LUNDS UNIVERSITET
STATISTISKA INSTITUTIONEN

# Patienters livssituation och vårdkostnad

## En klusteranalys av vårdkostnadsdata

Jonas Schwanbom

Uppsats i statistik
10 poäng
Nivå 41 – 60 poäng

Handledare: Mats Hagnell

# 1. Summary

The purpose of this paper is to test a procedure that would be useful for a certain, more final purpose.

Assume that you want to know if the differences in the result of a certain effort in health care have any connection with differences between the life situations of the patients. This would be easier to investigate if the life situations of the patients had connection to the easily available variable Health Care Cost (in Swedish VK). In that case the searching for connections between VK of patient groups and the result of health care efforts would be sufficient to gain knowledge of possible connection between health care efforts and life situations. Thus that is the final purpose.

The purpose of this paper is to find a connection between combinations of variable values (in Swedish VVK) of life situations and mean and variance of the variable VK.

Mean and variance for VK is let become the variables of a two-dimensional space. In such a space every observation contains data for a group of patients. If the observations for patient groups with similar life situation are more often in the same areas than they are hazardously spread out over the space, there probably is a connection between life situation and mean and variance of VK.

The concrete purpose of this paper is to try to find groups of similar life situations, groups represented by dots mostly found within certain areas (of the space mentioned).

The data material for the paper has a somewhat odd nature that makes it a bit awkward to work with. On the other hand data are easily achievable, updated every year and chosen all from the beginning for connecting differences in VK as statistically well as possible to life situation variables.

The oddness of the data makes the understanding of the methods require a lot of text. This is why the methods are excluded from this summary.

The results seems to mean that the groups of similar life situations are hazardously spread out over the space. On the good side this seems to be able to blame on the part of the method that was not thought through thoroughly, i.e. it is not proved that the method could not work. On the other hand the reliability of the results seems to fade away in the iterations of the method before these same iterations have put much data into the results. More clearly spoken the method seems to forbid that the results contain both high reliability and a lot of information about each patient group.

# 2. Inledning och Syfte

Antag att man vill veta om resultatet av en viss vårdinsats hänger samman med patienternas livssituation. Detta vore lättare att undersöka ifall data om livssituationen gömde sig i den i vården lättåtkomliga variabeln vårdkostnad (VK). Då skulle det räcka att söka samband mellan (medelvärde och varians för) patientgruppers VK och vårdinsatsers resultat, för att få kännedom om eventuellt samband mellan vårdinsatsers resultat och livssituation. Samband mellan livssituation och vårdkostnad är vad denna uppsats syftar till att finna.

Medelvärde och varians för vårdkostnad kan låtas utgöra variabler i en tvådimensionell rymd. I en sådan rymd innehåller varje inprickad observation data för en grupp av patienter. Ifall observationerna för patientgrupper med likartad livssituation ligger inom samma områden oftare än vad de ligger slumpmässigt fördelade över hela rymden så illustrerar detta ett samband mellan livssituation och (medelvärde och varians för) VK.

Syftet med denna uppsats är just att hitta grupper med likartad livssituation, grupper som representeras av punkter som främst återfinns inom vissa områden (i den nämnda rymden).

Datamaterialet för denna uppsats har en natur som gör det litet besvärligt att jobba med. I gengäld är data lättillgängliga, uppdaterade varje år och redan från början valda för att knyta skillnader i VK tydligast möjligt till livssituationsvariabler (se 3. ”*Data*”).

# 3. Data

Data för uppsatsen beskriver medborgare (d.v.s. potentiella patienter) med hjälp av livssituationsvariabler eller egentligen med hjälp av kombinationer av variabelvärden. Data gäller Stockholms landsting 2004 och varje variabelvärdeskombination (VVK) gäller för en grupp medborgare. För varje grupp finns också medelvärde och varians för det gångna årets (2004) Vårdkostnad VK. Data är framtagen för indelning av Stockholms befolkning efter vårdbehov och livssituation, där vårdbehovet har mätts med variabeln VK. Syftet med indelningen är att hitta de livssituationsvariabler som, när befolkningen indelas efter dem, bäst förklarar geografiska skillnader i vårdbehov. Syftet med detta i sin tur är att korrekt kunna fördela vårdresurser, mätt i pengar, mellan olika geografiska vårddistrikt (Andersson P-A et al., 2000). Att geografiska skillnader vägts in när livssituationsvariablerna valts ut innebär att det rena sambandet mellan VK och livssituation möjligen skulle ha skapat ett annorlunda val av variabler.

Data består av medelvärde och varians för variabeln VK, för c:a 250 kombinationer VVK av värden för livssituationsvariabler.

Det första som görs med data är att skapa en kolumn med standardavvikelserna för att få tal av liknande storlek på de två variablernas värden. Därmed blir också tvådimensionella samband mellan observationerna tydligare.

Det andra som görs med data är att plocka ut 20% av VVK:na slumpmässigt. Dessa 20% utgör Testdata som används att utvärdera resultatet med. Resultatet som avses följer naturligtvis av att låta resterande data (framöver benämnda Data) genomgå proceduren som beskrivs i 4. ”*Metod*”.

**Självklart finns det större utrymme för varians i grupper med högt medelvärde, varför data fördelar sig mer kvastformigt i rymden ju mindre samband som finns mellan medelvärde och varians.**

Det tredje som görs med data är att indela Data i överlappande datagrupper. Varje datagrupps innehåll blir samtliga de VVK som delar en viss grupp variabelvärden. Datagrupper kommer att slås ihop enligt 4. ”*Metod*”, varför det kan vara intressant att veta vilka de ursprungliga datagrupperna är. Detta särskilt som de ”grupper” av variabelvärden som kännetecknar i de olika, ursprungliga datagrupperna inte är några grupper utan endast består och namnges av ett delat variabelvärde för varje datagrupp. Listan över de ursprungliga datagrupperna är alltså också listan över samtliga variabelvärden.

Tabell 1: Ursprungliga datagrupper = rena variabelvärden

| Variabel | Samboende | | Bostadsområde | |
|---|---|---|---|---|
| Variabelvärde | ENS | PAR | FÖR | STAD |
| Utläses | Ensam | Parboende | Förort | Stad |

| Variabel | Bostadstyp | | Utbildning | | Sysselsättning | |
|---|---|---|---|---|---|---|
| Variabelvärde | LGH | HUS | LÅG | HÖG | FP | ÖV |
| Utläses | Lägenhet | Småhus | Låg | Hög | Folkpensionär | Övriga |

| Ålder | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 år | 1 år | 5 år | 10 år | 15 år | 20 år | 30 år | 40 år | 55 år | 65 år | 70 år | 75 år | 80 år |

Här utläses ”x år” som ”Med ålder (i år) mellan x och siffran i nästa variabelvärde”

# 4. Metod

## 4.1 Sammanfattning av använd metod

Vad som sökes är kluster (se 5. ”*Teori*”) av VVK, kluster i vilka VVK:na för det första delar så många variabelvärden som möjligt och där VVK för det andra är så väl samlade som möjligt inom varje kluster.

Ifall man gör en klustring av samtliga VVK blir det problematiskt att hitta gemensamma variabelvärden mellan VVK:na i varje kluster. För att lättare hitta gemensamma variabelvärden mellan VVK:na i varje kluster har en särskild iterationsprocess skapats:

- Först klustras alla VVK datagrupp för datagrupp. I varje datagrupp delar alla VVK:n då ett visst variabelvärde. När dessa första klustringar är gjorda tas mätvärden fram för att bestämma vilka VVK:n som skall få fortsätta till nästa klustringsomgång. Kriteriet är enligt ovan att VVK är så väl samlade som möjligt inom varje kluster; mätvärdena är ämnade att mäta hur väl samlade VVK är inom varje kluster.
- Till andra klustringsomgången sätts godkända VVK:n ihop i nya datagrupper. För var och en av dessa nya datagrupper delar alla dess VVK:n två variabelvärden som inte delas av alla VVK:n i någon annan datagrupp. När klustringarna av tvåvariabelsgrupperna är gjorda tas åter mätvärden fram för att bestämma vilka VVK:n som skall få fortsätta till nästa klustringsomgång.
- I teorin fortsätter proceduren på analogt sätt tills optimalt antal delade variabelvärden per grupp nåtts, för visst mått på god samling av VVK:n inom varje kluster. I praktiken stannar proceduren tidigare, när antalet VVK:n per grupp blir för litet (se 4.4 ”*Kommentarer till använd metod*”).

Mätvärdet $\Sigma_{XS'}$ som avgör om gruppen skall väljas eller inte *utgörs av summan av elementen i matrisen* $XS'$ *för var och en av första stegets grupper*.

$S$ är matrisen för (1/avstånd) för de euklidiska avstånden mellan punkterna som representerar VVK i diagrammet. $S'$ är transponatet till $S$ och $X$ är matrisen som talar om ifall avståndet mellan två punkter ligger inom samma kluster eller ej, medelst en $1$:a eller en $0$:a.

**Anledningen till att X*(1/avstånd) valts istället för 1/(X*avstånd) är att därmed kommer antalet element per kluster att inverka på resultatet lika mycket som antalet element per datagrupp, vilket ger rättvisare jämförelse mellan datagrupperna.**

**Här används begreppet ”element” eftersom exempelvis begreppet ”observation” kan sägas vara en smula missvisande att använda när varje datavärde gäller en grupp av personer.**

**I varje datagrupp delas ett variabelvärde, i ett kluster delas tätheten i rymden som spänns upp av standardavvikelse och medelvärde för VVK:na.**

**Ju fler element i klustringen, desto större blir $\Sigma_{XS'}$. För visst antal element i klustringen sjunker antalet element per kluster när antalet kluster ökar. För viss täthet inom klustren så minskar medelavståndet inom klustret när antalet element minskar. Därför gäller, vid konstant, totalt elementantal och konstant genomsnittlig täthet inom klustren men minskat genomsnittligt antal element per kluster, att desto större blir $\Sigma_{XS'}$. Därmed kommer antalet element per kluster att inverka på resultatet lika mycket som antalet element per datagrupp inverkar, vilket ger rättvisare jämförelse mellan data-grupperna.**

## 4.2 Val av klustermetod

Krav på procedurens klustermetod/er/:

1) Metod att välja antal kluster med. Vald metod måste
   a) Använda avstånd, inte likhet
   b) Kunna utvärderas (subjektivt) för olika klusterantal
2) Metod att bilda kluster till proceduren med måste
   a) Använda avstånd, inte likhet
   b) Ge utskrift över vilken observation som förs till vilket kluster
3) Metod som ger utskrift över avstånden mellan samtliga par observationer behövs men kan vara identisk med någon av de båda första

De här kraven visade sig uppfyllas alldeles utmärkt av MINITAB:s båda förstahandsval. Single linkage clustering valdes för valen av antal kluster. K-mean-metod valdes att bilda kluster till proceduren med. När K-mean-metoden användes med lika många kluster som observationer gavs en utskrift av avstånden mellan varje par observationer.

