Hemtentamen i Spatial data

Hampus Beijer

2025-12-19

Uppgift 1

I denna uppgift ska datamaterialet scallops analyseras. Datamaterialet består av fyra variabler, och presenteras i tabell 1:

- lat: lattitude.
- long: longitude .
- tcatch: antal fångade musslor.
- lgcatch: transformerade tcatch (log(tcatch)+1).

Table 1: Scallops datamaterial.

lat	long	tcatch	lgcatch
40.55000	-71.55000	0	0.0000000
40.46667	-71.51667	0	0.0000000
40.51667	-71.71667	0	0.0000000
40.38333	-71.85000	1	0.6931472
40.31667	-71.78333	0	0.0000000
40.26667	-71.88333	0	0.0000000
40.13333	-72.08333	2	1.0986120
40.06667	-72.16667	0	0.0000000
40.10000	-72.31667	7	2.0794420
40.01667	-72.40000	13	2.6390570
39.90000	-72.56667	530	6.2747620
39.81667	-72.48333	2750	7.9197200
39.85000	-72.58333	2060	7.6309470
39.81667	-72.68333	1016	6.9246120
39.76667	-72.73333	1206	7.0958930
39.68333	-72.61667	30	3.4339870
39.58333	-72.65000	110	4.7095300
39.50000	-72.53333	0	0.0000000
39.43333	-72.63333	0	0.0000000
39.46667	-72.66667	722	6.5834090

```
# Läser in datamaterialet
data <- read.table("/Users/hampusbeijer/Downloads/scallops.txt",
                   header = TRUE)
# Presenterar i en tabell
kable(head(data, n = 20), caption ="Scallops datamaterial.")
```

```
require(kableExtra)
require(maps)
require(sp)
require(gstat)
library(spdep)
library(spatialreg)
```

Deluppgift 1a

Rita upp en karta över området där data har samlats in, rita också in variabeln lgcatch på ett sätt som visar mätvärdets storlek på ett bra sätt.

Datamaterialet presenteras i två kartdiagram (figur 1). Datamaterialet kommer ursprungligen från väskusten i USA, vilket kan ses i den röda rutan i vänster figur. I figuren till höger presenteras kartdiagrammet inklusive observationerna. Dessa visar variabeln *lgcatch* som representerar antalet fångade musslor. Det kan noteras att observationerna har olika storlekar, där större ringar innebär fler musslor som har samlats in och mindre ringar menas att mindre musslor har fångats. Diagrammet till höger lyckas dock inte att fånga utspridningen av observationerna på ett tydligt sätt, därför visualiseras diagrammet igen i figur 2.

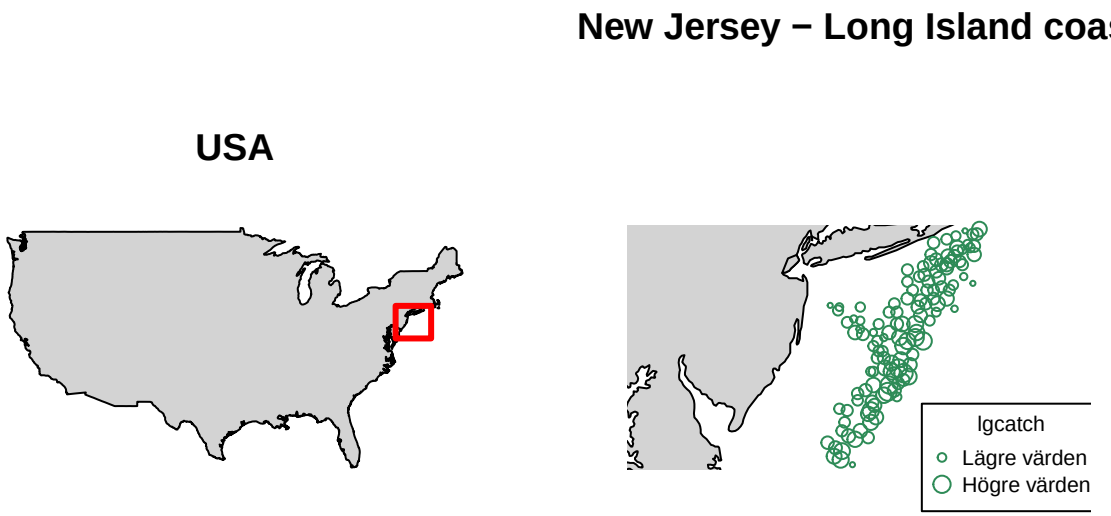


Figure 1: Presentation av datamaterialet (antal fångade musslor) i ett kartdiagram.

Nu presenteras datamaterialet åter igen (figur 2), men den här gången tydligare. Kartdiagrammet visar att var musslorna har blivit observerade längst kusten i New Jersey. Punktstorleken representerar den log-transformerade variabeln **lgcatch**, där större punkter indikerar på ett större antal musslor som fångats, och

mindre ringar mindre observationer. Sammanfattningsvis visar diagrammet att fångsterna avb musslorna varierar inom det studerade området.

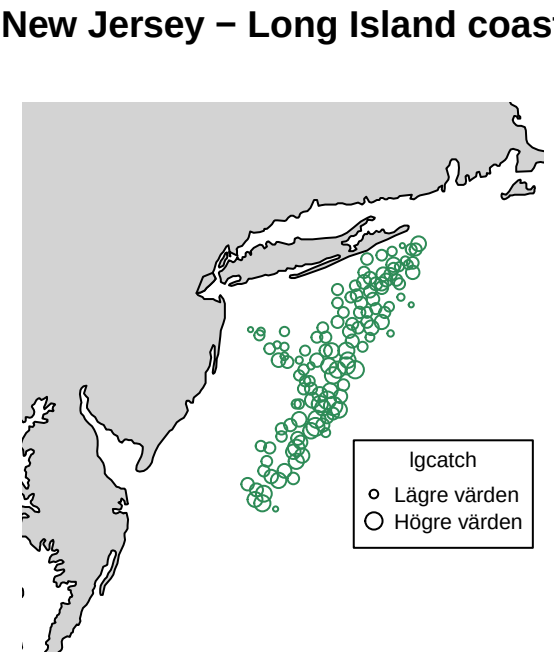


Figure 2: Presentation av datamaterialet (antal fångade musslor) i ett kartdiagram.

