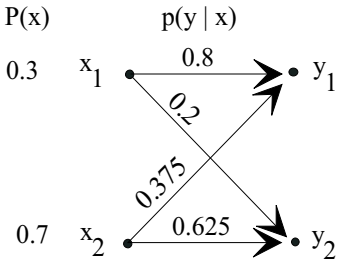
Tilsvarende for bilde b) får vi

$$H_b(X) = -106/112 \log_2 106/112 - 6/112 \log_2 6/112 = 0.301.$$

Etter denne beregningsmodellen har vi fått beregnet høyere entropi i bilde a) enn i bilde b). Den romlige samvariasjonen er introdusert i beregningene ved at vi som hendelser har benyttet differanser mellom to og to naboer. Modellen kan ytterligere utvides ved at vi trekker inn høyere ordens naboer. Videre kan vi skille mellom de fire tilstandene $\{ss, sh, hs, hh\}$, hvor s og h står for svart og hvit.

6.2.4 Informasjonstap

Vi vil nå demonstrere beregningen av informasjonstapet som følge av at symboler kan bli feiltolket, d.v.s. informasjonstap som følge av usikkerhet i den mottatte meldingen. Informasjonstapet vil bli beregnet på mottakersiden. Figur 6.2 illustrerer en situasjon.



Figur 6.2: $P(x)$ og $p(y \mid x)$ er henholdsvis sannsynligheten for symbol x og den betingede sannsynlighet for at symbol x tolkes som y.

Vi antar at vi har et kart med de to symbolene x_1 og x_2 og at deres sannsynligheter for forekomst i kartet er $p(x_1) = 0.3$ og $p(x_2) = 0.7$. Symbolenes tolkninger

benevnes y_1 og y_2 og deres sannsynligheter beregnes på grunnlag av de betingede sannsynligheter gitt i figur 6.2. Vi får

$$p(y_1) = 0.3 \cdot 0.8 + 0.7 \cdot 0.375 = 0.503 \quad \text{og} \quad p(y_2) = 0.3 \cdot 0.2 + 0.7 \cdot 0.625 = 0.497.$$

Som kontroll har vi at $p(y_1) + p(y_2) = 1.0$. Entropien i det tolkede kartet (den mottatte meldingen) beregnes fra likning 6.1

$$H(Y) = -0.503 \log_2 0.503 - 0.497 \log_2 0.497 = 0.99997.$$

Informasjonstapet (equivocation) $H(Y \mid X)$ som følge av usikkerhet i det tolkede kartet beregnes fra likning 6.7. For å gjøre likningen lettere å gjennomskue, vil summasjonene bli brutt ned i små steg. Informasjonstapet når symbol x_1 tolkes er

$$H(Y \mid x_1) = -0.8 \log_2 0.8 - 0.2 \log_2 0.2 = 0.72193.$$

Informasjonstapet når symbol x_2 tolkes er

$$H(Y \mid x_2) = -0.375 \log_2 0.375 - 0.625 \log_2 0.625 = 0.95443.$$

Det gjennomsnitlige informasjontapet beregnes ved å benytte sannsynlighetene for x_1 og x_2 som vekter. Dette gir oss

$$H(Y \mid X) = 0.3 \cdot 0.72193 + 0.7 \cdot 0.95443 = 0.88468$$

hvor $X = \{x_1, x_2\}$. Nyttig informasjon beregnes fra likning 6.12

$$R = H(Y) - H(Y \mid X) = 0.99997 - 0.88468 = 0.11529.$$

På grunn av feiltolkningene, må vi forvente at entropien til det tolkede kartet vil være forskjellig fra entropien til kildekartet, d.v.s. $H(X) \neq H(Y)$. Entropien til kildekartet beregnes fra likning 6.1 til

$$H(X) = -0.3 \log_2 0.3 - 0.7 \log_2 0.7 = 0.88129.$$

Tidligere har vi beregnet entropien $H(Y)$ til det tolkede kartet til 0.99997, altså er entropen på sendersiden forskjellig fra entropien på mottakersiden.

6.3 Informasjonskilder i kart

Vi vil nå gå over til å diskutere bruken av informasjonsteori på kartografiske problemstillinger. Framstillingen baserer seg i stor grad på [Bjø96]. Et første grunnleggende spørsmål vi vil stille oss er: "hvilke informasjonskilder eller hvilke klasser av variasjon finner vi i kart?La oss for eksempel vende tilbake til eksemplet i kapittel 6.2.2 om flere informasjonsvariable. Her beregnet vi entropien i kartet uten at vi innførte noe romlig konsept, men som vi illustrerte i eksempel 6.2.3 er romlig samvariasjon

et konsept av stor betydning for entropiberegninger. Dette viser at vi må addere et romlig konsept til Shannons informasjonsteori for at teorien skal bli operativ for kartografiske anvendelser.

Vi skal nærme oss det foreliggende problemet på en systematisk måte og skissere noen mulige løsninger. Det må imidlertid understrekes at vi har med en relativt ung teori å gjøre, og at det enda gjenstår forskning før vi kan si at vi har en beregningsmodell som løser de aktuelle kartografiske problemer på en fullt ut tilfredsstillende måte. Vår diskusjon vil begrenses til den syntaktiske komponenten til kart, altså kommunikasjonsproblemer på semantisk og pragmatisk nivå vil ikke bli diskutert.

Definisjon 8 *En kartentitet er et kartsymbol, en del av et kartsymbol, en gruppe av kartsymboler eller et virtuelt kartsymbol.*

I definisjon 8 trenger virtuelt kartsymbol en forklaring. I et senere eksempel om beregning av entropien til en linje, velges de såkalte δ-sirkler som kartentiteter. Siden δ-sirklene ikke finnes på kartet, men representerer en abstraksjon av en linje, er de et eksempel på konseptet virtuelt kartsymbol.

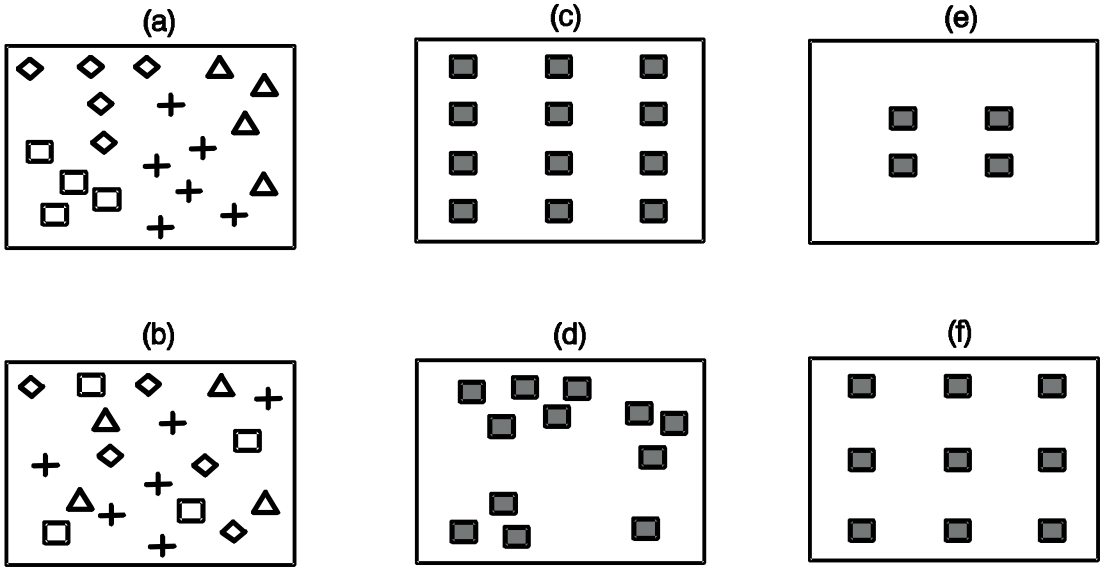
Definisjon 9 *En kartinformasjonskilde er et tuppel $(X, \mathcal{C})$ der X er en mengde av kartentiteter og C er en karakteristikk som er knyttet til kartentitetene.*

I definisjon 9 representer $\mathcal{C}$ den størrelse vi ønsker å beregne variasjonen for. La oss for eksempel anta et punktsymbolkart der symbolene kan anta ulike farger og ulike former. Mulige valg av $\mathcal{C}$ kan her være farge eller form. Dette vil i såfall gi oss de to informasjonskildene $(X, \mathcal{C}_1)$ og $(X, \mathcal{C}_2)$ der $\mathcal{C}$-ene representerer henholdsvis farge og form.

Bertin's klassifikasjon av visuelle variable [Ber81] kan benyttes som et utgangspunkt for klassifikasjon av kartinformasjonskilder. Bertin skiller mellom to hovedgrupper av variable: (1) symbolenes plassering i planet og (2) variable som benyttes til å utforme symbolene. For entropiberegninger er det nødvendig å splitte den første komponenten i tre delkomponenter:

1. metrisk komponent,
2. topologisk komponent,
3. antall symboler.

De tre konseptene er illustrert i figur 6.3. Bildene (a) og (b) tar sikte på å illustrere *topologisk entropi*. Ordningen i bildene (a) og (b) er klart forskjellig selv om antall symboler og de posisjoner symbolene okkuperer er lik i begge bildene. Det er naboskapene som er ulike. Bildene (c) og (d) tar sikte på å illustrere *metrisk* entropi. Variasjonen mellom bildene skyldes at avstanden mellom symbolene er ulik i de to bildene. Bildene (e) og (f) tar sikte på å illustrere *posisjonell* entropi. Variasjonen mellom bildene (e) og (f) skyldes at antall symboler er ulik i de to bildene.



Figur 6.3: Bildene (a) og (b) illustrerer topologisk entropi, bildene (c) og (d) illustrered metrisk entropi mens bildene (e) og (f) illustrerer enkel posisjonell entropi.