## 4.3 Metoden steg för steg

1. Val av datagrupper. Lägg till klustringsomgångar tills gränsen nås; se textstycke 4.4 ”*Kommentarer till använd metod*”. Här antas endast två klustringsomgångar göras.
   1.1. Klustringsomgång 1
      1.1.1. För varje sådant variabelvärde: Klustra varje grupp VVK (datagrupp) med 1 gemensamt variabelvärde (exempelvis Bostadstyp: Småhus). Använd härvid single linkage clustering för att välja antal kluster och en K-mean method för den slutgiltiga klusterbildningen (se *Teori* för beskrivning av klustermetoderna).
      1.1.2. Beräkna summan av elementen i matrisen XS´ för varje av första stegets datagrupper.
      1.1.3. Välj de datagrupper som skall behållas genom subjektiv klusterbildning av summorna från 1.1.2. ; datagrupperna vars summor har lägst värden behålls. Med subjektiv klusterbildning menas personlig bedömning av vilka värden som bildar kluster med vilka.
      1.1.4. Test av val av datagrupper
   1.2. Klustringsomgång 2
      1.2.1. Klustra, för varje sådant par variabelvärden, varje grupp VVK med 2 gemensamma variabelvärden, dock endast medtagande de variabelvärden som kan kombineras och som godkänts så här långt i proceduren (Förtidspension kan exempelvis inte kombineras med ålder under 20 år).
   1.3. Utför steg motsvarande 1.1.2., 1.1.3. och 1.1.4. .
2. Val av kluster för vardera av de valda datagrupperna
   2.1. Test av pålitligheten hos valet av kluster för vardera av de valda datagrupperna
3. Framtagning av resultaten.
4. Test av resultatens användbarhet. Testa klustringsmetoden på orörda delen av datamaterialet: Kolla om de klasser som här ligger inom de nu valda klustren innehåller mer av de variabelvärden som hör till respektive valda kluster, än slumpen skulle ge, åtminstone för de eventuella kluster som hör till alla tre använda variabelvärdena.
   4.1. Kontroll av pålitligheten hos testen av resultatens användbarhet

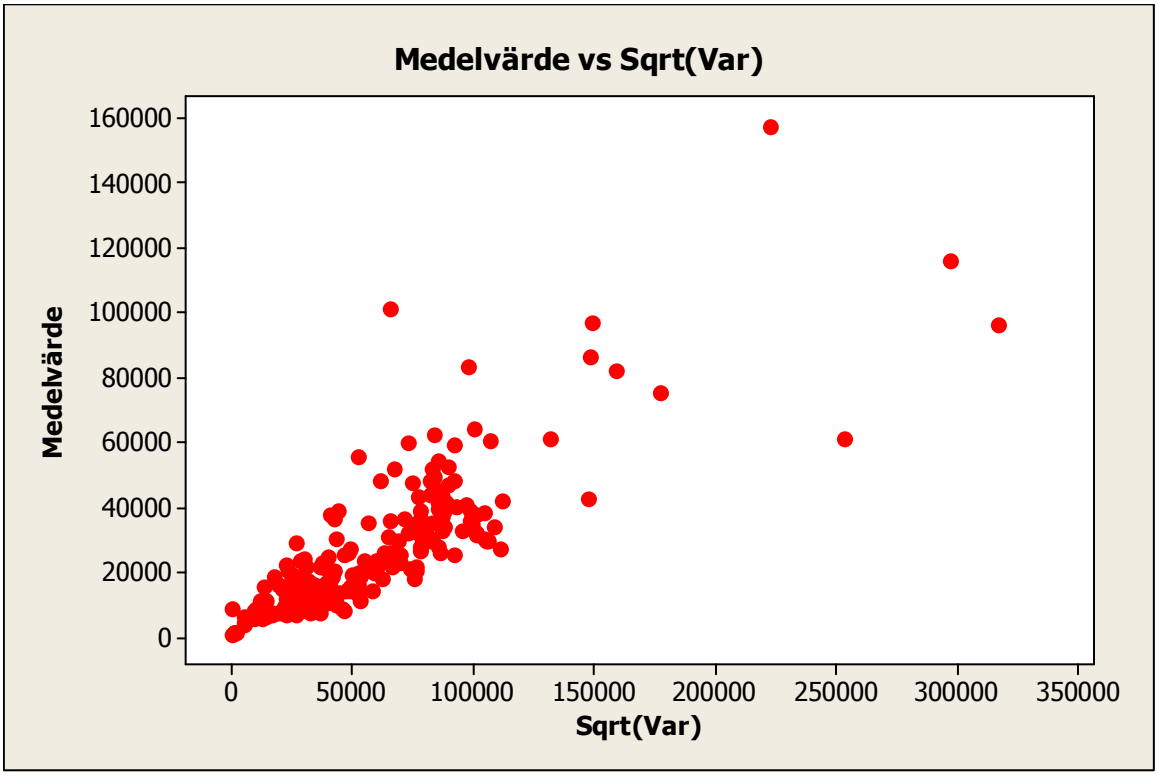
## 4.4 Kommentarer till använd metod

Helt enligt tidigare sökes kluster som för det första delar så många variabelvärden som möjligt och där VVK för det andra är så väl samlade som möjligt inom varje kluster. För att det skall kunna kontrolleras att VVK till en viss grad är väl samlade inom ett visst kluster så bör klustret ha hittats ur ett hyfsat stort antal VVK:n. Dessvärre minskar antalet VVK:n per grupp starkt vid varje utgallring av VVK:n. Kanske skall åter påpekas att utgallring av VVK:n görs för att endast de grupper där VVK:n delar så många variabelvärden som möjligt skall bidra med kluster till arbetets resultat. Meningen med att klustren skall dela så många variabelvärden som möjligt är, återigen, att när en grupp patienter, via medelvärde och varians för varibeln VK, knyts till ett visst kluster så skall detta kluster helst säga så mycket som möjligt om livssituationen för patienterna i fråga. Dessvärre stannar alltså proceduren i praktiken innan särskilt många variabelvärden hunnit komma att delas av klustren i varje grupp. I denna uppsats stannar proceduren redan före hopsättning av trevariabelsgrupper.

Trevligt nog är antalet potentiella patienter per grupp VVK:n större ju färre variabelvärden som delas av alla individer i gruppen. Det är trevligt eftersom det skulle öka sannolikheten för korrekt klassificering av varje samling patienter, efter VK d.v.s. efter livssituationsvariabler, om man slutligen kom att använda resultaten.

# 5. Teori

När man som i denna uppsats vill indela individer/observationer i grupper, utan att ha någon kunskap om efter vilka kriterier de borde indelas, tar man ofta till Klustermetoder. Lättast illustreras klustring när man som i denna uppsats kan pricka in observationerna i ett diagram med två axlar. Data visas inprickade på detta sätt i Figur 1.




Figur 1: Data
Varje punkt representerar en VVK. Sqrt(Var) betecknar standardavvikelsen. Variabeln är VK.

Den intuitiva tanken är att de observationer som är mer lika varandra än de är lika andra borde ligga närmre varandra än de ligger nära andra. Därför bildar man *kluster* av observationer som, allmänt uttryckt, ligger närmre varandra än de ligger nära andra. Oftast hittar man flera kluster i varje mängd data.
Det finns tyvärr många sätt för observationer att ligga närmre varandra på än vad de ligger nära andra. Till vilka kluster för man till exempel observationer som ligger glest, spridda emellan täta områden eller halvvägs in i hästskoformade, täta områden? Dessutom kan man göra exempelvis följande tankeexperiment:

> Fäst lysdioder på handleder och axlar hos en grupp personer, starta en dans, släck ljuset och ta en bild.

Det är tveksamt om det finns någon klusterdefinition som skulle tala om vilka dioder som hörde till vilken person.

> Häll ut lysdioder på ett sotat bord, släck ljuset och ta en bild.

Nu finns inga dioder som egentligen hör ihop. Hur ska man kunna skilja situationer i vilka det verkligen finns samband från den här ”uthällningssituationen”?
Det finns inga allmängiltiga teorier som besvarar sådana frågor.

Låt oss göra en sammanfattning som anknyter till de klustermetoder som använts denna gång. Det finns, sammanfattningsvis, två slags svårigheter i val av klustermetod. Det första slagets svårighet är definitionen av kluster. Ännu har det inte fastställts någon universell definition. Mest använda är klusterdefinitioner som bygger på euklidiska avstånd mellan observationer respektive på mått på likhet/olikhet mellan euklidiska avstånd mellan observationer.
Det andra slagets svårighet ligger i algoritmerna som behövs för att objektivt bestämma kluster. Det är svårt att hitta algoritmer som följer de klusterdefinitioner som algoritmmakaren har i tankarna och det är svårt att hitta rimligt snabba algoritmer som alltid hittar de kluster som bäst uppfyller den aktuella klusterdefinitionen.

Ett av många följdproblem är att *välja antal* kluster för klustringen man håller på med. Det nämndes ovan, just före exemplen med dioderna, att det kanske finns spridda observationer mellan tätare områden. Detta kan innebära att de spridda observationerna bildar en egen grupp ifall man väljer klusterantalet n+1 medan de utspridda fås att ingå i tätare områden ifall man väljer klusterantalet n.

För denna uppsats valdes metoden ”Single linkage clustering” att börja analysen med. Ur resultatet valdes antalet kluster subjektivt men med viss principfasthet. För den fortsatta analysen användes sedan en ”K-mean-metod” med det antal kluster som den första metoden hjälpt till att välja. Låt oss ta en snabb titt på dessa metoder, innan vi studerar hur de använts i uppsatsen.

## 5.1 Single linkage clustering

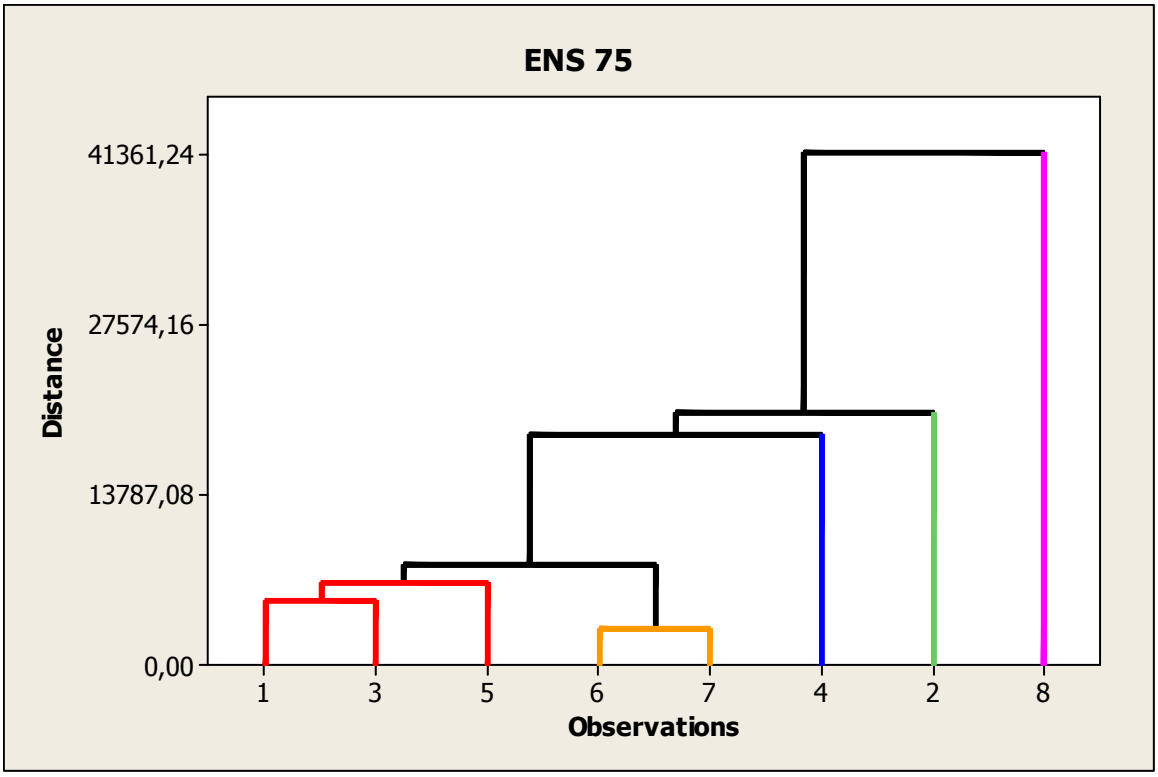
Hierarkiska klustermetoder kan inledas med att låta alla observationer tillhöra samma kluster som kan delas upp i underkluster som i sin tur kan delas upp o.s.v. Alternativt kan man börja med att varje observation bildar ett eget kluster, som enligt vissa regler kan sammanslås med andra, till överkluster, som i sin tur kan sammanslås o.s.v. . I Single linkage clustering klustrar man observationer en efter en, i ordning efter avståndet till närmsta observation. Det finns alternativ, några ges i 5.2 ”*K-mean-metoder*”.