Bifogad R kod

```r
################################################################################
# Börjar med att visualisera datamaterialet i ett kartdiagram
################################################################################

par(mfrow = c(1, 2))

# Börjar med att visualisera landet
map("usa", fill = TRUE, col = "lightgray")

# Lägger till titel i bilden
title("USA")

# Lägger till en röd rektangel som representerar
# var observationerna är observerade.
rect(
  -75.5, 38.2, -71, 41.5,
  density = NULL, angle = 45,
  col = NULL, lwd = 3, border = "red"
)

# Lägger till en yttligare karta men nu med delstaten
# tillsammans observationerna
map(
  # Väljer land
  "usa",
  # Koordinaterna för delstaten
  xlim = c(-76.2, -70.4),
  ylim = c(37.4, 42.1),
  # fill = TRUE måste finnas för att visa grafen
  fill = TRUE,
  # Gör bakgrundsfärgen till grå i bilden
  col = "lightgray"
)

# Lägger till titel
title("New Jersey - Long Island coast")

points(
  # Koordinaterna för observationerna
  data$long,
  data$lat,
  # Ändrar färg på punkterna
  col = "seagreen",
  # Väljer punkt stil
  pch = 1,
  # Nu skalar jag variabelbn för att visa skillnad på de
  # obsevationer som är stora/små
  cex = sqrt(data$lgcatch) / max(data$lgcatch)*3.8
)

legend(
```

```
  x = -72.5, y = 39.2,
  legend = c("Lägre värden", "Högre värden"),
  pch = 1,
  pt.cex = c(0.6, 1.2),
  title = "lgcatch",
  bg = "white",
  cex = 0.7,
  col ="seagreen"
)
```

```
##############################################################################
# Andra presentation av datamaterialet pga storleksproblemet på observationerna
##############################################################################
par(mfrow = c(1,1))
map(
  # Väljer land
  "usa",
  # Koordinaterna för delstaten
  xlim = c(-76.2, -70.4),
  ylim = c(37.4, 42.1),
  # fill = TRUE måste finnas för att visa grafen
  fill = TRUE,
  # Gör bakgrundsfärgen till grå i bilden
  col = "lightgray"
)

# Lägger till titel
title("New Jersey - Long Island coast")

points(
  # Koordinaterna för observationerna
  data$long,
  data$lat,
  # Ändrar färg på punkterna
  col = "seagreen",
  # Väljer punkt stil
  pch = 1,
  # Nu skalar jag variabelbn för att visa skillnad på de
  # obsevationer som är stora/små
  cex = sqrt(data$lgcatch) / max(data$lgcatch)*3.8
)

legend(
  x = -72.5, y = 39.2,
  legend = c("Lägre värden", "Högre värden"),
  pch = 1,
  pt.cex = c(0.6, 1.2),
  title = "lgcatch",
  bg = "white",
  cex = 0.7
)
```

Deluppgift 1b

Teori om interpolation (från föreläsningen)

I spatial data när man vill studera ett stort område är det omöjligt mäta hela området, på grund av kostnader och tidsbrist. En lösning till detta är att mäta ett område och genom interpolation uppskatta observationerna som inte finns. Ett alternativ är interpolation med det inversa avståndet.

Man samlar in observationer i ett rum, $Y(s_1), ...Y(s_n) \in D$. Sedan vill man uppskatta en ny observation, ($\hat{Y}(s_0)$ som tillhör rummet $s_0 \in D$ genom att utnyttja observationer i närområdet. Följande formel används för den viktade interpolationen:

$$\hat{Y}(s_0) = \frac{\sum_{i=1}^{n} \omega(s_i, s_0) \cdot Y(s_i)}{\sum_{i=1}^{n} \omega(s_i, s_0)} \quad \text{Där} \quad \omega(s_i, s_0) = ||s_i - s_0||^{-p}$$

Här är $\hat{Y}(s_0)$ den nya interpolerade observationen, $Y(s_i)$ är de observerade värdena, och $\omega(s_i, s_0)$ är vikten som ges av $||s_i - s_0||^{-p}$. Enkelt förklarat säger vikten hur stor påverkan en det observerade värdet har på skattningen i den nya observationen, där p kan väljas för om man vill att närmare observationer ska ha större påverkan än de längre bort. Om $\omega(s_i, s_0)$ blir stor innebär det att observationerna är nära den nya observationen, och konsekvent har vikten en större påverkan. Om $\omega(s_i, s_0)$ är litet är det observerade värdet längre bort från den nya observationen, och påverkan blir mindre (Alenlöv, s. 15, 2025)..

```
# Under är Johan Alenlövs kod för att skapa grid
llCRS <- CRS("+proj=longlat +ellps=WGS84") # WGS84 is the World Geodesic System from 1984
nGridPoints <- 100 # Decrease this if the computations takes too long.
dataGrid <- expand.grid(x = seq(-75.5, -71, length = nGridPoints),
                        y = seq(38.2, 41.5, length = nGridPoints))
coordinates(dataGrid) <- c("x", "y")
dataGrid <- SpatialPixels(dataGrid, proj4string = llCRS)
data_mat <- cbind(data$long, data$lat)
spdf <- SpatialPointsDataFrame(data_mat, data, proj4string = llCRS, match.ID = TRUE)

# Nu skapar jag IDW
par(mfrow = c(1,3))
int1 <- idw(lgcatch ~ 1, spdf, dataGrid, idp = 1)
```

```
## [inverse distance weighted interpolation]
```

```
int2 <- idw(lgcatch ~ 1, spdf, dataGrid, idp = 2)
```

```
## [inverse distance weighted interpolation]
```

```
int3 <- idw(lgcatch ~ 1, spdf, dataGrid, idp = 4)
```

```
## [inverse distance weighted interpolation]
```

```
int4 <- idw(lgcatch ~ 1, spdf, dataGrid, idp = 7)
```

```
## [inverse distance weighted interpolation]
```

```
int5 <- idw(lgcatch ~ 1, spdf, dataGrid, idp = 10)
```

```
## [inverse distance weighted interpolation]
```

```
int6 <- idw(lgcatch ~ 1, spdf, dataGrid, idp = 15)
```

```
## [inverse distance weighted interpolation]
```