Definisjon 10 *Den topologiske entropien i et kart måler variasjonen i det topologiske arrangement av kartentitetene. Informasjonskilden for en topologisk entropi skrives* (X, Top) *der karakteristikken Top baserer seg på: (1) en definisjon av en antributt til kartentiteten som visuell variabel o.l. og (2) en definisjon av naboskap i kartplanet.*

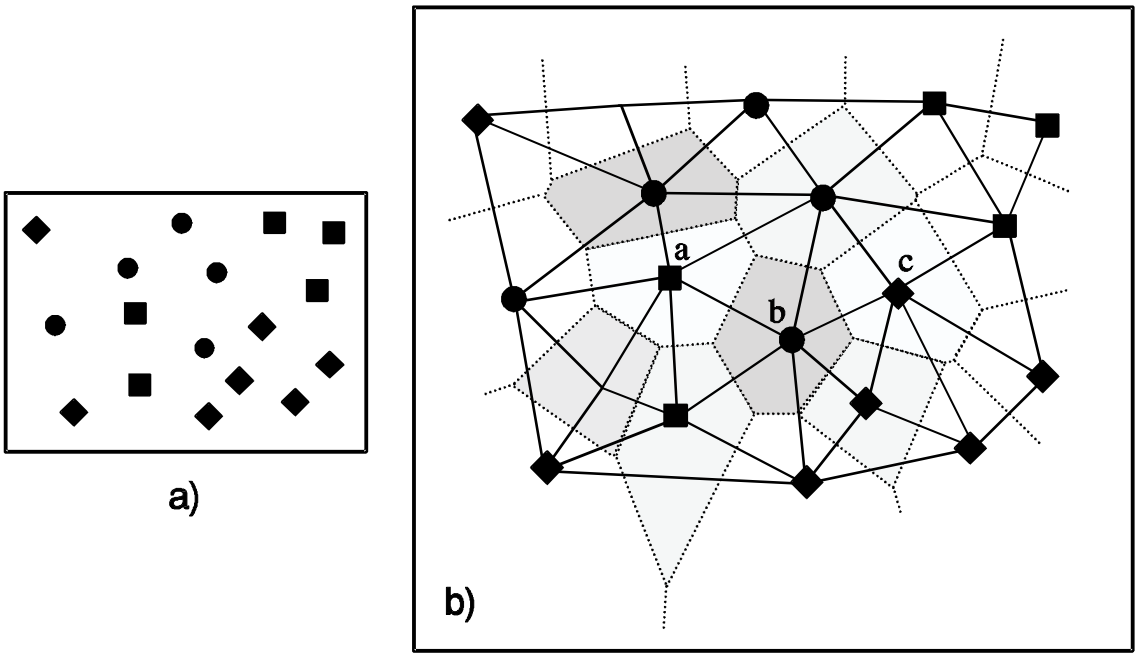
Definisjon 11 *Den metriske entropien i et kart måler variasjonen i avstand målt i kartplanet mellom kartentitetene. Informasjonskilden benevnes* (X, Met).

Definisjon 12 *Posisjonell entropi i et kart betrakter kartentitetenes posisjoner i planet som deres karakteristikk. Informasjonskilden benevnes* (X, Pos).

Siden vi i definisjon 12 definerer informasjonskilden's karakteristikk som planets (x, y)koordinater, fører det til at vi betrakter kartentiteter som unike hendelser dersom deres posisjoner i planet er forskjellige. I det spesielle tilfellet at alle kartentiteter er like sannsynlige og har ulike posisjoner, blir $H(X) = \log_2 n$; hvor n er antall kartentiteter.

Beregning av topologisk og metrisk entropi krever et romlig konsept. Vi skal belyse dette gjennom et eksempel. La oss ta utgangspunkt i punktsymbolkartet i ramme a) i figur 6.4 og stille oss til oppgave å definere hvilke symboler som er naboer til et gitt symbol. Her kan de såkalte Thiessen polygoner [LS80] komme til anvendelse. Vi vil kalle Thiessen polygonet til et punktsymbol for dets *visuelle område*. Resonnementet bak konseptet visuelt område er at når vi med øyet fokuserer på et lite område rundt et punktsymbol, er oppfattelsen lite forstyrret av nabosymboler.

Vi skal ikke gå nærmere inn på dette i denne forbindelsen, men nøye oss med å betrakte begrepet visuelt område som en definisjon. Siden en Delaunay triangulering står i et gjensidig avhengighetsforhold til Thiessen polygoner, vil vi basere nabodefinisjonen i et punktsymbolkart på Delaunay triangulering. Dette er illustrert i ramme b) i figure 6.4. La oss anta to noder i nettverket definert ved Delaunay trian-



Figur 6.4: Naboer i et punktsymbolkart defineres ved hjelp av en Delaunay triangulering. Et punkts Thiessen polygon kalles dets visuelle område.

guleringen. Vi definere graden av naboskap for de antatte nodene som antall kanter i den korteste stien mellom de to nodene. For eksempel er punkt (b) en 1. ordens nabo til punkt (a) mens punkt (c) er en 2. ordens nabo til punkt (a). Siden vi nå har en strategi for å definere naboer, kan vi anvende en differanseteknikk for å beregne den topologiske entropien i et punktsymbolkart. For eksempel vil vi i figur 6.4 få følgende mengde E av kartentiteter:

$$E = \left\{ \begin{array}{ccc} e_{\bigcirc\bigcirc} & e_{\bigcirc\diamond} & e_{\bigcirc\square} \\ e_{\diamond\bigcirc} & e_{\diamond\diamond} & e_{\diamond\square} \\ e_{\square\bigcirc} & e_{\square\diamond} & e_{\square\square} \end{array} \right\} \tag{6.16}$$

I det spesielle tilfellet at vi betrakter 0. ordens naboskap reduseres mengden av entiteter i likning 6.16 til:

$$E_0 = \{e_{\bigcirc\bigcirc}, e_{\diamond\diamond}, e_{\square\square}\} \tag{6.17}$$

Dette korresponderer forøvrig til valget at entiteter som benyttes av [Knö83]. Definisjonen av entiteter i likning 6.16 er mere fullstendig enn definisjonen av entiteter som

ble anvendt i kapittel 6.2.3 siden likning 6.16 definerer symmetrier som $(svart, hvit)$ og $(hvit, svart)$ som unike entiteter.

Den foreslåtte strategien for å beregne topologisk entropi, kan generaliseres til å gjelde alle nivåer av naboskap. En gjennomsnittsverdi kan beregnes for samtlige nivåer som et vektet middeltall. Problemet i denne forbindelsen vil i en gitt anvendelse være knyttet til valg av vekter. Vi vil senere presentere en teknikk som kalles seriation", og her vise betydningen av å trekke høyere ordens naboskap inn i entropiberegningen.

Den metriske entropien kan defineres etter en liknende måte som topologisk entropi. Vi velger i dette tilfellet avstanden mellom symbolene som informasjonskildens karakteristikk. Avstand kan i denne forbindelse knyttes til symbolenes tyngdepunkter. Entropien kan beregnes for 1. ordens naboer eller at beregningen utvides til å gjelde flere nivåer av naboer.

Før vi presenterer den neste definisjonen, vil vi vende tilbake til Bertin's visuelle variable. Hvorfor valgte Bertin det aktuelle settet som visuelle variable? Dette gir han ikke noen direkte begrunnelse for, men en måte å forklare valget på, er å si at de visuelle variable skal være ortogonale. Ortogonal i denne forbindelse betyr at vi kan endre en av de visuelle variable uten at det innvirker på de andre visuelle variable. Vi kan for eksemple endre fargetonen til et kartsymbol uten at det innvirker på symbolets form, retning eller størrelse. Med bakgrunn i nevnte forhold vil vi gjøre en generalisering og definere konseptet ortogonale informasjonskilder.

Definisjon 13 *Kartinformasjonskilder defineres som ortogonale dersom ingen av dem lar seg utlede ved å kombinere noen av de øvrige kartinformasjonskildene.*

Det må understrekes at ortogonal ikke må forveksles med statistisk uavhengighet. Ortogonal har med helt grunnleggende egenskaper til informasjonskilden å gjøre. Statistisk uavhengighet har med de aktuelle realisasjoner å gjøre.

6.4 Likhet og sannsynlighet for sammenblanding

Det er et veletablert prinsipp innen kartografien som sier at den "visuelle avstanden"mellom to kartsymboler skal være så stor at symbolene ikke blandes sammen. Visuell avstand kan dreie som to forhold: (1) avstand i kartplanet og (2) selve uformingen av kartsymbolet. Dersom for eksempel to linjer ligger nære hverandre i kartplanet, kan det være vanskelig å skille linjene fra hverandre og det vil som følge av dette være en viss sannsynlighet for sammenblanding av hele eller deler av linjene. Dersom for eksempel to linjer som representerer ulike informasjonsvariable har nær samme farge og utforming forøvrig, er også den visuelle avstanden liten selv om avstanden i kartplanet er stor.

Definisjon 14 *En funksjon $\mu(x, y)$ som definerer graden av visuell likhet mellom to kartentiteter x og y, vil bli kalt likhetsfunksjon. Likhet måles i intervallet* $[0, 1]$

av reelle tall. Dersom x og y er klart separerbare, er likheten 0. Desrom x og y er fullstendig umulig å adskille, settes likheten til 1.

Definisjonen av likhetsfunksjonen for et gitt kart er ikke noen triviell oppgave, fordi vi må ta hensyn til en rekke perseptuelle faktorer. Denne problemstillingen vil vi imidlertid la ligge i denne omgangen.

I entropiberegninger trenger vi å vite sannsynlighetene for sammenblanding. Det er uten videre klart at det må bestå en relasjon mellom grad av likhet og sannsynlighet for sammenblanding. For våre anvendelser kan vi basere oss på den følgende modellen. Anta en mengde X av kartentiteter og en mengde Y av tolkede kartentiteter. Kombinasjoner av elementer i X og Y defineres ved det kartesiske produktet $X \times Y$ der

$$X \times Y = \{(x, y) \mid x \in X \text{ og } y \in Y\}.$$

Kombinasjonen kan vi betrakte som en relasjon $R(X, Y)$ mellom X og Y. Vi tilordner en grad av likhet $\mu(x, y)$ til hvert par (x, y) i relasjonen. Siden en kartentitet skal være lik seg selv med maksimal grad av likhet, gis paret (x, y) likheten 1 i det tilfellet at $x = y$.