Ordet ”kan” är inskjutet i denna genomgång för att understryka att varje kluster *kan* men *ska* inte nödvändigtvis delas upp eller sammanslås, i varje steg av klustringen. Det är naturligtvis reglerna och data som avgör om ett kluster ska påverkas i ett visst steg eller inte.

Man kan rita träd över vilka kluster som hör till vilka över-/underkluster, när man använder hierarkiska metoder. Ifall man använder Single linkage klustering ser trädet likadant ut hur många eller få kluster man än väljer. Endast de färger eller noteringar som visar vilka kvistar, grenar, stammar etc. som tillhör samma kluster, skiftar med hur många kluster man valt. Därför kan man lätt jämföra olika val av antal kluster med varandra, mer eller mindre subjektivt.

Mer information hittas i exempelvis Cluster analysis av Brian Everitt.

Figur 2 är ett exempel på ett träddiagram från användning av Single linkage clustering.



Figur 2

Principerna som subjektivt använts i denna uppsats, för att välja klusterantal ur träddiagrammen från Single linkage clustering, är

- Välj så många kluster som möjligt/rimligt
- Välj så få kluster att avstånden mellan klustren blir maximala
- Välj kluster med så många observationer som möjligt i varje

Som antyds i exemplet på träddiagram fås ofta väldigt få observationer i klustren med störst avstånd mellan sig. I den procedur som använts (se 4. *”Metod”*) riskerar kluster med få observationer att rensas bort. Därmed försvinner en del av underlaget för procedurens resultat men resultatet störs inte i övrigt, därför är maximering av antalet observationer den minst viktiga principen.

## 5.2 K-mean-metoder

Metoder som utgår från en uppdelning i visst antal kluster och sedan låter observationerna byta klustertillhörighet tills uppdelningen uppfyller något kriterium kallas iterativa metoder. K-mean-metoder tillhör denna grupp. Kriteriet kan exempelvis vara att varje observation skall tillhöra det kluster vars centroid den ligger närmast (i motsats till att klustra efter vilken *observation* en viss observation ligger närmast, som i exempelvis den hierarkiska Single linkage clustering). Det finns även andra kriterier: Kriteriet avgör bl.a. vilken form som klustren kan få (Gordon, 1999).

I denna uppsats har K-mean-metod använts m.h.a. MINTAB. Obekant av programanvändaren är bl.a. vilket kriterium och hurdan första klustring som väljs när man använder K-mean-metod m.h.a. MINITAB. Däremot väljer man aktivt hur många kluster som skall letas upp. För att välja antal kluster har som sagt Single linkage clustering använts.

# 6. Resultat

Resultaten kan presenteras först efter ett antal tester, kontroller och förberedelser.

## 6.1 Test av val av grupper och Kontroll av resultatens användbarhet

Utöver att det inte får finnas några teoretiska brister i procedurens val av grupper och kluster så finns två krav att ställa på resultaten, för att de skall vara användbara.

- Det måste finnas ovanligt många av gruppens observationer i varje klusters område i /diagrammet över/ Testdata, för varje vald grupps alla valda kluster
- Det får inte finnas för stor andel observationer som inte tillhör gruppen, i varje klusters område i /diagrammet över/ Testdata, för något valt kluster i någon vald grupp

Först och främst måste kontrolleras att procedurens val av grupper inte, eller knappast, påverkats av

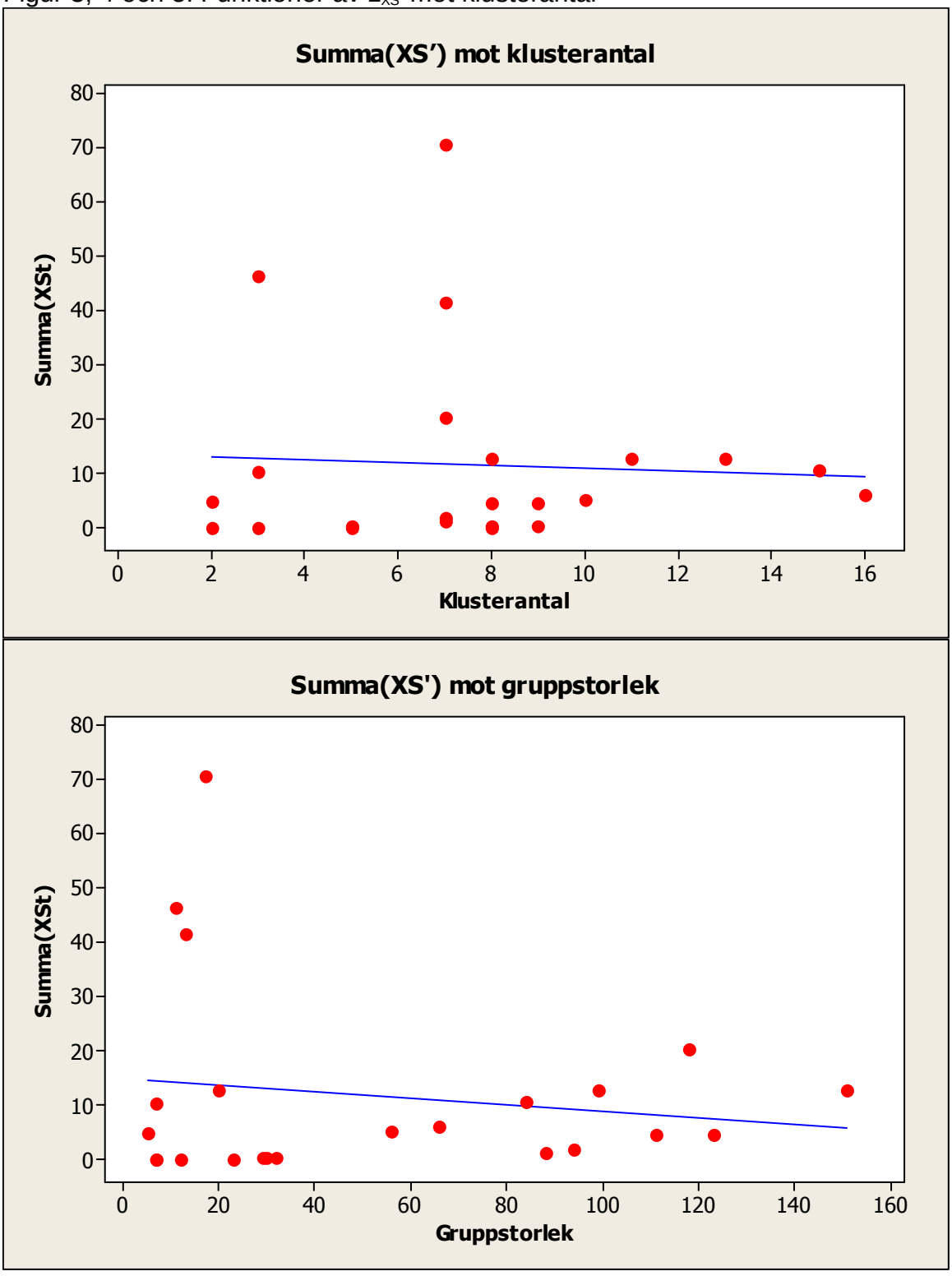
- Antal observationer för olika grupper (VVK)
- Antal kluster som valts för olika grupper (VVK)

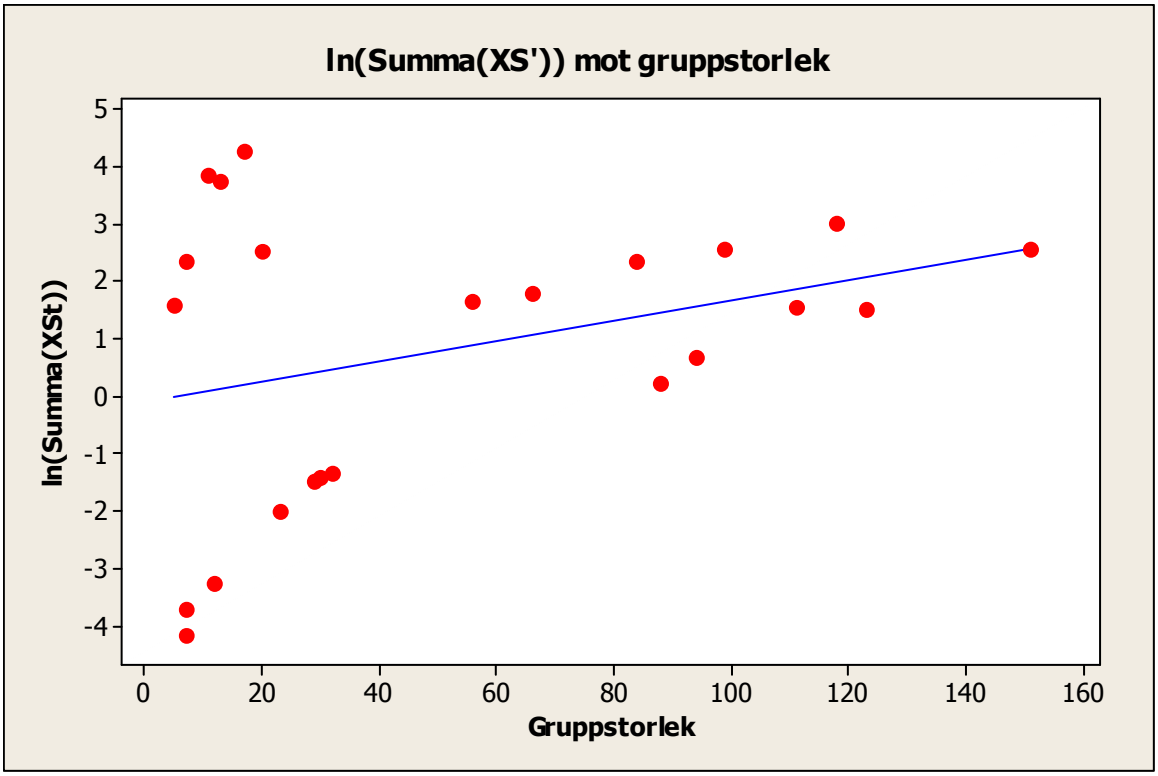
Värdet som använts för att välja grupper (VVK) är som bekant $\Sigma_{XS'}$. Låt oss plotta $\Sigma_{XS'}$ mot antal observationer för olika grupper respektive mot antal kluster som valts för olika grupper (VVK):

### 6.1.2. Första klustringsomgången

” ln ΣXS’” är programmets (Excel) beteckning för $\ln(\Sigma_{XS'})$. Denna kolumn och dess plott ingår för att misstanken att det finns ett exponentiellt samband att ta hänsyn till, mellan Gruppstorleken (Antal obs.) och ΣXS’, ska kunna avfärdas. Misstanken grundar sig på att plotten för Gruppstorlek har några höga ΣXS-värden för låga, och endast för låga, gruppstorlekar.