Följande figurer (3 och 4) visar reslutaten av interpolationen med det viktade inversa avståndet för samma datamaterial, där p varierar mellan 1 och 15. Parametern p styr hur stort vikt de närliggande observationerna har vid uppskattningarna, där större värden innebär att närliggande observationer har en större inflytning. Interpolationen visualiserar med hjälp av färger, där mörkare färger (röd/organge) indikterar lägre värden på **lgcatch**, medan ljusare färger (gul) indikerar högre värden. Det är värt att notera att om $p = 0$, har inte närliggande observationer någon betydelse alls. Detta kommer reslutera i att alla uppskattningar förväntas ha höga värden.

Det första diagrammet till vänster (figur 3), när $p = 1$, visar att den interpolerade ytan blivit mycket slät och värdena förväntas vara högra i nästan hela området. Det är svårt att urskilja någon stor variation i mellan mätpunkterna. I diagrammen där $p = 2$ och $p = 4$ framträder ett tydligare mönster kring det observerade området. Nu början uppskattningarna som är nära mätpunkterna få större betydelse vilket visar att värdena i det observerade området är högre medan, värdena utanför det observerade områder är lägre. Resultatet är rimligt eftersom det finns få observationer utanför det observerade området .

I de tre sista diagrammen (figur 4) ändras nu p till 7, 10, och 15. Man säger nu åt algortimen att närliggande observationer är mycket viktigare än de långt ifrån. Detta resulterar i att interpolationen blir tydligt dominerad i området där observationer finns, medan områden utan observationer får betydligt lägrde värden.

Sammanfattningsvis kan man observera att ju mer p ökar, desto större vikt läggs på de närliggande observatioerna. Detta leder till att interpolationen blir mer lokal och dominerad av områden med många observationer, medan utanför blir slätare och får lägre uppskattningar. Med det sagt verkar den bästa p för detta datamaterial vara $p = 2$.

```
# Ändrar bildstorleken i R
par(mfrow=c(1,3))

# Presenterar interpolationen mha image()
image(int1,
  main = "IDW: p = 1")

# Sedan lägger jag till kartan för tydligare visualisering
map(
  # Säger till att det är USA kartan jag vill ha
  "usa",
  # Lägger rätt koordinater
  xlim = c(-76.2, -70.4),
  ylim = c(37.4, 42.1),
  # Fill behöver vara T annars visas inte plotten
  fill = TRUE,
  # Färgen på landskapet.
  col = "lightgray",
  # add = T säger att jag vill lägga till i min tidigare plott
  add = TRUE)
```

```r
image(int2,
  main = "IDW: p = 2")

map(
  "usa",
   xlim = c(-76.2, -70.4),
   ylim = c(37.4, 42.1),
   fill = TRUE,
   col = "lightgray",
   add = TRUE)

image(int3,
  main = "IDW: p = 4")

map(
  "usa",
   xlim = c(-76.2, -70.4),
   ylim = c(37.4, 42.1),
   fill = TRUE,
   col = "lightgray",
   add = TRUE)
```

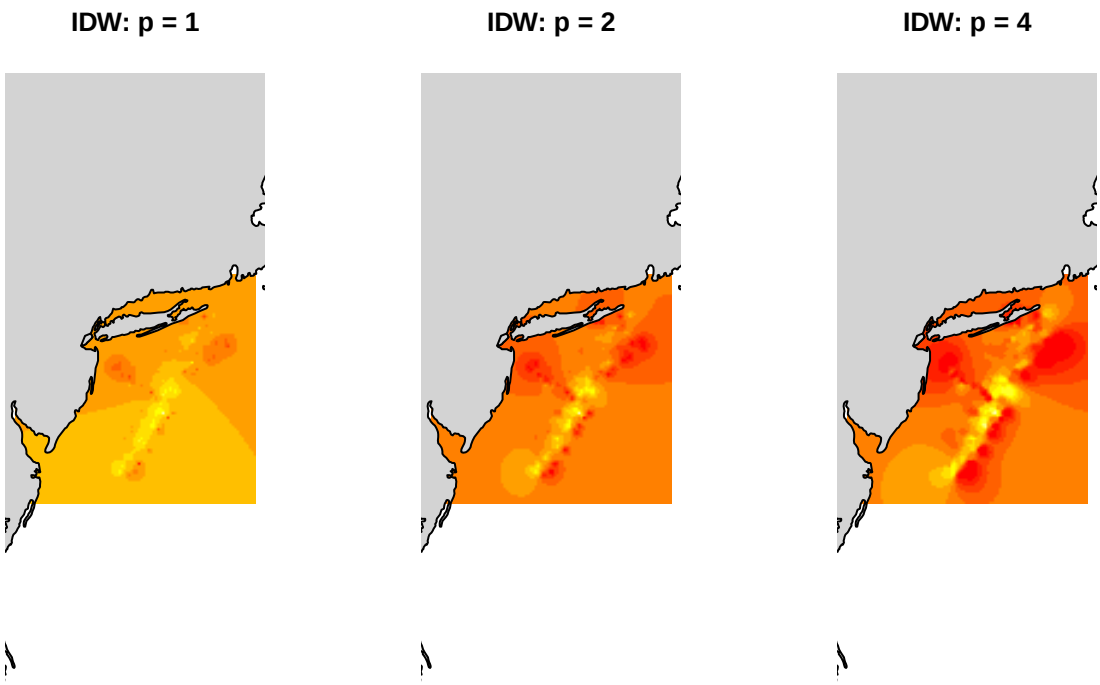
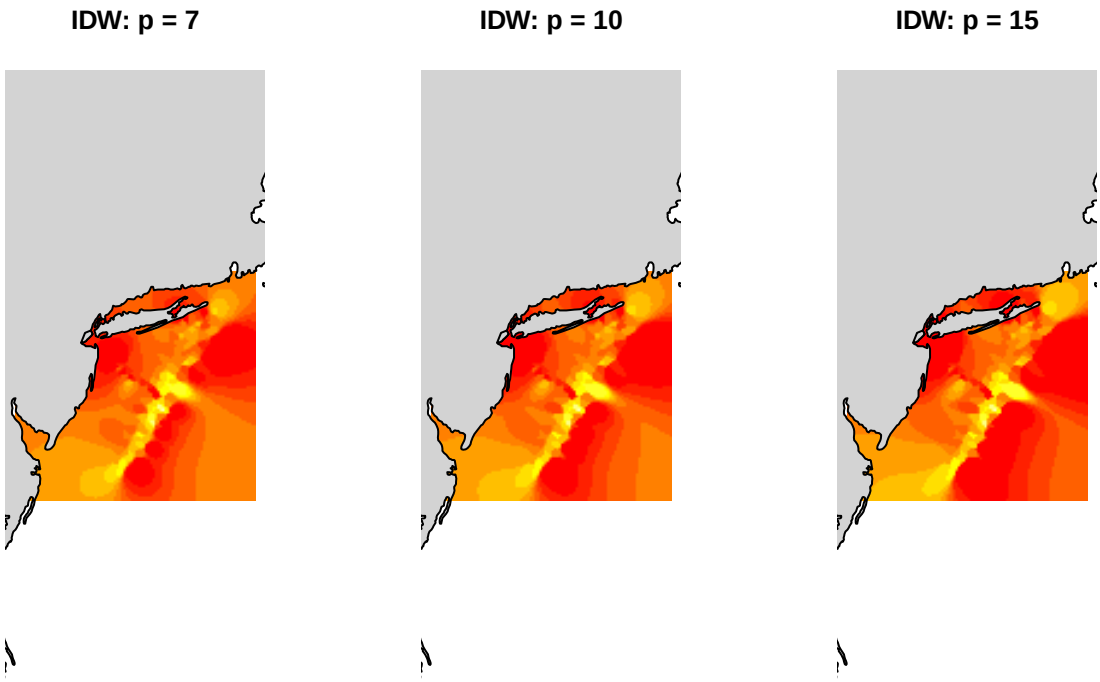