Fra relasjonen $R(X, Y)$ velger vi likhetsklassen av (x, y)-par gitt ved $x \times Y = \{(x, y) \mid y \in Y\}$. Vår definisjon av likhetsklasse svarer til begrepet likhetsklasse innen teorien om flytende mengder (eng. fuzzy set theory), jfr. [KF88] side 83. Overgangen fra likhet til sannsynlighet for sammenblanding kan gjøres med utgangspunkt i

$$p(y \mid x) = \frac{\mu(y \mid x)}{\sum_{y \in Y} \mu(y \mid x)} \qquad \text{for hver } y \in Y \tag{6.18}$$

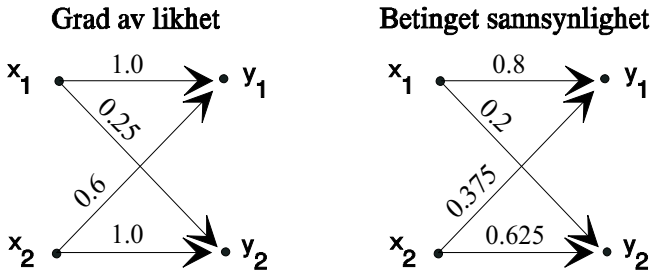
hvor $\mu(y \mid x)$ skal leses som "hvor lik x er y.Det er kun i spesielle tilfeller at $\mu(y \mid x) = \mu(x \mid y)$. Dette vil vi komme tilbake til senere i forbindelse med eksemplet i kapittel 6.7. Ved å anvende likning 6.18 på alle likhetsklasser til X, får vi

$$\sum_{y \in Y} p(y \mid x) = 1 \qquad \text{for hver } x \in X$$

som viser at den foreslåtte modellen tilfredsstiller sannsynlighetsregningens aksiom om at sannsynligheter skal summeres til 1. Figur 6.5 illustrere sammenhengen mellom likhet og sannsynlighet for sammenblandning. Ut fra figuren beregner vi de betingede sannsynligheter for sammenblanding til

$$p(y_1 \mid x_1) = \frac{1}{1 + 0.25} = 0.80 \qquad \text{and} \qquad p(y_2 \mid x_1) = \frac{0.25}{1 + 0.25} = 0.20$$

hvor vi ser at summen av de betingede sannsynligheter summeres til 1. En beregning etter likning 6.18 vil gi oss tidskompleksiteten $T = O(n^2)$ når likningen anvendes på alle likhetsklasser til X (n er antall elementer i X). For praktiske formål vil konflikter mellom nabosymboler begrenses til en liten omegn rundt x. Derfor kan vi



Figur 6.5: Sammenhengen mellom grad av likhet og betinget sannsynlighet for sammenblanding.

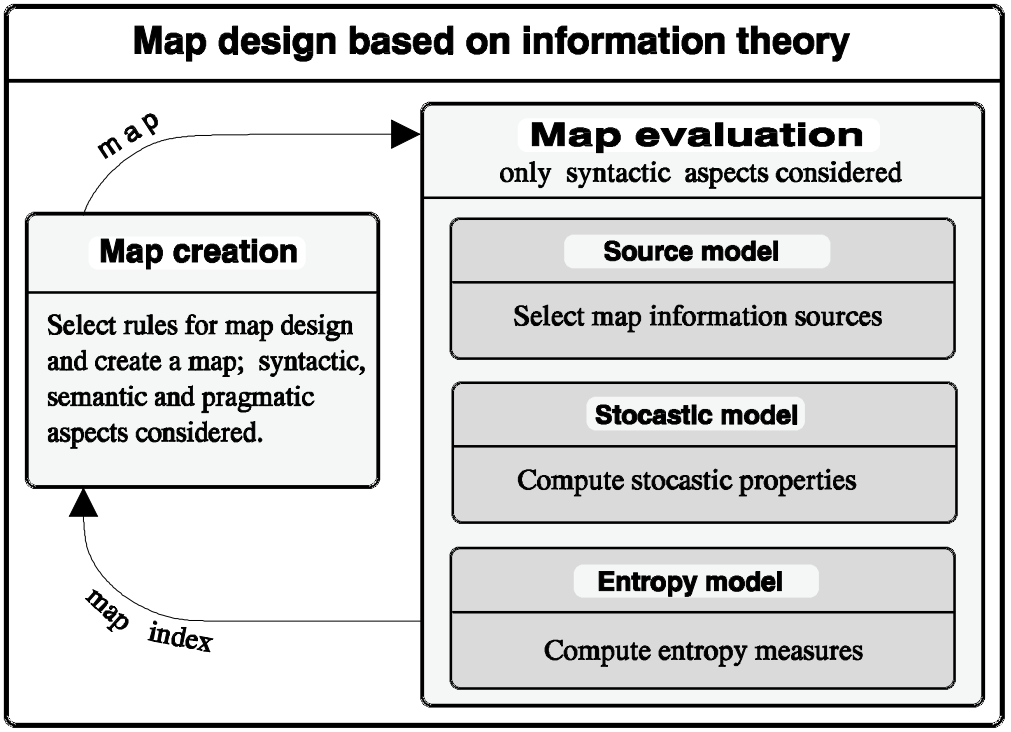
la løkken løpe over en liten omegn til x i stedet for over alle elementer til Y. Vi får da

$$Y_{s(x)} = \{y \in Y \mid y \text{ innenfor region } s(x)\} \tag{6.19}$$

hvor $s(x)$ er en omegn til x. Selv om tidskompleksisteten fortsatt er av 2. grad, kan vi som følge av søkeområdet's avgrensning håpe på at n blir liten.

6.5 Modell for kartdesign basert på informasjonsteori

Den modellen som her skal presenteres, baserer seg på et konsept av Bjørke [Bjø96]. Modellen, som er vist i figur 6.6, har to hovedprosesser: *kartgenerering* og *kartevaluering*. Basert på kartografisk kunnskap lager kartgeneratoren en strøm av kart. Kartevaluatoren vurderer kartene fortløpende og sender evalueringsindekser tilbake til kartgeneratoren. Indeksene utledes med utgangspunkt i Shannon informasjonsteori. Selv om i prinsippet alle kommunikasjonsnivåer kunne tenkes evaluert, begrenses likevel evalueringen til det syntaktiske nivået. Dette er fordi det pr. i dag ikke eksisterer noen operativ metodikk for å beregne entropier på det semantiske eller pragmatiske nivået i en kartografisk kommunikasjon.. Kartevalueringen dekomponeres til de tre operasjonelle områdene: (1) definere *kartinformasjonskilde*; (2) bestemme *stokastisk* modell; og (3) beregne *entropi* mål. Kartinformasjonskilden beskriver kartentitetene og deres karakteristikker. Den stokastiske modellen beskriver statistiske egenskaper til informasjonskilden som romlig samvariasjon, sannsynligheter for sammenblanding av kartentitetene etc.. Entropimodellen bestemmer hvilke entropimål som skal beregnes, som for eksempel R, $H(Y)$ og $H(Y \mid X)$. Et automatisert system basert på den foreliggende modellen, er en iterativ prosess. Etter hvert som kartgeneratoren får tilbakemeldinger fra kartevaluatoren, skal kartdesignet endres i retning av en optimal design. Dette vil senere bli belyst gjennom flere eksempler. Siden kartgeneratoren forutsettes å basere seg på kartografisk kunnskap, vil det totale systemet i vår modell ivareta alle de kartografiske vurderinger som er nødvendige for å få en effektiv kommunikasjon på alle de tre kommunikasjonsnivåene: syntak-



Figur 6.6: Modell for kartdesign etter Bjørke (1996)

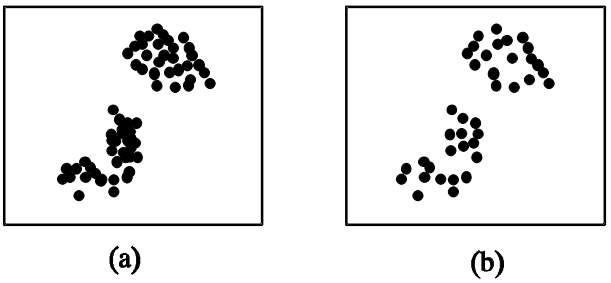
tisk, semantisk og pragmatisk. McMaster and Shea [MS92] beskriver en modell for kartgeneralisering der generaliseringen dekomponeres i tre operasjonelle områder: (1) hvorfor generalisere, (2) når generalisere, og (3) hvordan generalisere. Selv om informasjonsteori ikke kan gi direkte svar på hvordan generalisere, kan teorien likevel bidra til en forståelse for hvorfor generalisere. Anvendelsen av informasjonteori i vår modell dekker området når generalisere."

For å kunne definere modellene i kartevaluatoren, må det settes opp visse rammebetingelser for kartet. Disse rammebetingelsene vil vi kalle *designmål*. Et eksempel på designmål er at kartet skal inneholde så mange høydekurver som mulig ut fra visse perseptuelle kriterier. Det å sette opp designmålet vil vi oppfatte som en delprosess til kartgeneratoren. Etter at entropimålene er beregnet, må kartgeneratoren vurdere hva entropitallene betyr for kartdesignet og når en optimal løsning er funnet. Kartgeneratoren må derfor definere kriteriene for en optimal løsning. Disse kriteriene vil vi kalle *utvalgskriterier*.

6.6 Eksempel prikkekart

Prikkekart benyttes for å vise fordelingen av diskrete punktdata. De tradisjonelle designregler for prikkekart inkluderer: (1) valg av prikkestørrelse og (2) valg av antall enheter pr. prikk. Figur 6.7 viser to prikkekart og illustrerer betydningen av

designregel (2) siden kart (a) har en lavere enhetsverdi for prikkene enn hva tilfellet er i kart (b).



Figur 6.7: Prikkekart med forskjellig prikkeverdi

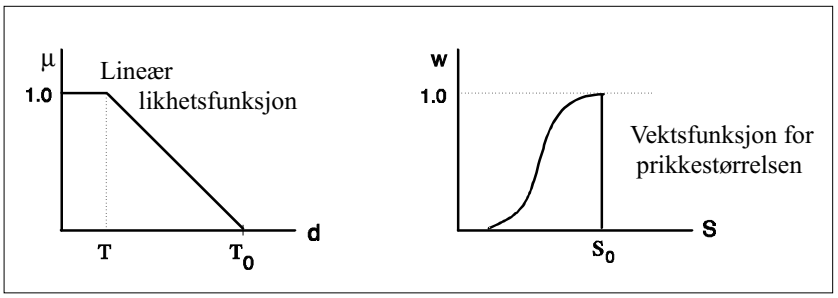
Designmål: (1) Gjør enhetsverdien q for prikkene så liten som mulig, d.v.s. lag så mange prikker som er mulig ut fra et perseptuelt synspunkt og (2) den best ønskelige prikkestørrelse er S_0.