Figur 3, 4 och 5: Funktioner av $\Sigma_{XS'}$ mot klusterantal






Det finns ingen uppenbart *otillbörlig* påverkan på valet av grupper från antalet observationer eller antalet kluster för grupperna; misstanken att det finns ett exponentiellt samband att ta hänsyn till, mellan Gruppstorleken (Antal obs.) och ΣXS', avfärdas eftersom de observationer som verkar följa ett exponentiellt samband gör detta längs en linje med riktningskoefficient nära noll. Dessutom påverkar deras (eventuella) följande av exponentiellt samband knappast ordningen i ln($\Sigma_{XS'}$)-led mellan sig själva och övriga observationer.

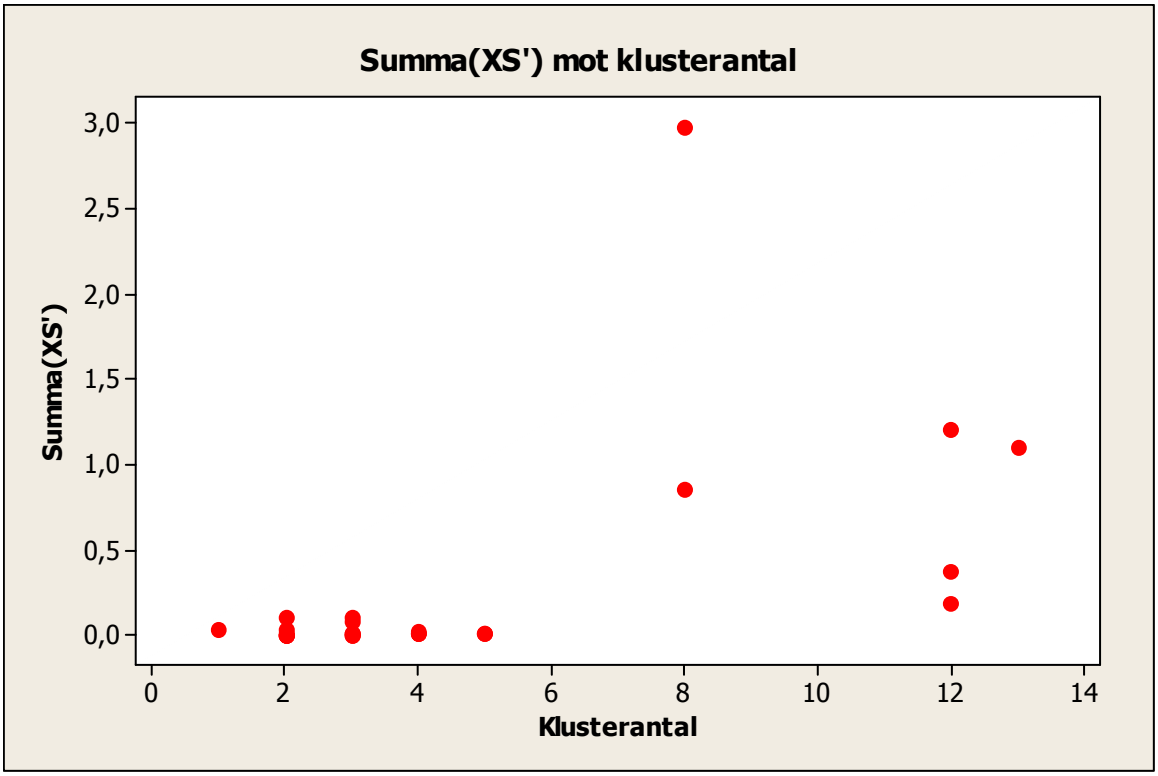
Den första klustringsomgångens val av grupper (VVK) blev dem med tabellens nio högsta $\Sigma_{XS'}$-värden. Valet är gjort genom subjektiv bedömning av vilka värden av de högsta som ligger tätast.

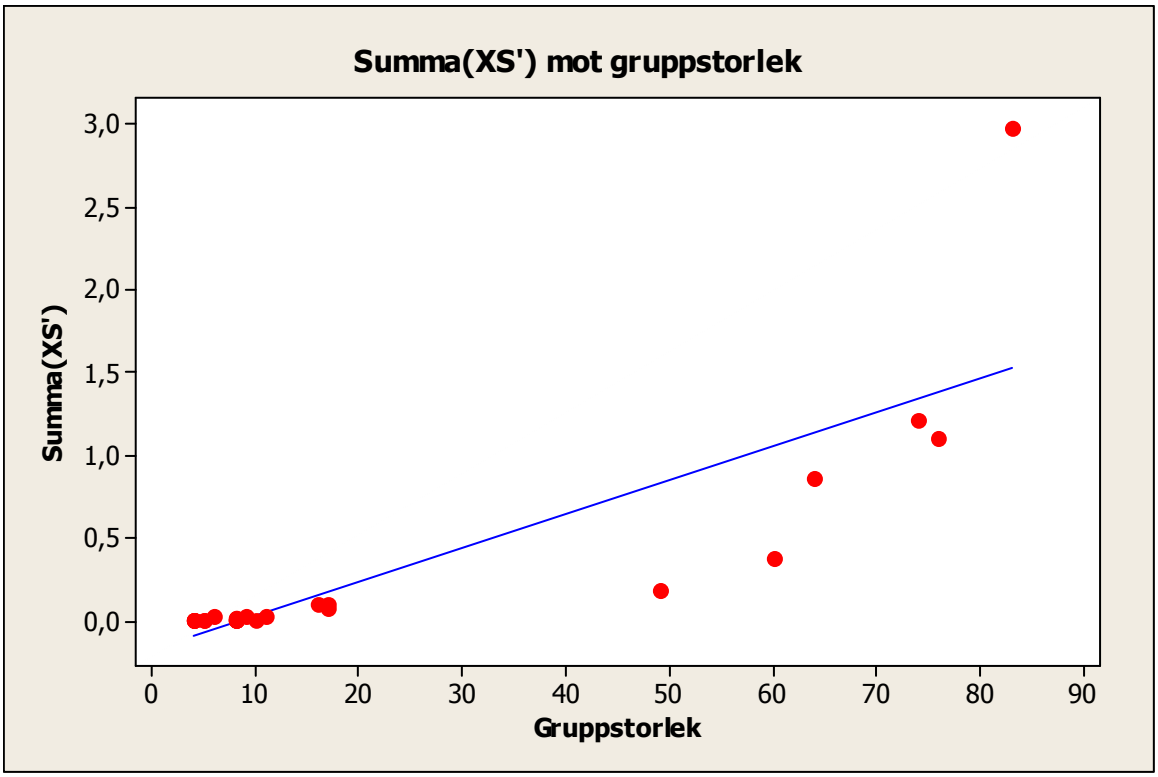
## 6.1.2 Andra klustringsomgången

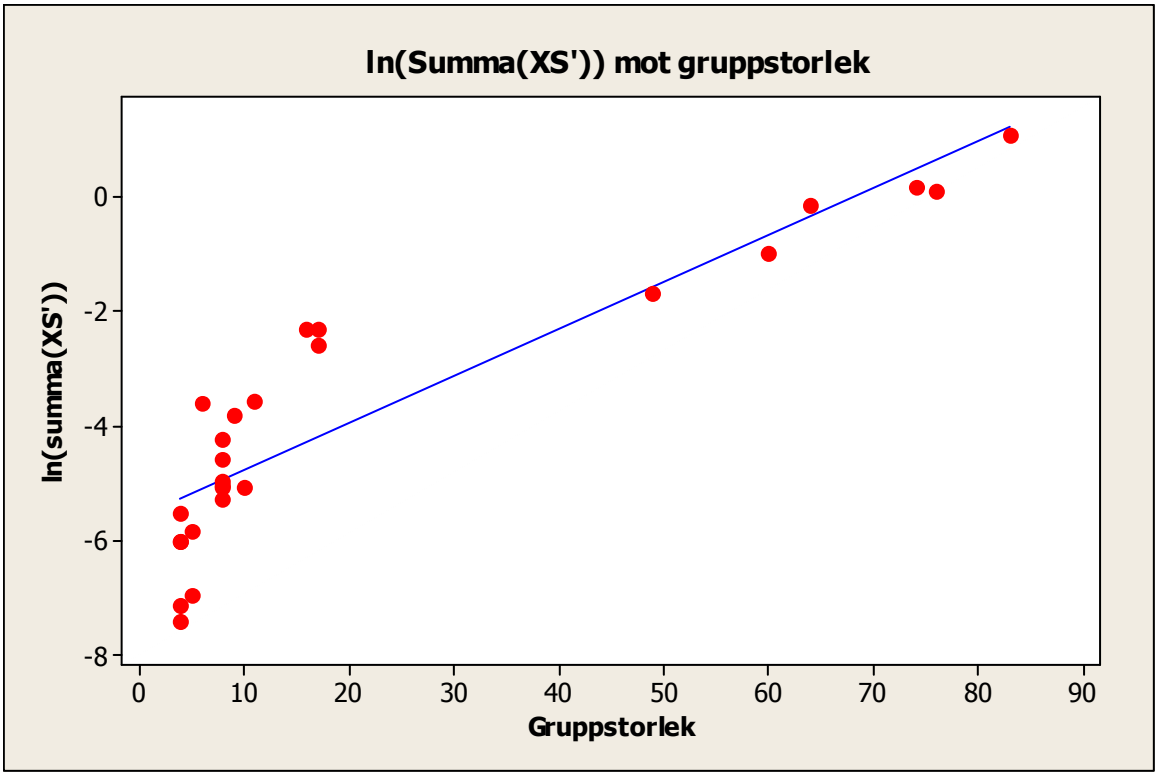
" ln ΣXS'" är programmets (Excel) beteckning för ln ($\Sigma_{XS'}$). Denna kolumn och dess plott ingår för att misstanken att det finns ett logaritmiskt samband att ta hänsyn till, mellan Gruppstorleken (Antal obs.) och ΣXS', ska kunna avfärdas. Misstanken grundar sig på att plotten för Gruppstorlek har några höga ΣXS-värden för låga, och endast för låga, gruppstorlekar.

Figur 6: $\Sigma_{XS'}$ mot klusterantal

Figur 7, 8 och 9: Funktioner av $\Sigma_{XS'}$ mot observationsantal = gruppstorlek


Summa(XS') mot klusterantal
Summa(XS')
3,0
2,5
2,0
1,5
1,0
0,5
0,0
0
2
4
6
8
10
12
14
Klusterantal


Summa(XS') mot gruppstorlek
Summa(XS')
3,0
2,5
2,0
1,5
1,0
0,5
0,0
0
10
20
30
40
50
60
70
80
90
Gruppstorlek




Proceduren bör avbrytas eftersom det nu finns ett tydligt samband mellan $\Sigma_{XS'}$ och Gruppstorlek. Den andra klustringsomgångens val av grupper av VVK blev ändå dem med tabellens sex högsta $\Sigma_{XS'}$-värden, trots att dessa också är värdena med ett närmast exakt exponentiellt samband med gruppstorlek. Detta eftersom detta deras följande av ett exponentiellt samband knappast påverkar ordningen i $\ln(\Sigma_{XS'})$-led mellan dem och övriga observationer.