Figure 3: Interpolation

```r
par(mfrow=c(1,3))
image(int4,
  main = "IDW: p = 7")

map(
  "usa",
   xlim = c(-76.2, -70.4),
   ylim = c(37.4, 42.1),
   fill = TRUE,
   col = "lightgray",
   add = TRUE)

image(int5,
  main = "IDW: p = 10")

map(
  "usa",
   xlim = c(-76.2, -70.4),
   ylim = c(37.4, 42.1),
   fill = TRUE,
   col = "lightgray",
   add = TRUE)

image(int6,
  main = "IDW: p = 15")

  map(
  "usa",
   xlim = c(-76.2, -70.4),
   ylim = c(37.4, 42.1),
   fill = TRUE,
   col = "lightgray",
   add = TRUE)
```



Deluppgift 1c)

Beräkna samplevariogrammet under antagandet om isotropisk korrelation med hjälp av momentmetoden.

I denna deluppgift (1c) beräknades samplevariogrammet med hjälp av momentmetoden under antagandet om isotropisk korrelaiton. När man använder momentskattningen för att uppskatta semivariogrammet används följande formel (Alenlöv, s.18, 2025):

$$\hat{\gamma}(I_j) = \frac{1}{2|N(I_j)|} \sum (Y(s_i) - Y(s_i))^2$$

Antgandet om isotropisk korrelation innebär att variogrammet bara beror på avståndet I_j, alltså att två punkter, oavsett avstånd, ska vara lika mycket korrelerade oavsett vart de är placerade (s. 17).

I figur 4 presenteras samplevariogrammet. Diagrammet vosar em skattning av det teoretiska variogrammet, där avstånd visas på x-axeln och semivarians på y. Man kan notera en tydligt ökande semivarians för små avstånd. Det vill säga, semivariansen ökar kraftigt vid ett tidigt stadie. Därefter noteras det att semivariansen stabiliserad vid avstånd 60-80 (detta motsvarar range), där semivariansen ligger runt 5 enheter; detta är sillen.

I figur 5 presenteras molnet för variogrammet, alltså parvisa semivarianser som funktion av avstånd. På x-axeln syns avstånd och semivarians på y-axeln. Man kan notera att spridningen är låg när semivariansen är låg, men hög när semivariansen är hög. Man kan konstantera att spridningen ökar med avståndet. På grund av att molnet är mycket "brusigt" blir diagrammet svårtolkat och därför är det viktigare att studera samplevariogrammet.

```r
# Skattar mitt samplevariogram
plot(
  # Under skattas modellen
  variogram(lgcatch ~ 1, spdf),
  # Definerar ett linjediagram
  type = "b")
```