Kildemodell: Velg informasjonskilden (X, Pos) og benytt prikkene som kartentiteter.

Stokastisk modell: Prikkene antas å ha lik sannsynlighet

$$p(x) = \frac{1}{N_x} \qquad \text{for hver } x \in X$$

hvor N_x er antall elementer i X. Sannsynligheten for sammenblandig utledes med utgangpunkt i likhetsfunksjonen i figur 6.8. I figuren er graden av likhet $\mu(y \mid x) = 1$



Figur 6.8: Likhetsfunksjon og vektsfunksjon for prikker i et prikkekart

dersom avstanden $d(x, y)$ mellom prikkene x og y er $\leq T$. Dersom $d(x, y) \geq T_0$, er prikkene lett å skille fra hverandre og $\mu(y \mid x) = 0$. Den betingede sannsynlighet $p(y \mid x)$ utledes med utgangspunkt i likning 6.18. Selve utformingen av likhetsfunksjonen krever at det tas omsyn til flere perseptuelle faktorer. Blant annet vil oppløsningen til kartet spille en rolle. Skal for eksempel kartet tegnes på en grafisk skjerm eller på et stykke papir.

Entropimodell: Beregn nyttig informasjon $R(X) = H(Y) - H(Y \mid X)$.

Utvalgskriterium: Krav (1) til designmålet modellers ved

$$\max[R(X) \mid M(q, S)]$$

Krav (2) til designmålet modelleres ved en vektet entropi gitt ved

$$K = w(S) \cdot R(X)$$

hvor $w(S)$ er en vektsfunksjon som tar prikkestørrelsen S som variabel. Figur 6.8 gir et eksempel på en slik vektsfunksjon. Vekten har her sin maksimumsverdi for den beste prikkestørrelsen S_0. Siden det ikke er noen god grunn til å akseptere prikker større enn S_0, er vektsfunksjonen utformet slik at $w(S) = 0$ når $S > S_0$. De verdier for (q, S) som tar hensyn til begge kravene som er stilt i designmålet, finnes ved

$$K_{max} = \max[w(S) \cdot R(X) \mid M(q, S)]$$

hvor kartgeneratoren $M(q, s)$ tar prikkenes enhetsverdi q og geometriske størrelse S som parametre.

6.7 Eksempel høydekurvekart

Et problem knyttet til høydekurvekart, eller isolinjekart generelt, er å fastlegge en gunstig verdi for den vertikale avstand mellom høydekurvene. Enkle formler basert på terrenghelning og minste akseptable horisontale avstand mellom høydekurvene benyttes i praksis, se for eksempel [Imh65] side 135. Vi vil imidlertid presentere en metode basert på informasjonsteori.

Designmål: Ut fra visse perseptuelle betingelser, gjør ekvidistansen så liten som mulig.

Kildemodell: Velg kartinformasjonskilden (X, Pos) der X tar de enkelte høydekurver som kartelementer.

Stokastisk modell: Det virker rimelig å benytte en modell der lange linjer tilordnes høyere sannsynlighet enn korte linjer (større sannsynlighet for å oppdage en lang linje enn en kort linje). Vi vil derfor benytte modellen

$$p(x) = \frac{l(x)}{\sum_{x \in X} l(x)} \qquad \text{for hver } x \in X$$

hvor $l(x)$ er lengden til høydekurve x og $\sum$ gir total lengde av samtlige høydekurver. Vi kan anta en tilsvarende likhetsfunksjon som vi benyttet for prikkekartet i kapittel 6.6. Med unntak av parallelle linjer, vil avstanden mellom høydekurvene variere. Dette gjør at bare deler av høydekurver kan være i konflikt med hverandre. Graden av likhet kan derfor beregnes som et vektet gjennomsnitt for de ulike seksjoner av kurvene.

Entropimodell: Beregn nyttig informasjon $R(X) = H(Y) - H(Y \mid X)$.

Utvalgskriterium: Den verdi for ekvidistansen som best tilfredsstiller designmålet, finnes ved

$$K_{max} = \max[R(X) \mid M(e)]$$

hvor kartgeneratoren $M(e)$ tar eksvidistansen e som parameter.

Et eksperiment [Bjø96] ble gjennomført med utgangspunkt i den presenterte modellen. Som terrengmodell ble benyttet et utsnitt av Kartverkets 100x100m-rutemodell over er kupert terrengområdet. Likhetsfunksjonen som ble benyttet, hadde følgende kritiske verdier: $T = 0.1$mm and $T_0 = 0.4$mm. Kartmålestokken var 1:120 000 og kartene ble tegnet ut med en linjetykkelse ca. 0.1mm. Resultatet av beregningen er vist i tabell 6.1. Av tabellen ser vi at kartets kanalkapasitet oppnås for ekvidistansen $e = 49$m. Den tilhørende R-verdi er 3.275. Vi må være klar

Tabell 6.1: R-verdier for ulike ekvidistanser. Kartmålestokken er 1:120 000.

ekvidistanse m	kartindeks $K = R(X)$	entropi $H(Y)$	grad av sammenblanding $H(Y \mid X)$
150	2.254	2.254	0.000
125	2.519	2.520	0.001
100	2.833	2.854	0.021
75	3.150	3.292	0.142
60	3.269	3.628	0.359
55	3.273	3.757	0.484
49	**3.275**	**3.918**	**0.643**
48	3.260	3.951	0.691
46	3.244	4.013	0.769

over at den valgte likhetsfunksjonen utgjør det springende punkt i turneringen"av sannsynligheten for sammenblandingen av kurver. En observasjon vi kan gjøre fra det foreliggende eksperimentet, er at optimal ekvidistanse oppnås når det er noe sammenblanding på kartet. Sammenblanding er selvsagt uønsket, men modellen vi benytter baserer seg på en optimalisering mellom antall kurver og graden av sammenblanding.

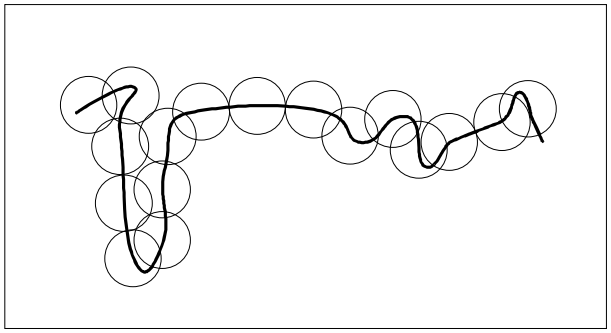
6.8 Eksempel linjegeneralisering

Vårt neste eksempel dreier seg om linjegeneralisering. Linjegeneralisering er i utgangspunktet en svært kompleks oppgave, fordi generaliseringen må vurdere den kontekst (sammenheng) linjen befinner seg i, se for eksempel [WM93]. Informasjonsteori vil i det kommende eksemplet kun bli anvendt til å vurdere forholdet mellom en linjes buktninger og graden av sammenblanding mellom buktningene. Et springende

punkt vil bli å finne en modell for å beregne entropien til en linje. En mulighet er å se på vinkelendringer langs linjen, men vi vil bruke en modell som baserer seg på en virtuell kartentitet, de såkalte $\delta-$sirkler. Figur 6.9 tar sikte på å illustrere en linjes $\delta-$sirkler. Sirklene er like store og plasseres langs linjen etter følgende regler: (1) sirkelsentrene ligger på linjen, (2) avstanden d mellom sirkelsentrene målt langs linjen er konstant og (3) diameteren δ til sirklene er lik d. Etter disse reglene vil δ-sirklene for en rett linje tangere hverandre. Det er først når linjen begynner å bukte seg at sirklene vil overlappe hverandre. Graden av overlapp kan utnyttes til å modellere graden av sammenblanding av linjens buktninger. Vårt konsept om $\delta-$sirkler har likhet til Perkal's ϵ-sirkler.

Designmål: Ut fra perseptuelle kriterier, behold så mye av variasjonen langs linjen som mulig.

Kildemodell: Velg kartinformasjonskilden (X, Pos) der X tar linjens $\delta-$sirkler som kartelementer, se figur 6.9.



Figur 6.9: δ-sirkler til en linje

Stokastisk modell:

$$p(x) = \frac{1}{N_x} \qquad \text{for hver } x \in X$$

hvor N_x er antall δ-sirkler til linjen. Graden av overlapp mellom δ-sirklene benyttes til å beregne sannsynligheten for sammenblanding. Det springende punkt blir å fastlegge δ-sirklenes størrelse.

Entropimodell: $R(X) = H(Y) - H(Y \mid X)$

Utvalgskriterium: Den generaliseringsparameter t (for eksempel t i D.P.-algoritmen) som tilfredsstiller designmålet, finnes ved

$$K_{max} = \max[R(X) \mid M(t)]$$

hvor kartgeneratoren $M(t)$ tar generaliseringsparameteren t som modellparameter.

Det foreligger foreløpig ingen forsøk basert på den presenterte modellen. Dette må gjøres før vi kan dra slutninger om hvor velegnet den er for praktiske anvendelser.

6.9 Eksempel koropletkart

Et sentralt designproblem for koropletkart er knyttet til: (1) valg av antall klasser og (2) valg av klassegrenser. Vi vil anvende informasjonsteori på den foreliggende problemstillingen.

Designmål: Ut fra perseptuelle betingelser, velg så mange klasser for koropletkartet som mulig.

Kildemodell: Velg kartinformasjonskilden (X,Top) der kartet tenkes inndelt i et raster og kartelementene er rutene i rasteret. Som karakteristikk velges differansen mellom rutene. Dette gir oss

$$X = \{x_{ij}\}$$

hvor x_{ij} er en rute i rasteret der rutens egen farge er i og fargen til naboruten er j. Dersom kartet kun har svarte og hvite ruter, får vi følgende kartentiteter: $\{x_{sh},\ x_{ss},\ x_{hs},\ x_{hh}\}$ hvor s og h refererer seg til svart og hvit.