Valet är gjort genom subjektiv bedömning av vilka värden av de högsta som ligger tätast, sedan antalet observationer har plottats mot en kolumn vars alla element är siffran 1.
Det slutliga valet av grupper av VVK blev enligt följande Tabell 2:

ÖV STAD
LGH ENS
LGH ÖV
ENS STAD
LGH STAD

### 6.1.3 Efter sista klustringsomgången

Nu vidtar förberedelser för kontroll av att VVK är tillräckligt väl samlade inom de utvalda klustren och deras områden. ”Tillräckligt” avser tillräckligt för att göra resultaten användbara för klassificering av patientgrupper om vilka man endast känner varians och medelvärde för vårdkostnaden VK. Med andra ord vidtar föreberedelser för kontroll av första punkten:

- Det måste finnas ovanligt många av den sökta sortens observationer just i varje klusters område, i diagrammet över Testdata

Som en bonus får man då föreberedelser för kontroll av andra punkten:

- Det får inte finnas för stor andel icke sökta observationer i varje klusters område i diagrammet över Testdata

Innan punkterna just ovan kan kontrolleras måste kluster väljas ut för varje grupp VVK. Utskriften från programmet, för varje grupp den har klusterindelat med hjälp av K-mean-metoden, ger kolumner med följande namn:

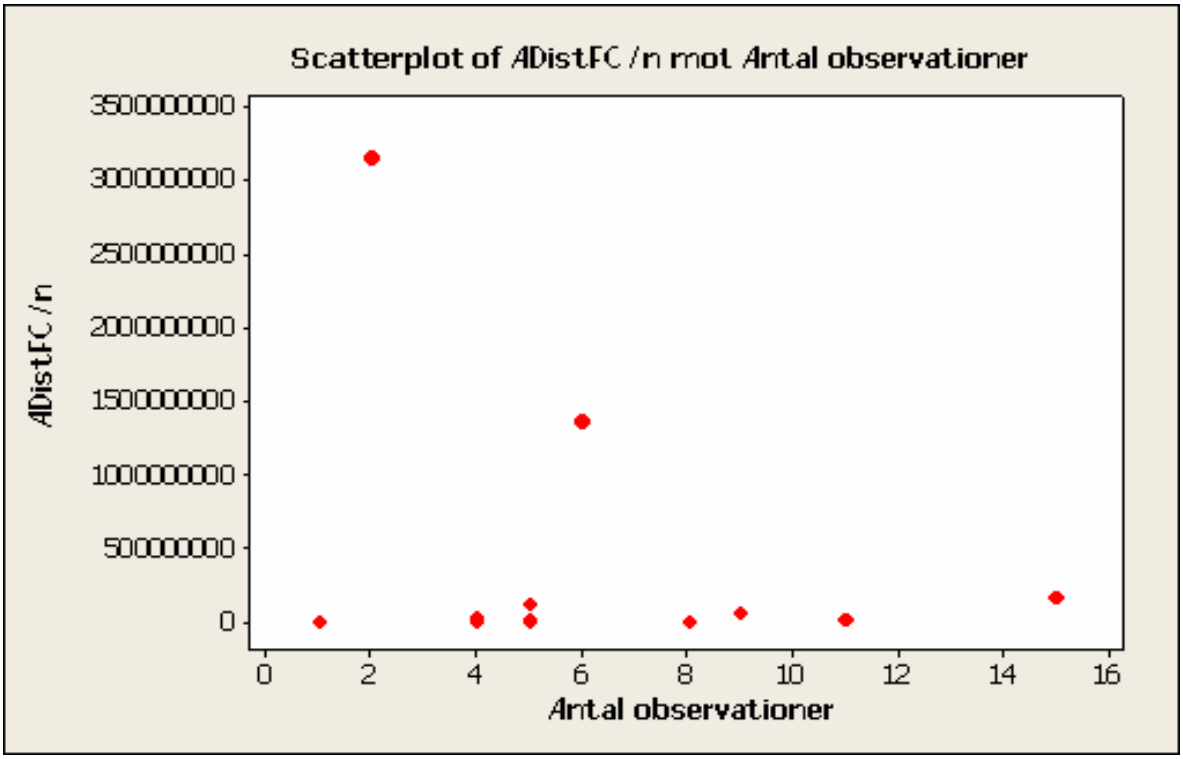
| Number of observations | Within cluster sum of squares | Average distance from centroid | Maximum distance from centroid |
|---|---|---|---|

Två av dessa kolumner har modifierats till att innehålla siffror att välja kluster efter. Resonemanget har varit att antalet observationer är ”ganska mycket” viktigare än hur löst sammanhållet klustret är. Till att börja med stryks alla kluster med färre än fyra observationer. Kluster med färre observationer än så har obefintlig sannolikhet att få observationer bland den praktiska tillämpningens data. Därefter har tagits fram de två kolumnerna ” Mean within cluster sum of squares” och “(Average distance from centroid)/ n”, förkortat namngivna enligt följande:

| Mean ClSS | ADistFC / n |
|---|---|

Innan kluster kan väljas för varje grupp med hjälp av värdena Mean ClSS och ADistFC / n måste kontrolleras att dessa värden *inte eller knappast påverkas av* Antal observationer *för respektive kluster*. Emellertid är det självklart att Mean ClSS påverkas av Antal observationer för respektive kluster. Eftersom påverkan blir större per observation ju större värde som Mean ClSS har, så väljes låga Mean ClSS.

Påverkan från Antal observationer på ADistFC / n kontrolleras grafiskt. Vi har här ett dåligt statistiskt underlag eftersom det är svårt att hitta ett sätt att justera så att jämförelser mellan värden från olika grupper blir meningsfulla. Därmed fås så få värden i varje kontroll att subjektiv bedömning av plotter känns lika säker som något annat. Den subjektiva bedömningen av plotter av ADistFC / n mot Antal observationer är att det valda värdet inte eller knappast påverkas av antal observationer för respektive kluster. Samtliga dessa plotter är väldigt lika följande exempel, som vi kallar Figur 10:




Valet av kluster har också den gjorts fullständigt subjektivt, med stöd av nämnda kolumner. Avvägningen har varit mellan att välja de tydligaste grupperna av lägsta ADistFC / n och att välja låga Mean ClSS, utan att välja fler än 3-4 stycken kluster per grupp.

Utskrifterna av nämnda kolumner redovisas tillsammans med beskrivningar av de kluster som väljs därur, i 9. ”*Bilaga*”.

## 6.2 Slutsats ur resultat med hjälp av Testdata

Testdata är de 20 % av data som plockades ut innan kluster började sökas. Testdata täcker i stort sett samma område som vad resten av data täcker, när man ritar upp dem i ett diagram. Proceduren som använts är avsedd att hitta områden med särskilt stor andel av respektive grupp variabelvärden. Om ett visst område i Data har särskilt hög andel av viss grupp variabelvärden så är det alltså ett gott resultat om samma område i Testdata också har särskilt hög andel av samma grupp variabelvärden. Så väl är det dessvärre inte, om man får tro Poisson i det här sammanhanget.

Poissonfördelningens sannolikhetsvärden talar som bekant om antalet observationer inom ett visst intervall av något slag. Goda resultat skulle ha varit osannolikt många av just den sort man sökte i varje visst område och osannolikt få av alla andra i just detta område.

### 6.2.1 Testdatas bearbetning

Centroiderna för klustren har använts för att klassificera Testdata runt, vartefter resultatet jämförts med det enligt ovan önskade. För varje valt kluster har de testobservationer plockats fram vars avstånd till klustrets medelpunkt är mindre än medelavståndet till samma medelpunkt för de observationer som bildat klustret.

För att jämföra resultatet har behövts mätvärden för ifall resultatet inom varje grupps klusters områden är bra eller dåligt. Här har valts Poissonfördelningens sannolikhetsvärden för det observerade antalet observationer inom ett medelavstånds radie från medelpunkten enligt ovan.

## 6.3 Slutsats ur resultat

Tabellerna visar sannolikhetsvärden för det funna antalet observationer runt respektive klusters centroid, inom det område i diagrammet som Testdata täcker.

De övre tabellerna, med observationsantal x, visar Poissonfördelningens sannolikhetsvärden för x stycken observationer i området runt respektive klusters centroid. x är det funna antalet observationer av alla VVK i området runt respektive klusters centroid, utom sådana som just det klustret är framtaget för.

De undre tabellerna, med observationsantal y, visar Poissonfördelningens sannolikhetsvärden för det funna antalet observationer av just de VVK som klustren är framtagna för.

Resultat som betecknas ” rätt ” innebär att slh är liten att det skall finnas så många observationer som det finns i det aktuella området av det slag som dess kluster är framtaget för, samtidigt som slh är stor att det det skall finnas så många observationer som det finns i det aktuella området av andra slag än det som dess kluster är framtaget för.

Över respektive tabell anges det vilka VVK som motsvarande kluster är framtagna för.

ICKE KUMULATIVT, ÖV STAD

| Kluster X | 8 | 9 | 10 | 11 |
|---|---|---|---|---|
| 0 | 0,979219 | | 0,975583 | 0,985655 |
| 1 | | 0,132455 | | |
| 2 | | | | |
| 3 | | | | |
| 4 | | | | |
| 5 | | | | |
| 6 | | | | |
| 7 | | | | |
| 8 | | | | |
| Y | | | | |
| 0 | 0,959656 | 0,741577 | | 0,972369 |
| 1 | | | 0,045551 | |
| 2 | | | | |
| 3 | | | | |
| 4 | | | | |
| 5 | | | | |
| 6 | | | | |
| 7 | | | | |
| 8 | | | | |
| | fel | rätt | fel | fel |

ICKE KUMULATIVT, ENS LGH

| Kluster X | 1 | 2 | 8 | 10 |
|---|---|---|---|---|
| 0 | | 0,572381 | 0,27468 | |
| 1 | | | | 0,084766 |
| 2 | | | | |
| 3 | | | | |
| 4 | 0,000133 | | | |
| 5 | | | | |
| 6 | | | | |
| 7 | | | | |
| 8 | | | | |
| Y | | | | |
| 0 | | 0,314821 | 0,071549 | |
| 1 | 0,310406 | | | |
| 2 | | | | |
| 3 | | | | 0,000984 |
| 4 | | | | |
| 5 | | | | |
| 6 | | | | |
| 7 | | | | |
| 8 | | | | |
| | rätt | fel | fel | fel |

Tabell 3

ICKE KUMULATIVT LGH ÖV

| Kluster X | 5 | 6 | 7 | 2 |
|---|---|---|---|---|
| 0 | | | 0,241142 | |
| 1 | | | | |
| 2 | | 0,02883 | | 0,193629 |
| 3 | | | | |
| 4 | 8,06E-05 | | | |
| 5 | | | | |
| 6 | | | | |
| 7 | | | | |
| 8 | | | | |
| Y | | | | |
| 0 | | 0,738479 | | |
| 1 | | | 0,256436 | 0,363665 |
| 2 | | | | |
| 3 | | | | |
| 4 | | | | |
| 5 | 5,63E-06 | | | |
| 6 | | | | |
| 7 | | | | |
| 8 | | | | |
| | fel | rätt | rätt | rätt |