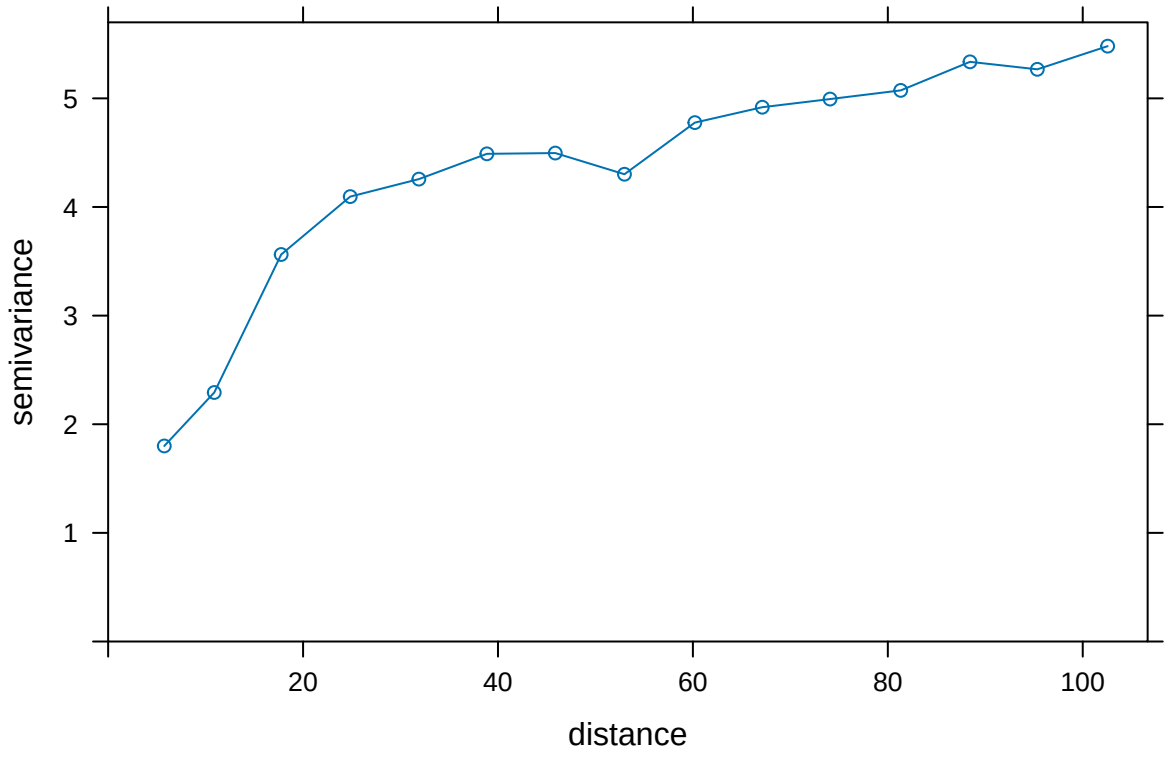


Figure 4: Skattning av samplevariogrammet med hjälp av momentskattning under antagandet om isotropisk korrelation.

```r
# Skattar samplevariogram där alla paren syns
plot(variogram(lgcatch ~ 1, spdf, cloud = T))
```

Deluppgift 1d)

Undersök om det spatiala beroendet ser likadant ut i olika riktningar

För att besvara om det spatiala beroendet är likadant i olika ritkning behöver man studera flera diagram i olika riktningar. Detta typ av diagram används för att undersöka istropi eller anisotropi. Det vill säga om beroendet är samma i alla riktningar eller olika (Alenlöv, s.22, 2025).

I figur 6 presenteras samplevariogrammet med riktningarna $\alpha = 0, 45, 90, 135$. Det man vill se är att samplevariogrammen är lika oavsett den riktning de pekar i. Man kan snabbt notera att när $\alpha = 45$ noteras en mycket lägre semivarians över avstånd än för resterande diagram. För $\alpha = 90$ och $\alpha = 0$ är dessa diagram

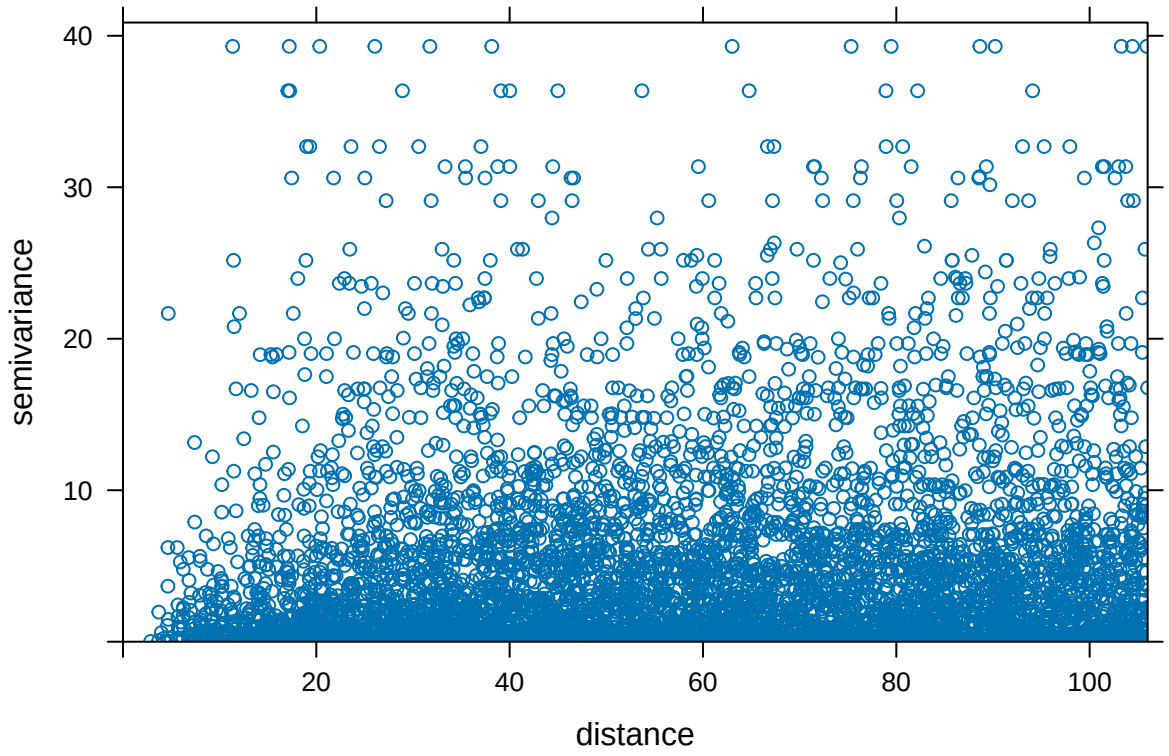


Figure 5: Samplevariogrammet med alla par.

relativt lika varandra. Men när $\alpha = 135$ ökar semivariansen drastiskt från avståndet 60-100. Slutsatsen kan dras att samplevariogrammen inte är lika i olika riktningar. Detta uppvisar anisotropi.

```
# Nu undersöks om det spatiala beroende i alla riktningar
# är densamma
plot(
  # Skattar samplevariogrammet
  variogram(
    # Skattar först variogrammet
    lgcatch ~ 1, spdf,
    # Sedan väljer jag riktningarna i grader
    alpha = c(0,45,90,135
              )
    )
  )
```