Stokastisk modell:

$$p(x_{ij}) = \frac{N(i,j)}{\sum_{i \times j \in F^2} N(i,j)} \quad \text{for hver } (i,j) \in F^2$$

hvor $N(i,j)$ er antall naboruter med fargekombinasjonen (i,j) og F er mengden av gråtoner (farger) i kartet.

Entropimodell: Beregn nyttig informasjon $R(X) = H(Y) - H(Y \mid X)$.

Utvalgskriterium: Det antall gråtoner eller klasser som best tilfredsstiller designmålet, finnes ved

$$K_{max} = \max[R(X) \mid M(h)]$$

hvor kartgeneratoren $M(h)$ tar antall gråtoner h som parameter.

Et eksperiment [Bjø92] ble utført med utgangspunkt i den foreslåtte modellen. Her ble et tredvetalls personer bedt om evaluere forskjellen mellom gråtoner. Dette forsøket dannet så grunnlaget for å sette opp modellen for sannsynlighet for sammenblanding av gråtonenene. Deretter ble entropimål beregnet for to datasett, ett datsett med en sterk romlig samvariasjon og et annet med en tilfeldig romlig fordelig. Beregningen for de korresponderende koropletkartene er vist i tabell 6.2. Tabellen viser at det korrelerte kartet oppnår kanalkapasiteten ved 5 klasser, mens det random kartet oppnår kanalkapasiteten ved 4 klasser. Dette forsøket stemmer forbausende bra overens med de tommelfingerregler som benyttes for koropletkart. Her sies det at antall klasser helst ikke bør overstige 3-5 klasser, maksimum 7 klasser. Det forhold at det korrelerte kartet tåler større antall klasser enn det random kartet, stemmer også overens med hva vi kunne ventet oss.

Tabell 6.2: Entropimål for ulike antall klasser. Kanalkapasiteten er vist med uthevet skrift.

det korrelerte kartet			
ant. klasser	$H(Y)$	$H(Y \mid X)$	R
3	2.04	0.16	1.88
4	2.71	0.51	2.20
5	**3.32**	**0.98**	**2.34**
6	3.92	1.71	2.21

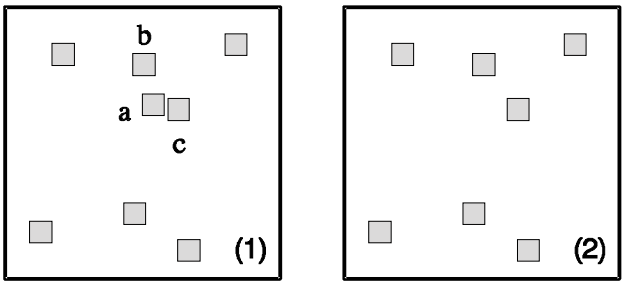
det random kartet			
Class no.	$H(Y)$	$H(Y \mid X)$	R
3	2.48	0.16	2.32
4	**3.24**	**0.63**	**2.61**
5	3.84	1.28	2.56
6	4.33	2.29	2.05

6.10 Eksempel eliminasjon av objekter

En metode for å løse generaliseringsproblemer er å eliminere objekter fra kartet. Kriteriene kan være (1) minimumsstørrelse eller (2) avstand til naboelementer, se for eksempel [RMM+95], side 466. La oss anta et kart med like store punktsymboler som vist i figur 6.10. For å løse de visuelle konfliktene, setter vi opp følgende modell.

Designmål: (1) Behold så mye av variasjonen som mulig og (2) fjern kartsymboler som ligger nær nabosymboler.

Kildemodell: Benytt kartinformasjonskilden (X, Pos) hvor elementene til X er kartsymbolene.



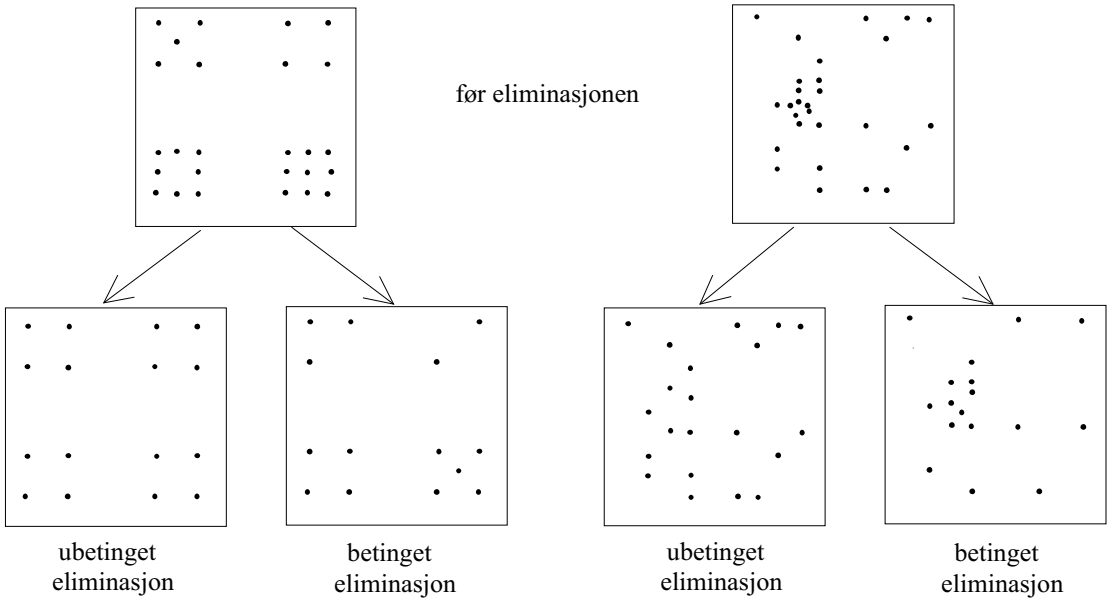
Figur 6.10: Forenkling ved eliminasjon av objekter. Symbol a er en kandidat for eliminasjon i kart (1). I kart (2) er a fjernet.

Stokastisk modell: Vi antar at symbolene er like store og benytter derfor den enkle modellen

$$p(x) = \frac{1}{N_x} \qquad \text{for hver } x \in X$$

hvor N_x er antall elementer. Beregningen av de betingede sannsynsynligheter for sammenblandig kan utledes på tilsvarende måte som for prikkekartet i kapittet 6.6

Entropimodell: Beregn lokal sammenblandig $H(Y \mid x)$ for elementet x og beregn nyttig informasjon for hele kartet ved $R(X) = H(Y) - H(Y \mid X)$.



Figur 6.11: Ubetinget utvalg kan gi feilaktig inntrykk at objektenes romlige fordeling

Utvalgskriterium: Krav (2) til designmålet tilfredsstilles ved å eliminere det elementet som har størst lokal sammenblanding. Krav (1) tilfredsstilles ved å maksimere $R(X)$. Det utvalg av kartsymboler som tilfredsstiller begge kravene, finnes ved

$$K_{max} = \max[R(X) \mid M(x)]$$

hvor kartgeneratoren $M(x)$ eliminerer fra kartet det symbolet x som har størst lokal sammenblanding. Størst lokal sammenblanding beregnes med utgangspunkt i

$$K_x = \max_{x \in X}[H(Y \mid x)].$$

Beregning etter den foreslåtte modellen vil bli illustrert ved et eksempel. La oss anta at kartsymbolene i figur 6.10 har følgende likheter: $\mu(a \mid a) = \mu(b \mid b) = \mu(c \mid c) = 1$, $\mu(a \mid b) = \mu(b \mid a) = 0.1$ og $\mu(a \mid c) = \mu(c \mid a) = 0.4$. Alle øvrige likheter antas å ha verdien 0. De tilhørende betingede sannsynligheter beregnes fra likning 6.18: $p(a \mid a) = 0.667$, $p(b \mid a) = 0.067$, $p(c \mid a) = 0.266$; $p(b \mid b) = 0.909$, $p(a \mid b) = 0.091$; $p(c \mid c) = 0.714$ og $p(a \mid c) = 0.286$.

Den lokale sammenblandingen for symbol a finnes så ved:

$$H(Y \mid a) = -0.667 \log_2 0.667 - 0.067 \log_2 0.067 - 0.266 \log_2 0.266 = 1.16.$$

Tilsvarende får vi for symbolene b og c: $H(Y \mid b) = 0.44$ og $H(Y \mid c) = 0.86$. Dette gir oss prioritetslisten (a, c, b), d.v.s. symbol a skal fjernes siden dette symbolet gir størst lokal sammenblanding.

Den foreslåtte prosedyren fører til et ubetinget utvalg. Dette har blant annet det problemet knyttet til seg at tynningen vil kunne gi inntrykk av en mere homogen

fordeling enn hva grunnlagsmaterialet skulle tilsi, se figur 6.11. Dette kan rettes opp ved å innføre betingelser slik at det relative forholdet mellom glisne og tette områder beholdes som illustrert i figur 6.11. Formuleringen av disse betingelsene har fortsatt flere uløste problemer.

6.11 Seriering (eng. seriation)

Vi skal her beskrive en metode der informasjonsteori kommer til anvendelse. Selv om metoden har en noe løs kobling til informasjonteori, har vi likevel valgt å beskrive metoden i et underkapittel til informasjonteori.

Klyngenalyse (eng. cluster analysis) anvendes innen geografisk informasjonsbehandling for å finne områder med like egenskaper. La oss anta at vi som geografisk enhet har valgt eiendommer og at vi har en rekke data om hus som befinner seg på eiendommene, som for eksempel antall etasjer, antall bad, antall soverom, antall stuer, grunnflatens størrelse, type takkledning, eiendomstomt eller bygslet tomt, osv.. Det å vise alle disse informasjonsvariable på ett enkelt kart, kan lett gi et kart med lavt persepsjonsnivå. Ofte er vi mere interessert i variasjoner enn absolutte verdier. Derfor kunne vi i vårt eksempel forsøke å finne ut om visse egenskaper (karakteristikker) har en tendens til å opptre sammen. Problemet kan angripes ved bruk av klyngeanalyse. En slik metode som kalles *seriation*, beskrives av Bertin [Ber81]. Vi vil oversette seriation (eng.) med seriering.