ICKE KUMULATIVT ENS ÖV

| Kluster X | 9 | 11 | 1 |
|---|---|---|---|
| 0 | 0,979219 | 0,856758 | |
| 1 | | | 0,024116 |
| 2 | | | |
| 3 | | | |
| 4 | | | |
| 5 | | | |
| 6 | | | |
| 7 | | | |
| 8 | | | |
| Y | | | |
| 0 | | 1,85E-01 | |
| 1 | | | 0,045551 |
| 2 | | | |
| 3 | | | |
| 4 | | | |
| 5 | 2,91E+00 | | |
| 6 | | | |
| 7 | | | |
| 8 | | | |
| | fel | fel | rätt |

Tabell 4

| | ICKE KUMULATIVT | | ENS STAD | |
|---|---|---|---|---|
| Kluster | 4 | 3 | 1 | |
| X | | | | |
| 0 | 0,170333 | 0,921576 | | |
| 1 | | | | 0,052097 |
| 2 | | | | |
| 3 | | | | |
| 4 | | | | |
| 5 | | | | |
| 6 | | | | |
| 7 | | | | |
| 8 | | | | |
| Y | | | | |
| 0 | | 0,710992 | | 0,794618 |
| 1 | | | | |
| 2 | 0,016675 | | | |
| 3 | | | | |
| 4 | | | | |
| 5 | | | | |
| 6 | | | | |
| 7 | | | | |
| 8 | | | | |
| | fel | fel | rätt | |

| | ICKE KUMULATIVT | | LGH STAD |
|---|---|---|---|
| Kluster | 6 | 5 | 1 |
| X | | | |
| 0 | | 0,977884 | 0,933743 |
| 1 | | | |
| 2 | 0,096711 | | |
| 3 | | | |
| 4 | | | |
| 5 | | | |
| 6 | | | |
| 7 | | | |
| 8 | | | |
| Y | | | |
| 0 | | 0,916703 | |
| 1 | | | |
| 2 | | | |
| 3 | | | |
| 4 | | | |
| 5 | 0,053775 | | |
| 6 | | | |
| 7 | | | |
| 8 | | | |
| 9 | | | 1,44E-11 |
| | fel | fel | fel |

Tabell 5

# 7 Slutsats och diskussion

Resultaten som helhet avslöjar knappast huruvida uppsatsens syfte uppnåtts. Möjligen tyder resultaten på att uppsatsens syfte inte uppnåtts men det som tyder därpå är främst nollresultat. Som väl är har bristen på slutsats en ursäkt: Det har på sluttampen visat sig att klustermetoden skulle ha valts med större omsorg.

| | |
|---|---|
| Antal resultatobservationer: | 21 |
| Antal resultatobservationer visande på icke uppnått syfte: | 14 |
| Antal nollresultat bland resultatobservationer visande på icke uppnått syfte: | 8 |

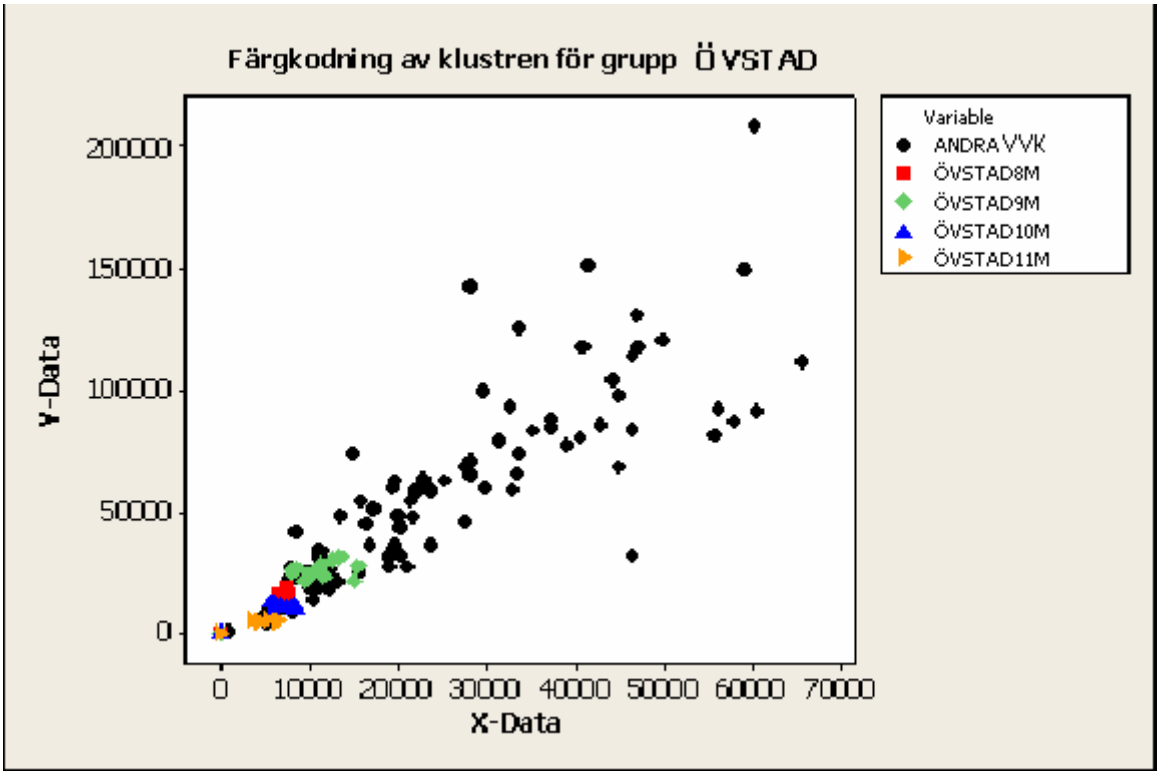
Tolkningen av utfallet noll funna observationer är värd en eftertanke:
Det kan mycket väl vara tunnsått med observationer i ett klusters område utan att klustret är felaktigt valt eller framtaget med olämplig metod. Det kan ju nämligen vara tunnsått med observationer i området som klustret definierats i medan de observationer som ändå finns nästan uteslutande är av gruppens sort. Noll observationer funna betyder alltså inte nödvändigtvis att klustermetoden är missvisande eller ens illa vald. Så länge det är betydligt lägre sannolikhet att hitta vad man söker än att hitta vad man inte söker så kan inte området förkastas som olämpligt eller illa valt. Å andra sidan innebär noll observationer naturligtvis inga belägg för vare sig hypotesen eller mothypotesen.

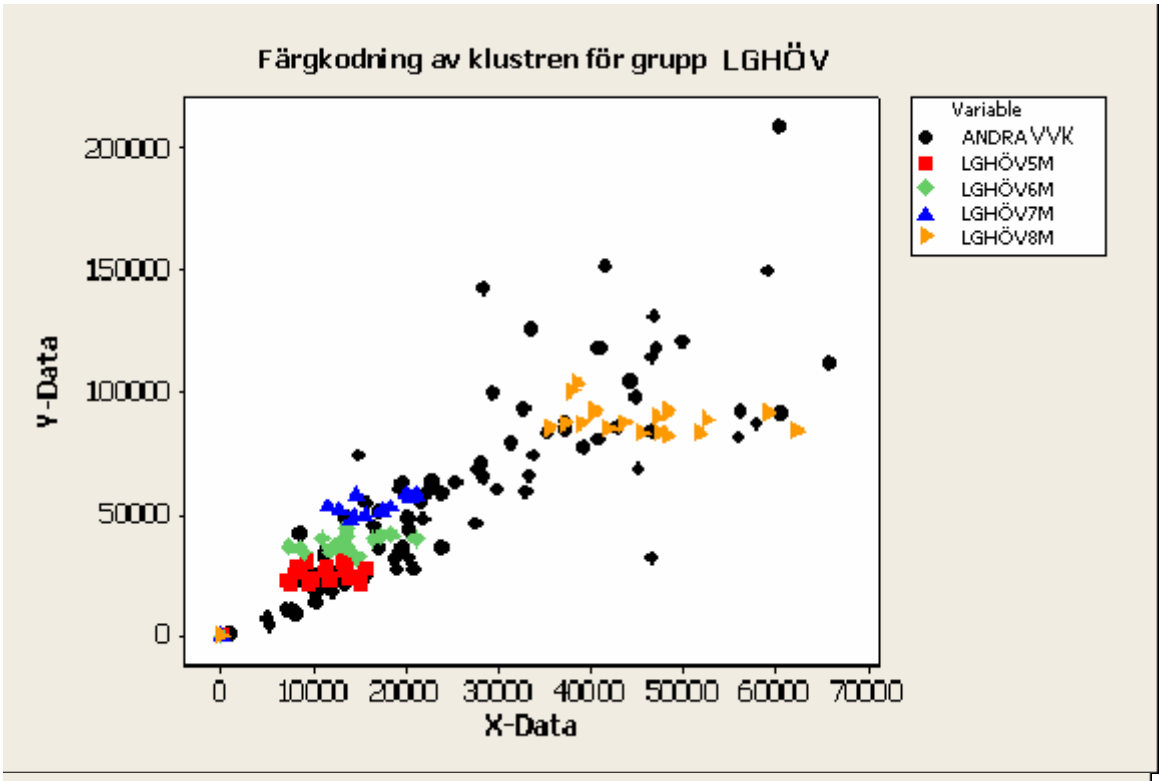
Förutom behov av djupare avväganden inför metodvalet har det enligt tidigare framkommit behov av resonemang kring samband mellan $\Sigma_{XS'}$ och gruppstorlek. Det vore väl om det gick att justera för detta samband utan att resultaten blev missvisande. Det är just nu okänt om sådan missvisning har del i bristerna hos denna uppsats' resultat.

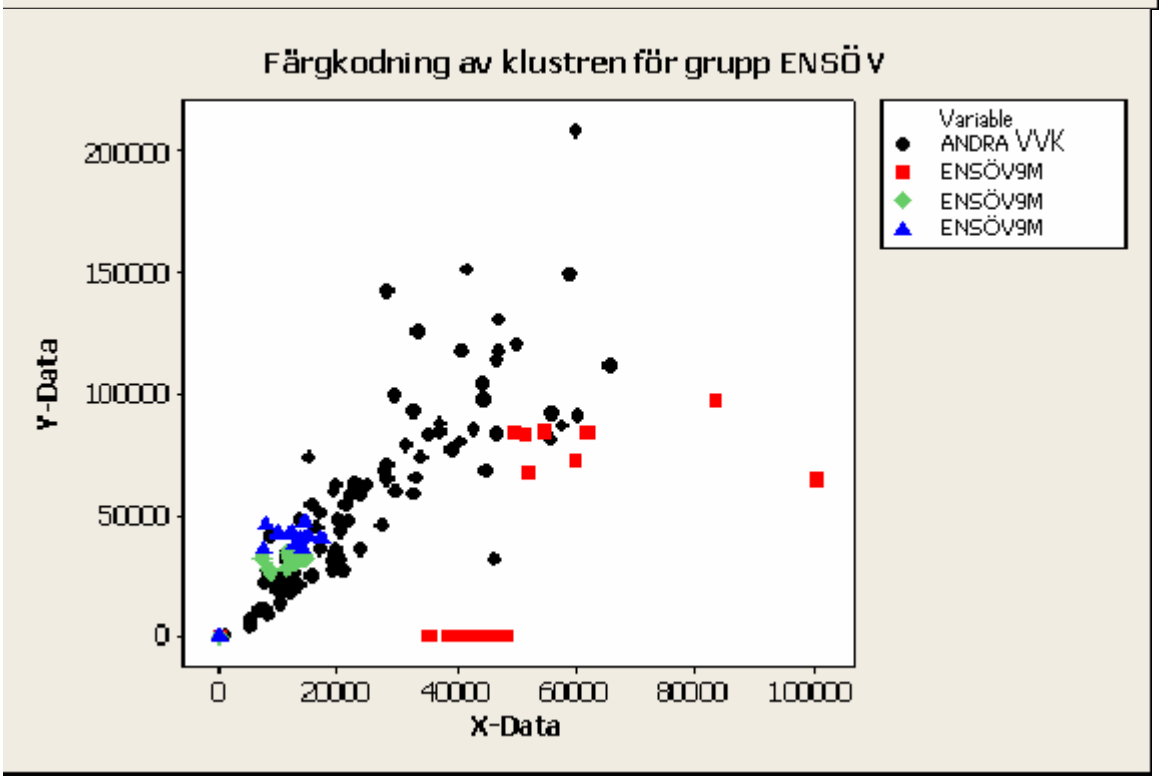
Centroiderna för klustren har använts för att klassificera Testdata runt, vartefter resultatet jämförts med det enligt ovan önskade. För varje valt kluster har de testobservationer plockats fram vars avstånd till klustrets medelpunkt är mindre än medelavståndet till samma medelpunkt för de observationer som bildat klustret. Förmodligen skulle resultatet ha stöttat hypotesen bättre ifall de testobservationer plockats fram som ligger inom de ellipser som täcker de observationer som bildat respektive kluster.