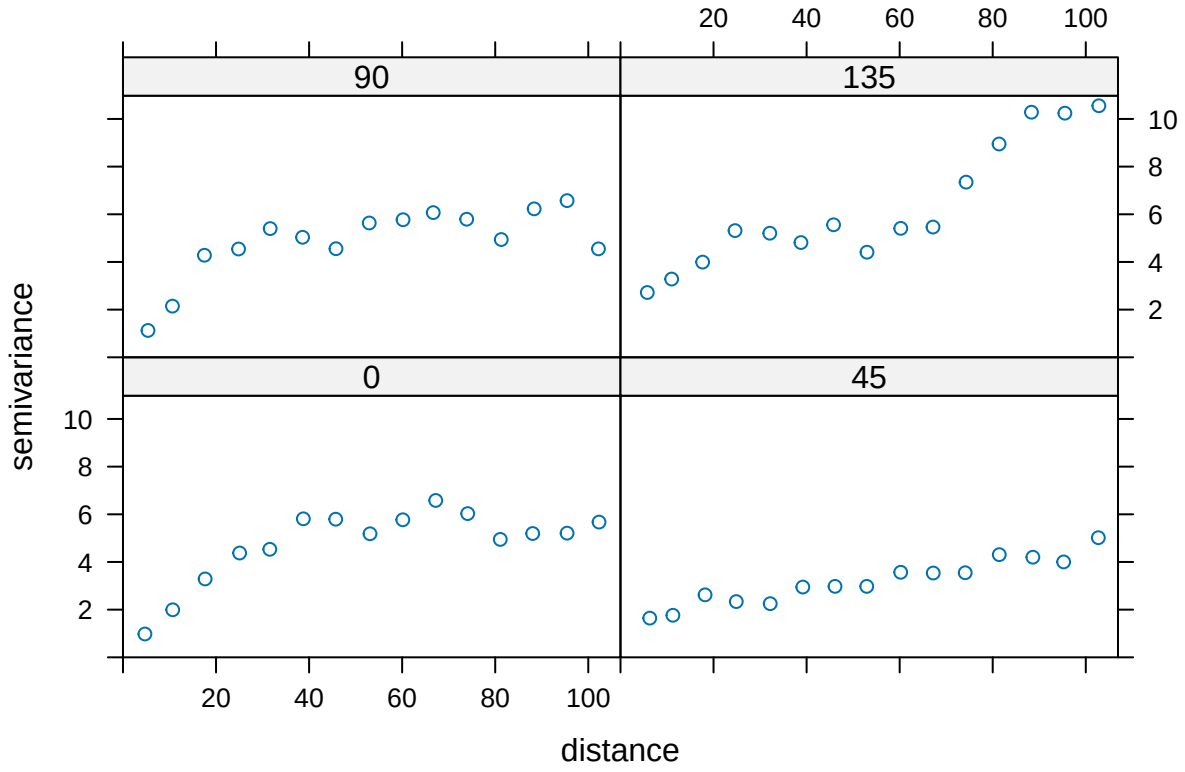


Figure 6: Samplevariogrammet i olika riktningar, med alpha = 0, 45, 90, 135 grader.

Deluppgift 1e)

Anpassa minst fyra olika isotropiska variogrammodeller till datamaterialet. Tolka resultaten. Visa vilken modell passar bäst till data?

Nu anpassas fyra isotropiska variogrammodeller. De modeller jag väljer är linjär, sfärisk, powered exponential, och Matérn. För att skatta dessa modeller kräver funktionen vgm() att man anger startvärden på

partiell sill, range, och nugget. I detta fall är nugget värdet i variogrammet (figur 4) där avståndet är som minst vilket är drygt 1.8. Sill är det värde (semivarians) där variogrammet planar ut, vilket är drygt 5.2. Range är värdet på y-axeln då variogrammet planat ut, och det ser ut att vara drygt 70. Detta innebär att följande värden används för modellerna:

- Nugget: 1.8.
- Sill: 5.2.
- Range: 70.
- Partiell Sill = Sill - Nugget = 5.2-1.8=3.4.

```
# Nu skattar jag de fyra olika modellerna

    variogram <- variogram(lgcatch ~ 1, spdf)
  # Linjär modell
    linjär <- vgm(model = "Lin", psill = 3.4,range = 70,nugget = 1.8)
    linjär_variogram <- fit.variogram(variogram,linjär)
  # Sfärsisk
    sfärisk <- vgm(model = "Sph", psill = 3.4,range = 70,nugget = 1.8)
    sfärisk_variogram <- fit.variogram(variogram,sfärisk)
  # Powered Exponential
    p_exponential <- vgm(model = "Exp", psill = 3.4,range = 70,nugget = 1.8)
    p_exponential_variogram <- fit.variogram(variogram,p_exponential)
  # Matern
    matern <- vgm(model = "Mat", psill = 3.4,range = 70,nugget = 1.8)
    matern_variogram <- fit.variogram(variogram,matern)
```

I figur 7 presenteras samplevariogrammet med den anpassade linjära modellen. Man kan notera att modellen ökar linjärt fram till avstånd = 25, men är därefter konstant, och saknar sill. Det vill säga att den linjära modellen lyckas inte beskriva variogrammets form, varken vid stora eller små avstånd, och är därför inte lämplig.

```
# Sedan kan man presentera plottarna
plot(variogram, linjär_variogram, main = "Linjär modell")
```

```
kable(linjär_variogram, caption ="Outputs för den linjära modellen.")
```

Table 2: Outputs för den linjära modellen.

model	psill	range	kappa	ang1	ang2	ang3	anis1	anis2
Nug	0.9943411	0.00000	0.0	0	0	0	1	1
Lin	3.5808144	27.37524	0.5	0	0	0	1	1

Variogrammet med den sfäriska modellen presenteras i figur 8 och tabell 3. Denna modell har en nugget på drygt 0.84 och sill = $0.84 + 3.807 \approx 4.65$, och en range = 37.5. Sillen berättar att när avståndet överstiger räckvidden är semivariansen drygt 4.65. När $h > 37.5$ sker ingen mer ökning i semivarians (enligt modellen), alltså att observationerna inom 37.5 enheter är rumsligt korrelerade. Utifrån diagrammet (figur 7) ser man att man lcykas fång den snabba ökning vid korta avstånd, men det planar av tidigare än de sista punkterna i diagrammet.