Teknikken går ut på at datamaterialet presenteres i en toveis tabell og at denne tabellen reorganiseres ved å bytte om henholdsvis rader og kolonner til den når maksimal ordning. Radene kan for eksempel representere de enkelte eiendommer mens kolonnene representerer de enkelte atributter (karakteristikker). La oss nå anta at atributtene har verdiområdet ja/nei. Det betyr at vi kan framstille tabellen som et binært bilde der rutene enten er svarte eller hvite. Det er uten videre klart at om vi bytter om rekkefølgen på radene, endres ikke meningsinnholdet i tabellen. Tilsvarende har vi for kolonnene. Seriering går ut på å reorganisere det binære bildet (tabellen) slik at rader (kolonner) som ligner hverandre, flyttes slik at de kommer nær hverandre i bildet (tabellen).

Figur 6.12 tar sikte på å illustrere teknikken med seriering. Først sorteres linjene i tabellen, deretter sorteres kolonnene. Sorteringen viser at vi har to hovedgrupper av geografiske områder: by og land. Landområdene har vi imidlertid valgt å dele i de to gruppene bygder og utkantstrøk. Dette gir oss tre klasser: bymessige områder, bygder og utkantstrøk. Disse tre klassene danner så grunnlag for det korresponderende kartet. I utgangspunktet hadde vi sju informasjonsvariable, men ved hjelp av seriering har vi klart å redusere antallet til bare tre generaliserte informasjonsvariable. Den serierte tabellen kan også danne grunnlag for å klassifisere informasjonsvariablene. Disse danner to grupper. Skole og kino danner den ene gruppen og forekommer i nesten samtlige områder. Disse informasjonsvariable kunne kanskje slås sammen i en gruppe vi kunne kalle primære tjenestetilbud"mens de øvrige kunne benevnes

sekundære tjenestetilbud."

Bertin utviklet i sin tid en serieringsmaskin. En operatør styrte hele prosessen ved å sørge for at bildet gikk mot stadig høyere grad av ordning. Teknikken ligger imidlertid til rette for å bli implementert i en datamaskin. Vi skal beskrive en algoritme som er utviklet av Bjørke og Smith [BS97] og belyse den gjennom et eksempel. Eksemplet baserer seg på en 144×53 datatabell over informasjon om boliger i 144 geografiske soner i byen Ibarra, Ecuador, Syd Amerika. Det er to hovedproblemer knyttet til seriering:

- finne en algoritme som kan seriere store tabeller med et akseptabelt tidsforbruk
- tolke det serierte bildet (tabellen).

Noe som kjennetegner seriering er:

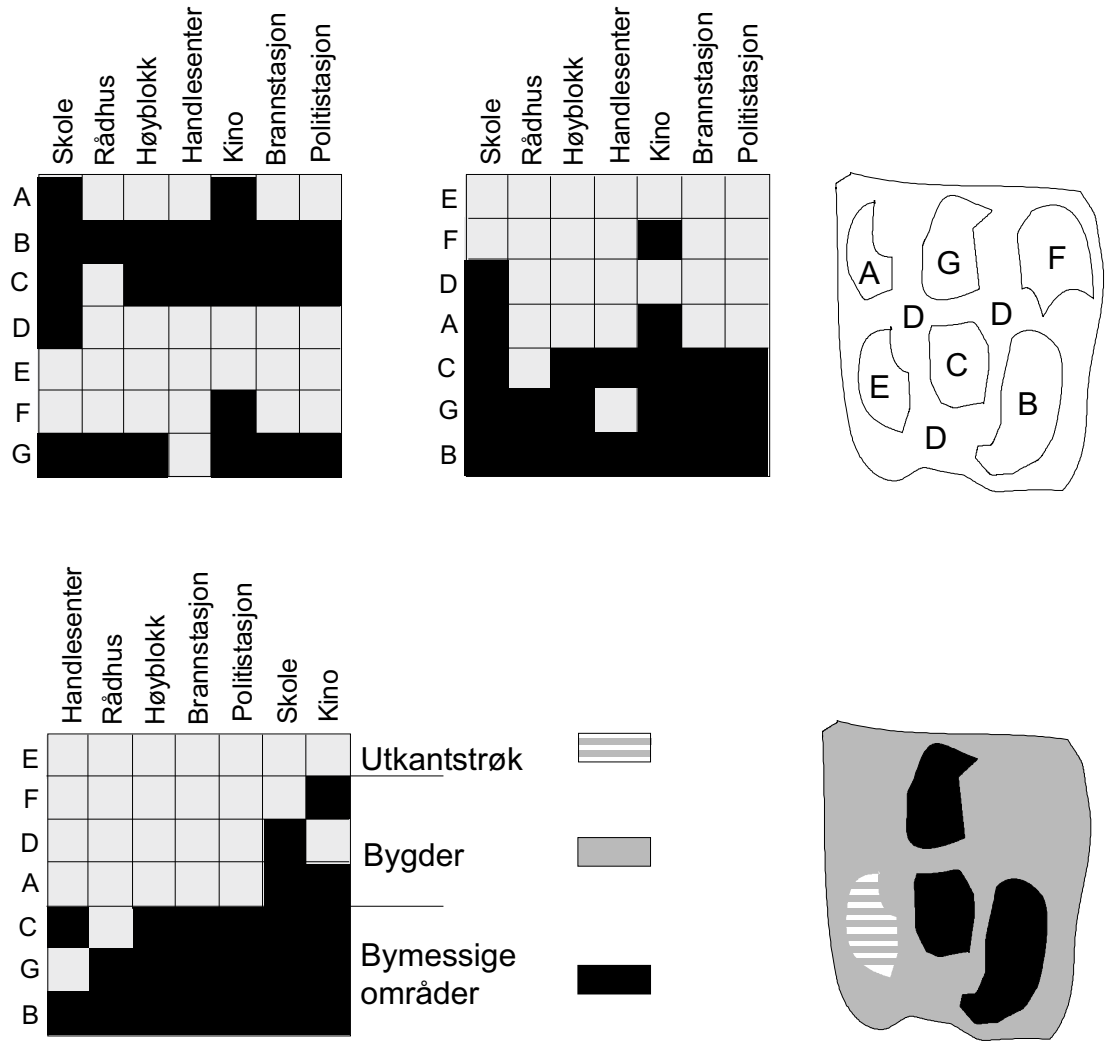
1. seriering filtrerer ikke ut informasjon,
2. seriering endrer ikke informasjonsinnholdet i dataene,
3. den presenterer dataenes struktur som et ordnet binærbilde,
4. tolkningen av dataene baserer seg på visuell evaluering av bildet.

6.11.1 Kriteriet for en seriert tabell

Selv om konseptet for seriering er svært enkelt, er det likevel vanskelig å finne en realiserbar algoritme, fordi problemet er i utgangspunktet et $O(n!)$ problem. Dersom vi har en rad med n elementer, finnes $(n-1)!$ unike rekkefølger for elementene. En $n \times m$ tabell vil derfor gi tidskompleksiteten $O(n! + m!)$ dersom alle mulige permutasjoner av såvel rader som kolonner skal evalueres.

La oss anta at vi har en prosess som bytter om rader (kolonner) i et binært bilde. For at denne prosessen skal kunne dra ombyttingen i en retning som stadig gir et mere seriert bilde, må prosessen ha et kriterium for hvor godt seriert et bilde er. Et slikt kriterium kan formuleres med utgangspunkt i den topologiske entropien til et binært bilde gitt i likning 6.15, $H(X) = -p^+ \cdot \log_2 p^+ - p^- \cdot \log_2 p^-$. Her vil $H(X) \rightarrow 0$ når $p^+ \rightarrow 1$ eventuelt når $p^- \rightarrow 1$. I det sistnevnte tilfellet vil $p^+ \rightarrow 0$ og rasterrutene vil danne en ordning som nærmer seg arrangementet på et sjakkbrett. Dette er selvsagt en løsning vi ikke ønsker for det serierte bildet. Derimot når $p^+ \rightarrow 1 \Longleftrightarrow p^- \rightarrow 0$, får vi et bilde som nærmer seg maksimal ordning ved at alle naboruter er av samme farge. Dette er den ordning vi er ute etter for det serierte bildet. Kriteriet for det serierte bildet må derfor inneholde en betingelse om

$$\max\left\{p^+\right\}.$$



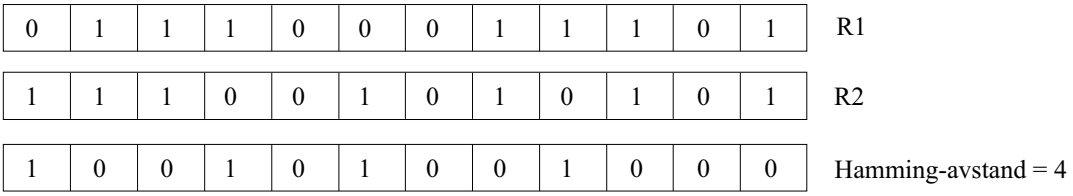
Figur 6.12: Eksempel på seriering

Dette svarer til å minimere Hamming-avstanden for det binære bildet. Hamming-avstanden mellom to linjevektorer $\mathbf{g}_i$ og $\mathbf{g}_{i+\Delta}$ eller to søylevektorer $\mathbf{g}_j$ og $\mathbf{g}_{j+\Delta}$ defineres ved

$$d_{i,i+\Delta} = \sum_{k=1}^{n} \mid g_{i,k} - g_{i+\Delta,k} \mid, \qquad d_{j,j+\Delta} = \sum_{k=1}^{m} \mid g_{j,k} - g_{j+\Delta,k} \mid \tag{6.20}$$

hvor n og m er lengden av henholdsvis linje- og søylevektorene og hvor $g_{i,k}$ og $g_{j,k}$ er vektorkomponentene. Figur 6.13 demonstrerer beregningen av Hamming-avstanden mellom to vektorer. Basert på likning 6.20 definerer vi første ordens Hamming-avstand for et $n \times m$ bilde ved

$$R_1 = \sum_{i=1}^{n-1} d_{i,i+1}, \qquad C_1 = \sum_{j=1}^{m-1} d_{j,j+1} \tag{6.21}$$



Figur 6.13: Beregning av Hamming-avstand. Hamming-avstanden mellom vektorene R_1 og R_2 er 4.