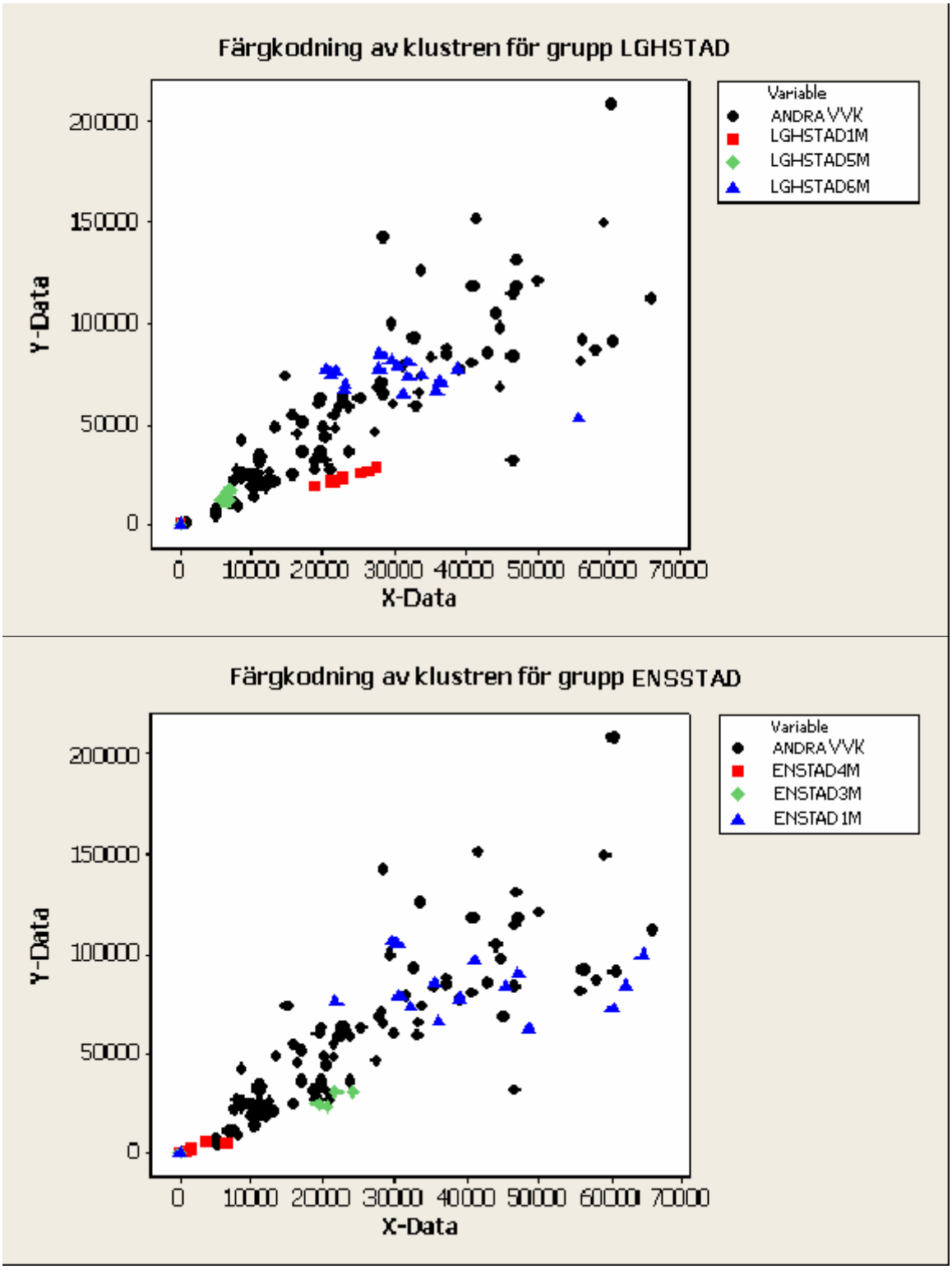
## 7.1 Den valda klustermetodens brister

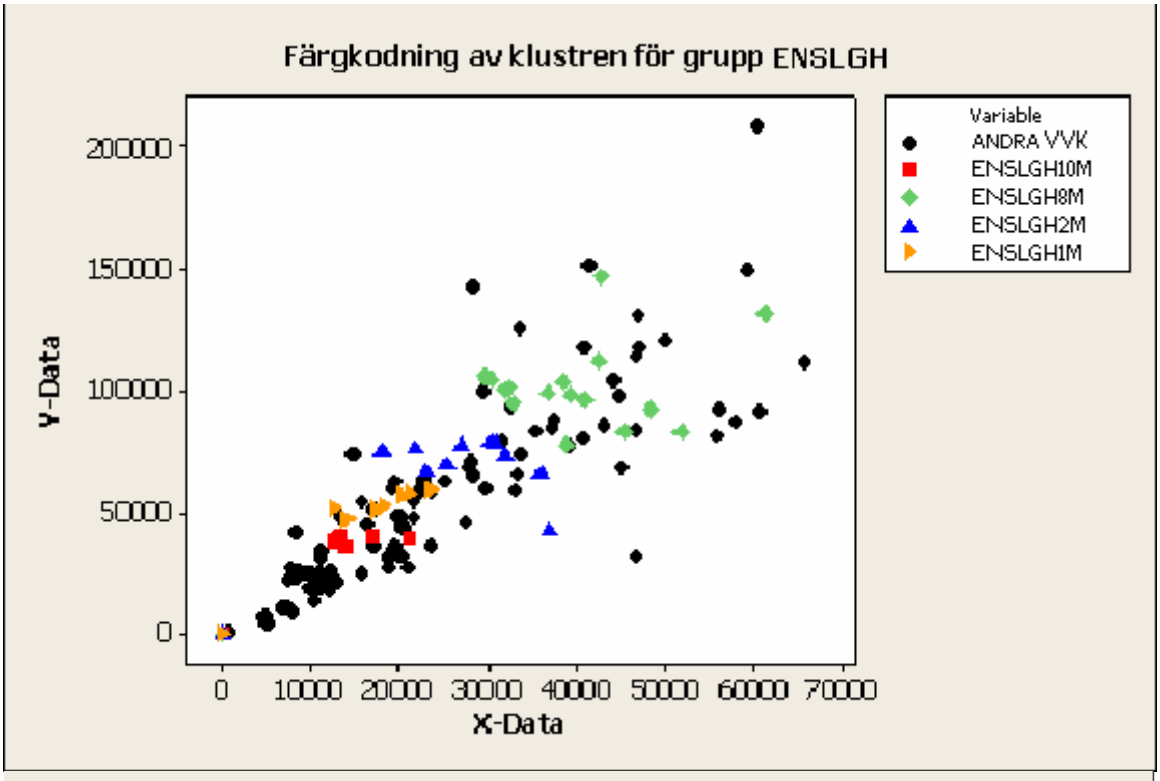
Låt oss titta på samtliga de utvalda klustren för respektive grupp, inritade i diagrammet över Testdata. Låt oss alltså titta på Figur 11, 12, 13, 14, 15, 16 och 17, där varje kluster har en viss färg och innehåller en viss av de VVK som delar de variabelvärden (exempelvis ÖVSTAD), som kännetecknar gruppen som det aktuella diagrammet beskriver.

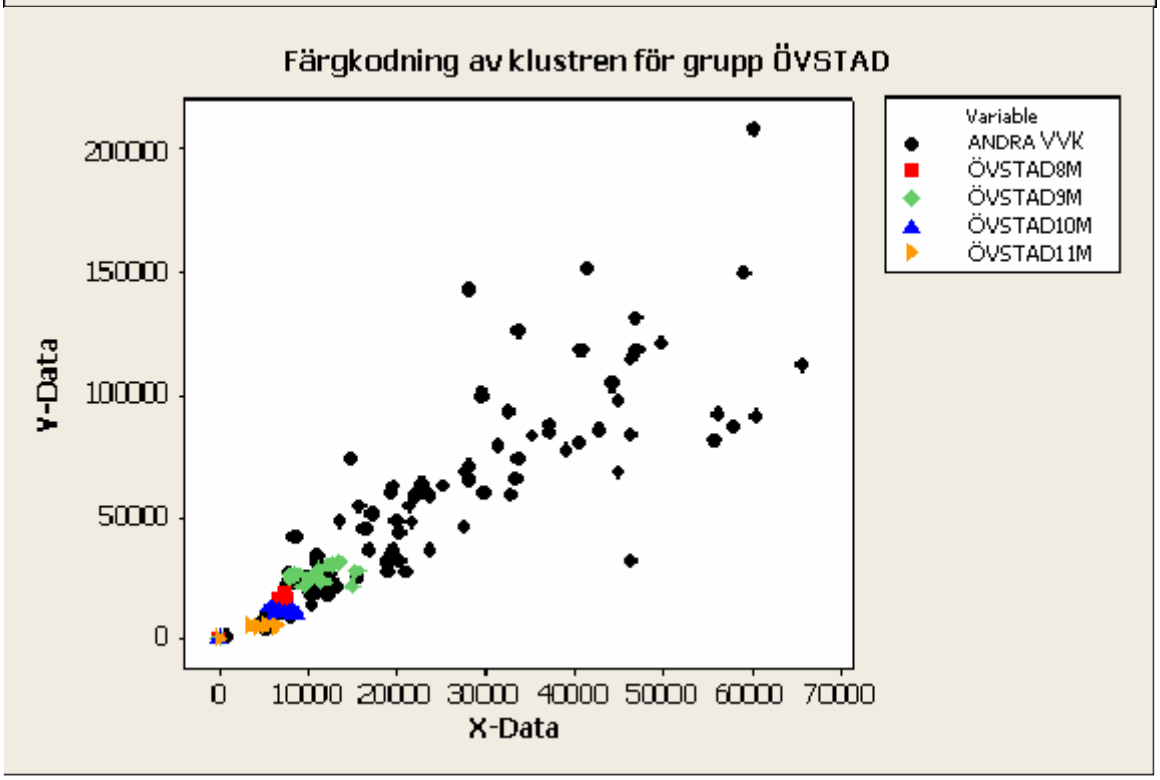



Färgkodning av klustren för grupp LGHÖV
Variable
ANDRA VVK
LGHÖV5M
LGHÖV6M
LGHÖV7M
LGHÖV8M
Y-Data
X-Data
0
50000
100000
150000
200000
0
10000
20000
30000
40000
50000
60000
70000