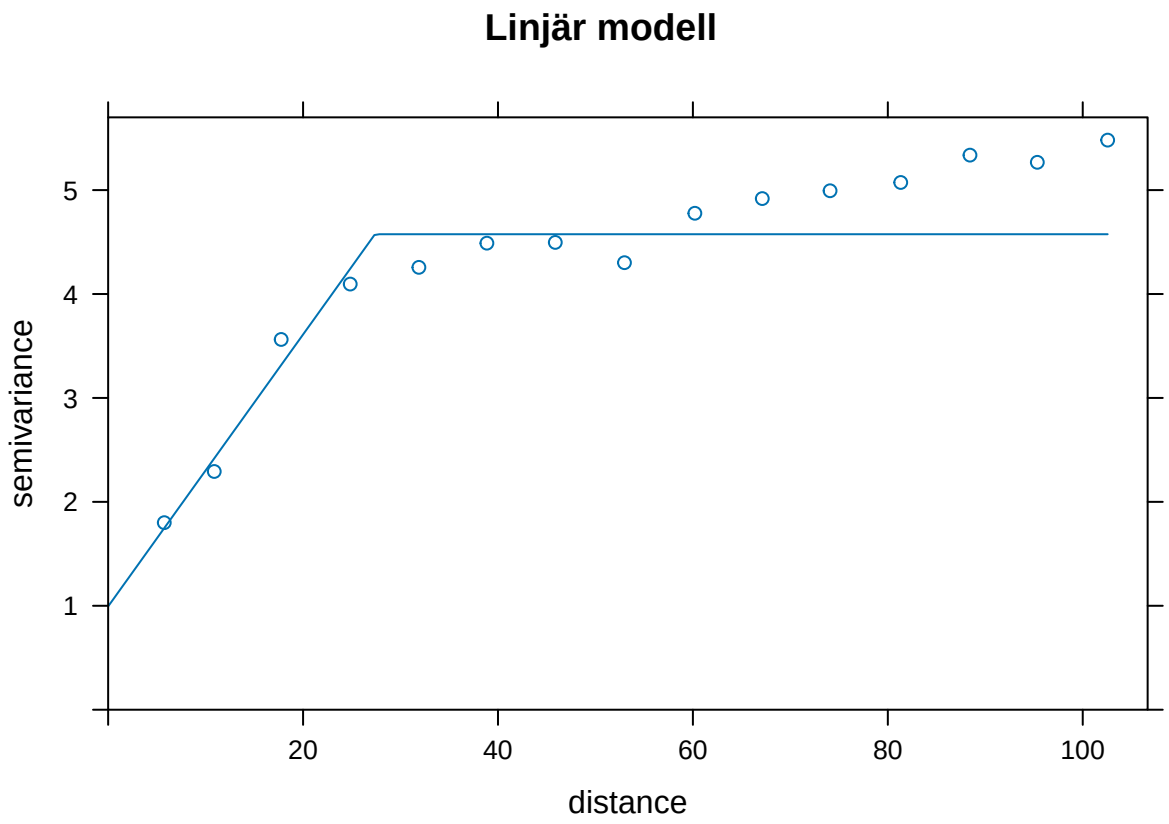


Figure 7: Samplevariogram för den linjära modellen

```
plot(variogram, sfärisk_variogram, main = "Sfärisk modell")
```