hvor R_1 og C_1 er summen for henholdsvis rader og kolonner og n og m er henholdsvis antall rader og kolonner. Som vi senere vil se, er det for enkelt kun å basere seg på første ordens naboskap i bildet. Høyere ordens naboskap må også evalueres av serieringsalgoritmen. Likning 6.21 lar seg generalisere til å definere k-te ordens Hamming-avstand ved

$$R_k = \sum_{i=1}^{n-k} d_{i,i+k}, \qquad C_k = \sum_{j=1}^{m-k} d_{j,j+k}. \tag{6.22}$$

Vi vil belyse betydningen av høyere ordens naboskap ved et eksempel. For dette eksemplet vil vi definere summen av Hamming-avstander over alle naboskap fra første til k-te orden, gitt ved:

$$\sum R_k = \sum_{i=1}^{k} R_i, \qquad \sum C_k = \sum_{i=1}^{k} C_i \tag{6.23}$$

hvor R_i og C_i finnes av likning 6.22. I tabell 6.3 er størrelsen $\sum R_k$ beregnet for $k = 1$ og $k = 36$ for bildene i figur 6.14. Fra tabellen ser vi at $\sum R_1$ til bildene (a) og (c) er henholdsvis 834 og 828, d.v.s. de numeriske verdiene er nesten like (forhold 0.98), men fra et visuelt synspunkt er det klart at bilde (c) er betraktelig mindre random enn bilde (a). Dette indikerer at første ordens Hammingavstand ikke gir et godt mål for den globale ordningen i et bilde. Derimot ved å se på $\sum R_{36}$ finner vi at verdiene avviker mere fra hverandre, 37054 og 31188 (forhold 0.84), noe som indikerer at høyere ordens naboskap må evalueres i forbindelse med serieringen. Dersom vi sammenligner bildene (b) og (c), får vi det interessante resultat at $\sum R_1$ sier at bilde (b) er mere ordnet enn bilde (c) mens $\sum R_{36}$ sier det motsatte. Vi forklarer dette også ved at bilde (c) har et globalt mønster som ikke fanges opp av beregningen basert på kun første ordens naboskap. I bilde (d) har vi det ferdig serierte bildet. Her ser vi at $\sum R_1$ faktisk er høyere enn for mellomstadiet i bilde (b). Derimot finner vi at $\sum R_{36}$ har sin minste verdi i bilde (d).

På bakgrunn av de resultater som hittil er presentert, definerer [BS97] kriteriet for det serierte bildet ved

$$D_{\min} = \min\left\{\sum_{k=1}^{n-1} w_r(k) \cdot R_k + \sum_{k=1}^{m-1} w_c(k) \cdot C_k\right\} \tag{6.24}$$

Tabell 6.3: Sammenligning av lokale og globale parametre for å måle ordningen i et bilde. $\sum R_1$ representerer en lokal parameter mens $\sum R_{36}$ representerer en global parameter.

	Bilder i figur 6.14			
	bilde (a)	bilde (b)	bilde (c)	bilde (d)
$\sum R_1$	834	316	828	348
$\sum R_{36}$	37054	35920	31188	29528

hvor R_k og C_k er definert i likning 6.22 og w er en vektsfunksjon. Det å designe vektsfunksjonen representerer et problem, fordi det er vanskelig å gi en generell regel for hvordan den skal designes. Basert på en antakelse om regulariteten i det serierte bildet, ble det for serieringen av bildet i figur 6.14 valgt den enkle regel at $w(k) = 0$ dersom $k > n/4$ ellers $k = 1$. Dette er grunnen til valget av $\sum R_{36}$ i tabell 6.3.

6.11.2 Algoritmen

Basert på kriteriet for et seriert bilde, vil vi beskrive en halvautomatisk algoritme. Brukerinteraksjonen gjør at vi oppnår en tidskompleksitet på $O(n^2)$. I forhold til problemets slemme $O(n!)$-natur, har vi med en reduksjon til $O(n^2)$ funnet en realiserbar løsning på problemet. Vår algoritme for å seriere et bilde består av to faser:

1. Kvikk-ordner (eng. Quick reorder),
2. Gruppe-ordner (eng. Group-reorder).

Algoritmen forutsetter at vi har med binære data å gjøre. Dersom dataene kan anta flere enn to verdier, er det mulig å takle dette ved å innføre flere kolonner i tabellen. For eksempel la oss anta at vi har tre kategorier for hustype. Vi kan da innføre én kolonne for hver kategori, altså vil vi i vårt eksempel innføre tre kolonner. Vi har ikke mistet noe informasjon ved denne transformasjonen.

Kvikk-ordner

Kvikk-ordner gjør den tilnærming at hva som er best lokalt også er best globalt. Anta at vi har et $n \times m$ binærbilde.

1. Flytt den mest hvite raden til toppen av bildet og den mest hvite kolonnen til venstre i bildet. Dersom alle radene (eller alle kolonnene) har likt antall svarte rasterruter, gjøres et tilfeldig valg blant radene (kolonnene). Tidskompleksiteten til dette steget er $O(n + m)$.
2. De resterende $(n - 1)$ radene og $(m - 1)$ kolonnene reorganiseres i henhold til en lokal evaluering som baseres på henholdsvis R_1 og C_1. Algoritmen har en

dobbel løkke for radene (kolonnene). Tidskompleksiteten er derfor $O(n^2+m^2)$. Først beregnes Hammingavstanden for rad 1 mot alle de andre $(n-1)$ radene. Herav velger vi raden med minst Hamming-avstand og lar denne raden bytte plass med rad 2. Prosedyren gjentas så for rad 2 i det vi lar den av de resterende $(n-2)$ radene som har kortest Hamming-avstand til rad 2, bytte plass med rad 3. Slik fortsetter prosedyren inntil den ytre løkken er inkrementert til rad $(n-1)$. Tilsvarende prosedyre benyttes for kolonnene.

Siden Kvikk-ordner kun tester på 1. ordens naboskap, vil det serierte bildet kunne få en syklisk tendens, altså like rader (kolonner) kan bli plassert i det serierte bildet med visse mellomrom. Denne effekten kan kompenseres ved å innføre en *antisykel-faktor* α. Denne faktoren gjør at alle rader som har har en Hammingavstand til rad i mindre enn α, flyttes til en sekvens av rader som starter med rad i. En tilsvarende prosedyre gjentas for kolonnene. Selv om antisykelfaktoren til en viss grad vurderer høyere ordens naboskap, er likevel ikke Kvikk-ordner i stand til å garantere at bildet blir tilstrekkelig seriert, noe som framkommer av bildene i figur 6.14. Bildene (b) og (c) viser effekten av antisykelfaktoren. Bilde (c) har et globalt mønster som viser fire grupper av rader, noe vi ikke oppdager i bilde (b). I bilde (b) er $\alpha = 0$ mens i bilde (c) er $\alpha = 12$. Valget av en hensiktsmessig verdi for α kan gjøres med utgangspunkt i en prøve- og feilemetode basert på visuell inspeksjon av bildet.

Gruppe-ordner

Det neste steget i vår algoritme er *Gruppe-ordner*.

1. Kjør Kvikk-ordner på bildet.

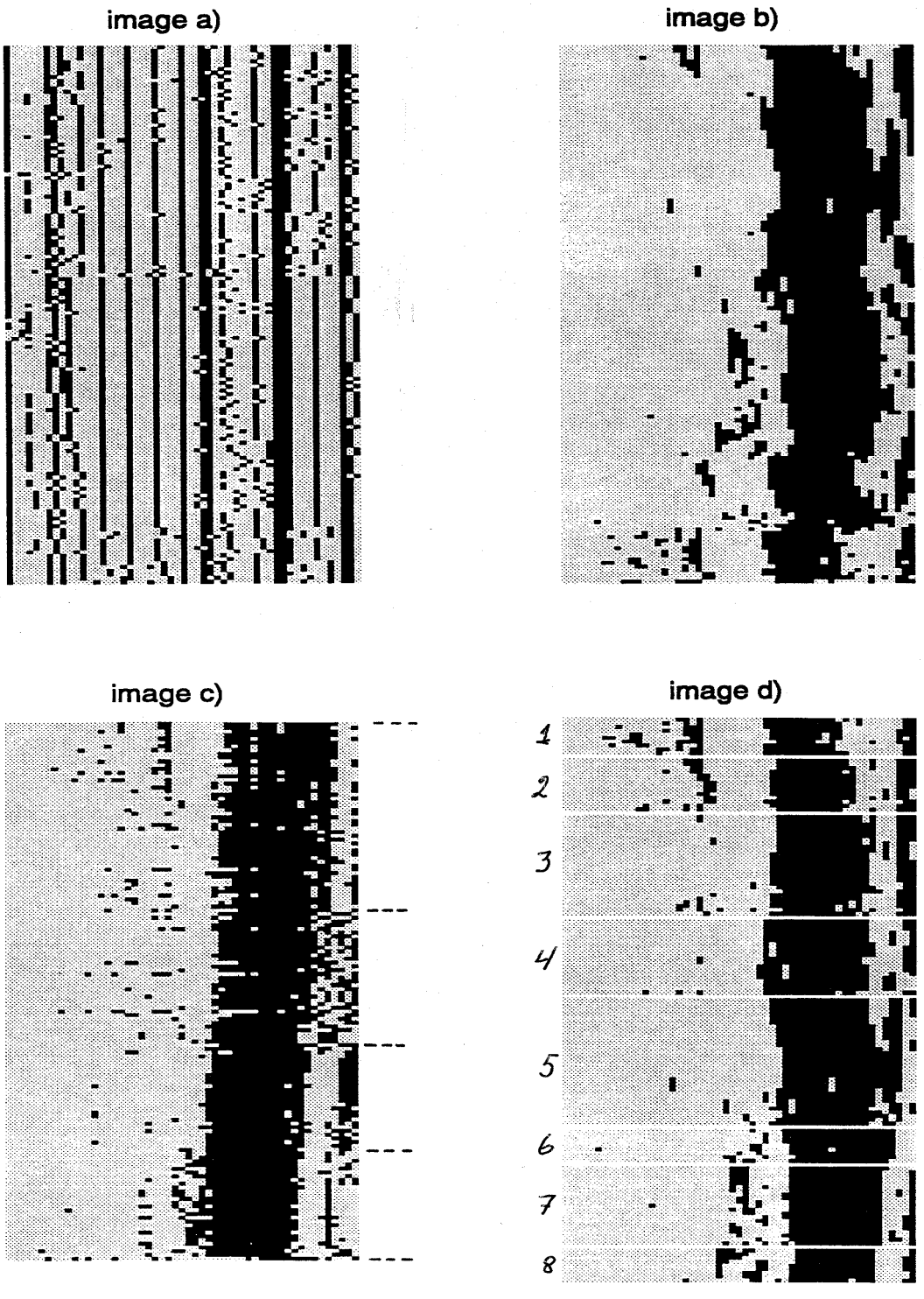
2. Med det Kvikk-serierte bildet på skjermen, skal operatøren peke på rader (kolonner) som synes å tilhøre en homogen gruppe av rader (kolonner), d.v.s. å oppsøke brudd i bildet. I bilde (c) i figur 6.14 markerer de prikkede linjene bruddsteder for radene.