Färgkodning av klustren för grupp ENSÖV
Variable
ANDRA VVK
ENSÖV9M
ENSÖV9M
ENSÖV9M
Y-Data
X-Data
0
50000
100000
150000
200000
0
20000
40000
60000
80000
100000


Färgkodning av klustren för grupp LGHSTAD
Variable
ANDRA VVK
LGHSTAD1M
LGHSTAD5M
LGHSTAD6M
Y-Data
X-Data
0
50000
100000
150000
200000
0
10000
20000
30000
40000
50000
60000
70000
Färgkodning av klustren för grupp ENSSTAD
Variable
ANDRA VVK
ENSTAD4M
ENSTAD3M
ENSTAD 1M
Y-Data
X-Data


Färgkodning av klustren för grupp ENSLGH
Y-Data
X-Data
0
50000
100000
150000
200000
0
10000
20000
30000
40000
50000
60000
70000
Variable
ANDRA VVK
ENSLGH10M
ENSLGH8M
ENSLGH2M
ENSLGH1M


Färgkodning av klustren för grupp ÖVSTAD
Y-Data
X-Data
0
50000
100000
150000
200000
0
10000
20000
30000
40000
50000
60000
70000
Variable
ANDRA VVK
ÖVSTAD8M
ÖVSTAD9M
ÖVSTAD10M
ÖVSTAD11M

Figurer 11-17 visar den K-mean-metod som valts oftast ger långsmala kluster när den används på det aktuella materialet. Litteraturen och gissningar ger vid hand att metoden är skapt så att den företrädesvis hittar kluster som liknar varandra till formen, sedan algoritmen på någon grund valt en favoritform i just det aktuella materialet. Gissningsvis har algoritmen valt strecklika kluster delvis på grund av att materialet bildar en långsmal kvast i materialet. Kanske borde ett program ha använts som visar tillräckligt med detaljer om hur dess K-mean-metod är uppbyggd för att kunna låta användaren hitta fakta i litteraturen om metodens uppförande. Kanske borde de teoretiska avvägandena och insamlingen av goda råd inför valet av klustermetod ha varit grundligare.

# 8. Litteraturförteckning


Andersson P -A; Varde E; Diderichsen F, *Modelling of resource allocation to health care authorities in Stockholm County*, Health Care Management Science, 2000, 141-149

Hjälpfunktion i MINITAB 14

Everitt Brian, Cluster analysis, London (1974)

Gordon A. D., Classification, Chapman & Hall/CRC, Boca Raton (1999)


# 9. Bilaga

Val av kluster och värden för val av kluster, för respektive grupp (VVK):

Tabell 4, Tabell 5, VVK: ÖV STAD
K-means Cluster Analysis: C2; SQRT(C3)
Final Partition

Number of clusters: 13

| | Number of observations | Within cluster sum of squares | Average distance from centroid | Maximum distance from centroid |
|---|---|---|---|---|
| Cluster1 | 6 | 8,22523E+09 | 33788,573 | 51425,252 |
| Cluster2 | 4 | 7,25485E+07 | 4149,343 | 5373,349 |
| Cluster5 | 15 | 2,59180E+09 | 11818,834 | 25440,583 |
| Cluster7 | 5 | 4,34415E+07 | 2386,816 | 5819,677 |
| Cluster8 | 5 | 1,27935E+07 | 1350,840 | 2459,665 |
| Cluster9 | 11 | 1,61571E+08 | 3561,592 | 5622,715 |
| Cluster10 | 8 | 1,98019E+07 | 1463,694 | 2093,587 |
| Cluster11 | 4 | 5683431,362 | 1129,151 | 1675,556 |
| Cluster12 | 9 | 5,23500E+08 | 6961,739 | 12313,056 |
| Cluster13 | 5 | 5,87275E+08 | 10455,015 | 14408,292 |

| | Mean ClSS | | ADistFC / n |
|---|---|---|---|
| Cluster1 | 1,37E+09 | | 5631,429 |
| Cluster2 | 1,81E+07 | | 1037,336 |
| Cluster5 | 1,73E+08 | | 787,9223 |
| Cluster7 | 8,69E+06 | | 477,3632 |
| Cluster8 | 2,56E+06 | | 270,168 |
| Cluster9 | 1,47E+07 | | 323,7811 |
| Cluster10 | 2,48E+06 | | 182,9618 |
| Cluster11 | 1,42E+06 | | 282,2878 |
| Cluster12 | 5,82E+07 | | 773,5266 |
| Cluster13 | 1,17E+08 | | 2091,003 |

Val: 10
8
9
11

Tabell 6, Tabell 7, VVK: ENS LGH
K-means Cluster Analysis: C13; C16
Final Partition

Number of clusters: 12

| | Number of observations | Within cluster sum of squares | Average distance from centroid | Maximum distance from centroid | |
|---|---|---|---|---|---|
| Cluster1 | 7 | | 2,02503E+08 | 4850,155 | 7971,972 |
| Cluster2 | 10 | | 5,96317E+08 | 7200,117 | 11253,204 |
| Cluster3 | 4 | | 5,93740E+07 | 3771,236 | 4664,973 |
| Cluster8 | 18 | | 2,70496E+09 | 10950,596 | 25416,953 |
| Cluster9 | 4 | | 2,07437E+09 | 22238,154 | 28033,818 |
| Cluster10 | 7 | | 7,44626E+07 | 2939,898 | 6102,420 |

| | Mean ClSS | | ADistFC / n |
|---|---|---|---|
| Cluster1 | 2,89E+07 | | 692,8793 |
| Cluster2 | 5,96E+07 | | 720,0117 |
| Cluster3 | 1,48E+07 | | 942,809 |
| Cluster8 | 1,50E+08 | | 608,3664 |
| Cluster9 | 5,19E+08 | | 5559,539 |
| Cluster10 | 1,06E+07 | | 419,9854 |

Val: 10
8
1
2

Tabell 8, Tabell 9, VVK: LGH ÖV
K-means Cluster Analysis: C23; C26
Final Partition

Number of clusters: 8

| | Number of Observations | Within cluster sum of squares | Average distance from centroid | Maximum distance from centroid | |
|---|---|---|---|---|---|
| Cluster2 | 18 | 1,58293E+09 | 8280,174 | 17592,840 | |
| Cluster3 | 12 | 6,19265E+08 | 6709,627 | 9409,291 | |
| Cluster5 | 18 | 2,97929E+08 | 3797,078 | 6165,625 | |
| Cluster6 | 16 | 3,52773E+08 | 4233,748 | 7708,693 | |
| Cluster7 | 11 | 2,68144E+08 | 4689,796 | 6596,581 | |
| Cluster8 | 5 | 4,46595E+07 | 2829,942 | 4707,853 | |

| | Mean ClSS | | ADistFC / n |
|---|---|---|---|
| Cluster2 | 8,79E+07 | | 460,0097 |
| Cluster3 | 5,16E+07 | | 559,1356 |
| Cluster5 | 1,66E+07 | | 210,9488 |
| Cluster6 | 2,20E+07 | | 264,6093 |
| Cluster7 | 2,44E+07 | | 426,3451 |
| Cluster8 | 8,93E+06 | | 565,9884 |

Val: 5
6
7
2

Tabell 10, Tabell 11, VVK: LGH STAD
K-means Cluster Analysis: C33; C36
Final Partition

Number of clusters: 8

| | Number of Observations | Within cluster sum of squares | Average distance from centroid | Maximum distance from centroid | |
|---|---|---|---|---|---|
| Cluster1 | 17 | 2,15290E+09 | 9245,798 | 32592,782 | |
| Cluster2 | 11 | 1,85744E+09 | 11768,227 | 20726,503 | |
| Cluster3 | 11 | 1,06365E+09 | 9133,290 | 18591,225 | |
| Cluster5 | 4 | 1,69546E+07 | 1798,769 | 3484,354 | |
| Cluster6 | 9 | 1,09197E+08 | 3149,299 | 5548,714 | |
| Cluster7 | 7 | 2,08689E+08 | 5044,014 | 8563,351 | |
| Cluster8 | 4 | 2,11706E+09 | 21123,943 | 33245,035 | |

| | Mean ClSS | | ADistFC / n |
|---|---|---|---|
| Cluster1 | 1,27E+08 | | 543,8705 |
| Cluster2 | 1,69E+08 | | 1069,839 |
| Cluster3 | 9,67E+07 | | 830,2991 |
| Cluster5 | 4,24E+06 | | 449,6923 |
| Cluster6 | 1,21E+07 | | 349,9221 |
| Cluster7 | 2,98E+07 | | 720,5734 |
| Cluster8 | 5,29E+08 | | 5280,986 |

Val: 6
5
1

Tabell 12, 13, VVK: ENS ÖV
K-means Cluster Analysis: C44; C47
Final Partition

Number of clusters: 12

| | Number of observations | Within cluster sum of squares | Average distance from centroid | Maximum distance from centroid | |
|---|---|---|---|---|---|
| Cluster1 | 18 | 6,59658E+09 | 14935,128 | 51663,467 | |
| Cluster2 | 4 | 1,25363E+10 | 48396,659 | 92417,375 | |
| Cluster4 | 4 | 5,24055E+07 | 3184,089 | 5116,187 | |
| Cluster5 | 6 | 5,23461E+08 | 8834,561 | 13758,560 | |
| Cluster6 | 4 | 3,56309E+07 | 2903,818 | 3744,360 | |
| Cluster7 | 9 | 9,34501E+08 | 9500,094 | 16425,678 | |
| Cluster9 | 12 | 2,26108E+08 | 3765,450 | 7126,629 | |
| Cluster11 | 10 | 1,14124E+08 | 3189,554 | 4702,870 | |

| Mean ClSS | | ADistFC / n | |
|---|---|---|---|
| Cluster1 | 3,66E+08 | | 829,7293 |
| Cluster2 | 3,13E+09 | | 12099,16 |
| Cluster4 | 1,31E+07 | | 796,0223 |
| Cluster5 | 8,72E+07 | | 1472,427 |
| Cluster6 | 8,91E+06 | | 725,9545 |
| Cluster7 | 1,04E+08 | | 1055,566 |
| Cluster9 | 1,88E+07 | | 313,7875 |
| Cluster11 | 1,14E+07 | | 318,9554 |

Val: 9
11
1

Tabell 14, Tabell 15. VVK: ENS STAD
K-means Cluster Analysis: C54; C57
Final Partition

Number of clusters: 12

| | Number of observations | Within cluster sum of squares | Average distance from centroid | Maximum distance from centroid | |
|---|---|---|---|---|---|
| Cluster1 | 15 | 4,90859E+09 | 16480,391 | 28080,303 | |
| Cluster3 | 4 | 5,07152E+07 | 3537,059 | 3979,518 | |
| Cluster4 | 4 | 3,56309E+07 | 2903,818 | 3744,360 | |
| Cluster8 | 4 | 1,77920E+09 | 18508,320 | 34560,994 | |
| Cluster10 | 8 | 1,21757E+09 | 11559,447 | 17404,068 | |

| Mean ClSS | | ADistFC / n | |
|---|---|---|---|
| Cluster1 | 3,27E+08 | | 1098,693 |
| Cluster3 | 1,27E+07 | | 884,2648 |
| Cluster4 | 8,91E+06 | | 725,9545 |
| Cluster8 | 4,45E+08 | | 4627,08 |
| Cluster10 | 1,52E+08 | | 1444,931 |

Val: 4
3
1