3. Basert på gruppedefinisjonene beregnes en gjennomsnittsvektor for hver gruppe. Deretter flyttes rader (kolonner) til den gruppen de har størst tilhørighet til, basert på avstanden til den beregnede gjennomsnittsvektoren. Det kan innføres en maksimal avstand s. Dersom det for en gitt vektor ikke er mulig å finne en gruppe som gir tilhørighet bedre enn s, flyttes vedkommende vektor til en *støygruppe*.

4. Basert på gruppenes gjennomsnittvektorer, organiseres gruppene i en ordnet rekkefølge (tilsvarende prosedyre som for Kvikk-ordner).

5. Gjenta punktene 2-4 inntil gruppene er så homogene vi ønsker de skal bli.

Vi kan spørre om vi kunne hoppet over Kvikk-ordner og gått direkte på trinn 2. i Gruppe-ordner. Problemet med dette er at det på forhånd er vanskelig å vite



Figur 6.14: Eksempel på seriering av et bilde, etter [BS97]

Tabell 6.4: Tolkning av grupper av "Housing Variablesfor byen Ibarra i Ecuador, etter [BS97]. Se bilde (d) i den aktuelle figuren.

Group	**Group characteristics**
Group 1	1) lack of city services
Rural poor	2) some homes have earth floors and adobe walls
	3) all owners
Group 2	1) city services available
In town poor	2) no private bathrooms
	3) all one bedroom homes
	4) adobe walls
Group 3	1) no two room homes
In town small	2) predominantly one room homes
	3) city services available
	4) mostly owners
Group 4	1) nearly all have private bathroom and shower
In town better	2) city services available
	3) mostly owners
Group 5	1) only group with rooms for rent
Room Rent	2) mostly renters
	3) most homes have floorboards, rather than cement
Group 6	1) only group with some apartments
Apartment Rent	2) all renters
	3) all homes have private bathrooms and showers
	4) all cement and brick walls and cement roof
	5) some two and three bedrooms homes
	6) most homes have floorboards, rather than cement
Group 7	1) only group with some sectors with five rooms/home
Good Quality	2) more than half predominantly three bedroom
Big Home	3) all have private bathroom and shower
	4) nearly all are owners
Group 8	1) only group which has no one bedroom homes
Best Quality	2) all owners
	3) mostly two and three bedrooms

hvor mange grupper som finnes i bildet og hvordan vi skal finne et godt estimat for gjennomsnittsvektoren for en gruppe. Dersom vi ikke treffer gjennomsnittsvektoren ganske bra, kan vi risikere å danne grupper som består av vektorer som egentlig tilhører klart ulike grupper.

Den foreslåtte algoritmen har en viss likhet til K-means clusteringsom er beskrevet hos [AC90]. Det første steget i K-means clusteringer å dele dataene inn i K initielle klynger. Dette er det vi oppnår med vår Kvikk-ordner. Deretter utfører vår Gruppe-ordner de øvrige trinnene i en K-means clustering."

6.11.3 Tolkning av et seriert bilde

Den beskrevne metoden ble anvendt på data over eiendommer i byen Ibarra i Ecuador [BS97]. Dataene er fra 1990. Datatabellen har 144 rader og 53 kolonner der radene er geografiske soner og kolonnene er data som karakteriserer sonene. Ulike faser fram mot det serierte bildet er vist i figur 6.14. Her viser bilde (a) den initielle matrisen, bildene (b) og (c) to mellomsteg og bilde (d) den endelig serierte matrisen. I den serierte matrisen framkom åtte grupper av linjer. Det gjøres oppmerksom på at programmet som ble benyttet, var en prototype. Derfor ble ikke kolonnene fullstendig seriert, men dette innvirket ikke på tolkningen av de åtte gruppene av linjer i matrisen.

For å evaluere gruppene ble følgende spørsmål stilt:

1. Hvilke variable er representert på 100-prosentnivå for samtlige grupper (de fullstendig svarte kolonnene)?

2. Hvilke variable er ikke representert i noen av gruppene (de fullstendig hvite kolonnene)?

3. Hvilke variable er representert i bare én gruppe?

4. Hvilke variable mangler i en gruppe, men finnes i de øvrige gruppene?

Svarene på spørsmål en og to gir variable som er svake indikatorer for gruppene, mens svarene på spørsmåk tre og fire gir oss de sterke gruppe-indikatorene. Basert på en sosio-økonomisk tolkning av dataene, ble de åtte gruppene i bilde (d) gitt de tolkninger som er oppstilt i tabell 6.4.

Bibliografi

[AC90] A. A. Afifi and V. Clarck. *Computer-aided Multivariate Analysis*. Van Nostrand Reinhold, New York, 1990.

[AO93] R. W. Anson and F. J Ormeling. *Basic cartography for students and technicians*, volume 1. Elsevier Applied Science Publishers, London and New York, 2 edition, 1993.

[BA84] Axel Baudouin and Peder Anker. Kartpersepsjon og edb-assistert kartografi. Technical report, Norsk Regnesentral, Oslo, 1984. ISBN 82-539-0266-3.

[Ber81] Jacques Bertin. *Graphics and Graphic Information Processing*. Walter de Gruyter, Berlin, New York, 1981.

[Bjø87] Jan Terje Bjørke. *Cartographic Communication in Computer-based Environments*. PhD thesis, the Norwegian Institute of Technology, Department of Surveying and Mapping, N-7034 Trondheim, June 1987.

[Bjø92] Jan Terje Bjørke. Towards a formal basis for cartographic generalization. In JT Bjørke, editor, *Proceedings Neste generasjon GIS*, pages 101–109, 7034 Trondheim, Norway, December 1992. The Norwegian Institute of Technology, Department of Surveying and Mapping.

[Bjø96] Jan T. Bjørke. Framework for entropy-based map evaluation. *Cartography and Geographic Information Systems*, 23(2):78–95, 1996.

[BS97] Jan T. Bjørke and Betty Smith. Seriation: An implementation and case study. *Computers, Environments and Urban Systems*, ??(??):??, ?? 1997. Akseptert for publikasjon.

[CC80] G. J. Chamberlin and D. G. Chamberlin. *Colour: Its measurement, computation and application*. Heyden, London, 1980.

[FvDFH93] James D. Foley, Andries van Dam, Steven K. Feiner, and John F. Hughes. *Computer Graphics: Principles and Practice*. Addison-Wesley Publishing Company, Inc., New York, 2 edition, 1993.

[Imh65] Eduard Imhof. *Kartographische Geländedarstellung*. Walter de Greyter & Co., Berlin, 1965. Boken finnes i en engelsk oversettelse.

[Int73] International Cartographic Association, Wiesbaden/Germany. *Multilingual Dictionary of Technical Terms in Cartography*, 1973 edition, 1973.

[Kea82] J. S. Keates. *Understanding maps*. John Wiley & Sons, Inc., New York, 1982.

[KF88] George J. Klir and Tina A. Folger. *Fuzzy sets, uncertainty, and information*. Prentice Hall, New York, 1988.

[Knö83] Rudolf Knöpfli. Communication theory and generalization. In D.R.F. Taylor, editor, *Graphic Communication and Design in Contemporary Cartography*, pages 177–218. John Wiley & Sons Ltd., 1983.

[Kol69] A. Koláčny. Cartographic information - a fundamental concept and term in modern cartography. *The Cartographic Journal*, 6:47–49, 1969.

[LS80] D.T. Lee and B.J. Schachter. Two algorithms for constructing a delaunay triangulation. *International Journal of Computer and Information Sciences*, 9(3):219–242, 1980.

[Mac94] A. MachEachren. Time as a cartographic variable. In Hilary M. Hearnshaw and David J. Unwin, editors, *Visualization in Geographical Information Systems*, pages 115–130. John Wiley & Sons Ltd, Chichester, 1994.

[McL86] K. McLaren. *The colour science of dyes and pigments*. Adam Hilger, Bristol, 2 edition, 1986.

[Mor74] Joel L. Morrison. Cartographic generalization with emphasis on the process of symbolization. In *International Yearbook of Cartographhy*, volume 14, pages 115–127. Kirshbaum Verlag, 1974.

[Mor76] Joel L. Morrison. The science of cartography and its essential processes. In *International Yearbook of Cartographhy*, pages 84–97. Kirshbaum Verlag, 1976.

[Mor86] Joel L. Morrison. Cartography: A milestone and its future. In Michael Blakemore, editor, *Auto Carto London*, volume 1, pages 1–12, London, 1986. ICA ACI, Auto Carto London Ltd.

[MS92] Robert B. McMaster and Stuart K. Shea. *Generalization in Digital Cartography*. Association of American Geographers, 1710 16th Street NW, Washington D.C. 20009-3198, 1992.

[RMM+95] A. H. Robinson, J. L. Morrison, P. C. Muehrcke, A. J. Kimerling, and S. C. Guptill. *Elements of Cartography.* John Wiley & Sons, Inc., New York, sixth edition, 1995.

[RP76] Arthur H. Robinson and Barbara Bartz Petchenik. *The Nature of Maps, Essays toward Understanding Maps and Mapping.* The University of Chicago Press, 1976. ISBN 0-226-72281-3.

[SW64] Claude E. Shannon and Warren Weaver. *The Mathematical Theory of Communication.* The University of Illinois Press, Urbana, 1964.

[WM93] Ze-shen Wang and J.C. Muller. Complex coastline generalization. *Cartography and Geographic Information Systems*, 20(2):96–106, April 1993.