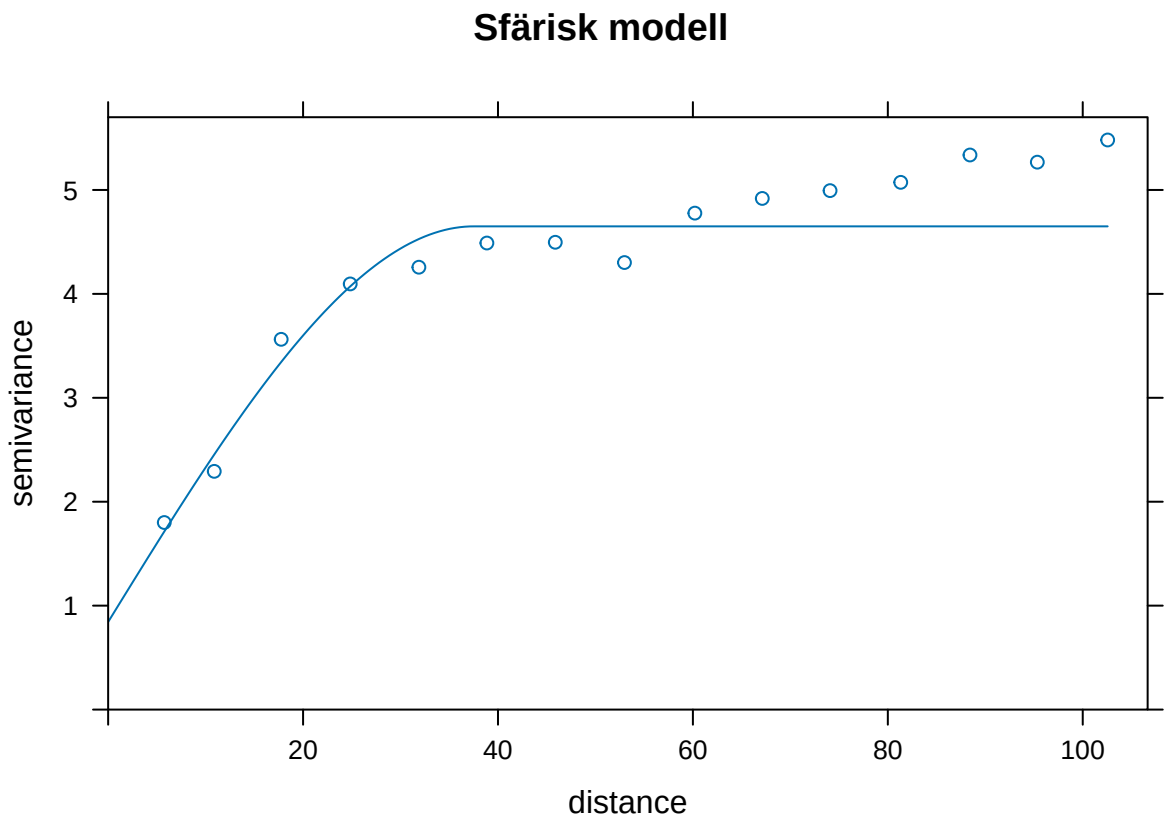


Figure 8: Samplevariogram för den sfäriska modellen.

```
kable(sfärisk_variogram, caption ="Outputs för den sfäriska modellen.")
```

Table 3: Outputs för den sfäriska modellen.

model	psill	range	kappa	ang1	ang2	ang3	anis1	anis2
Nug	0.843171	0.00000	0.0	0	0	0	1	1
Sph	3.806523	37.50432	0.5	0	0	0	1	1

I figur 9 och tabell 4 presenteras den exponentiella modellen. Denna visar att nugget är 0.36, partial sill = 4.67, sill är 5.02, och range är 17.38. Detta innebär att det finns en liten variation eller mätfel, men den största variationen i datan har en rumslig struktur (se partial sill). Range = 17.38 visar att det spatiala beroendet avtar mycket snabbt, jämfört med resterande modeller. Men hur som helst försvinner det inte helt. I diagrammet (figur 9) kan man notera specifikt att nugget är mycket liten eftersom den är relativt nära 0. Beroendet avtar snabbt med avståndet och efter 17.4 enheter är observationerna nästan helt oberoende, men ej helt.

```
plot(variogram, p_exponential_variogram, main = "Powered exponential")
```

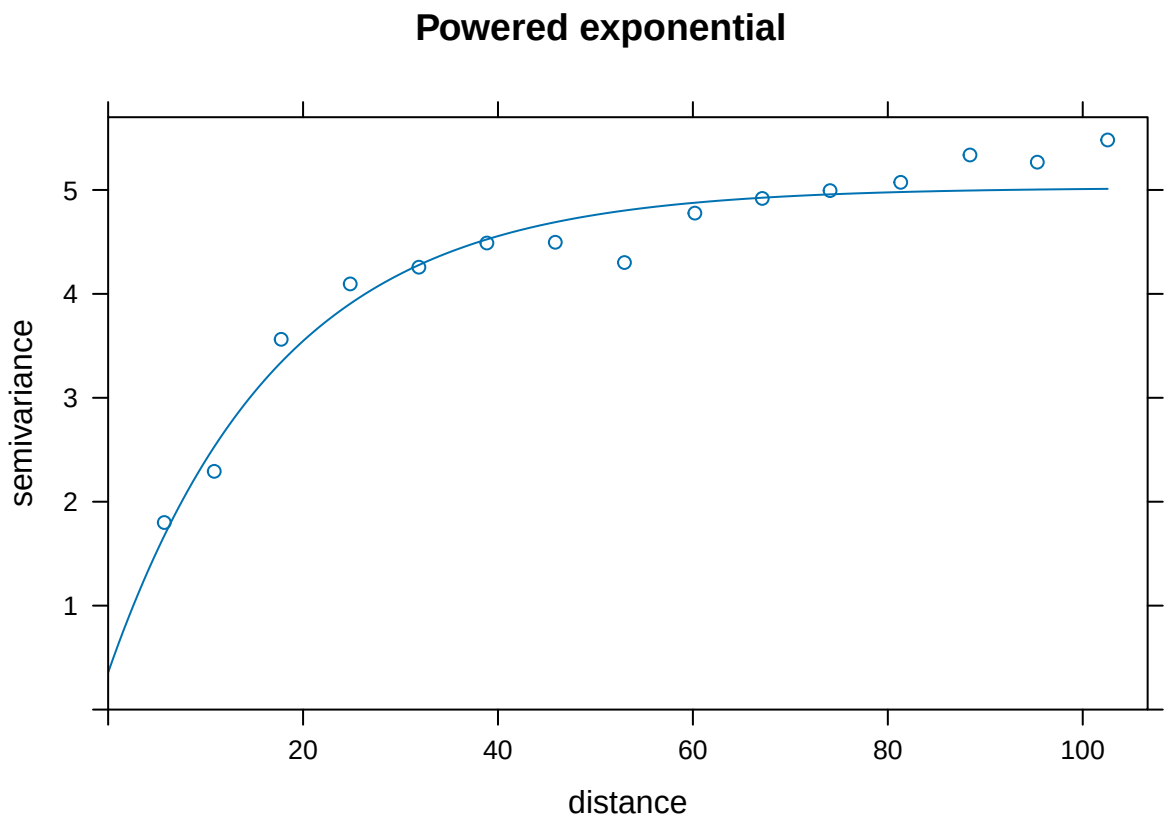


Figure 9: Samplevariogram för den exponentiala modellen.

```
kable(p_exponential_variogram, caption ="Outputs för den powered exponentiala modellen.")
```

Table 4: Outputs för den powered exponentiala modellen.

model	psill	range	kappa	ang1	ang2	ang3	anis1	anis2
Nug	0.3570203	0.00000	0.0	0	0	0	1	1
Exp	4.6662492	17.38368	0.5	0	0	0	1	1

För Matérn modeller (figur 10) noteras det att modellen, i princip, är exakt samma som exponential modellen. Detta ser man även i kappa värdet som är 0.5. Därför är det inte viktigt att beskriva Matern modellen eftersom den är nästan exakt likadant som modellen som beskrev tidigare.

```
plot(variogram, matern_variogram, main = "Matérn modell")
```

```
kable(matern_variogram, caption ="Outputs för Matern modellen.")
```

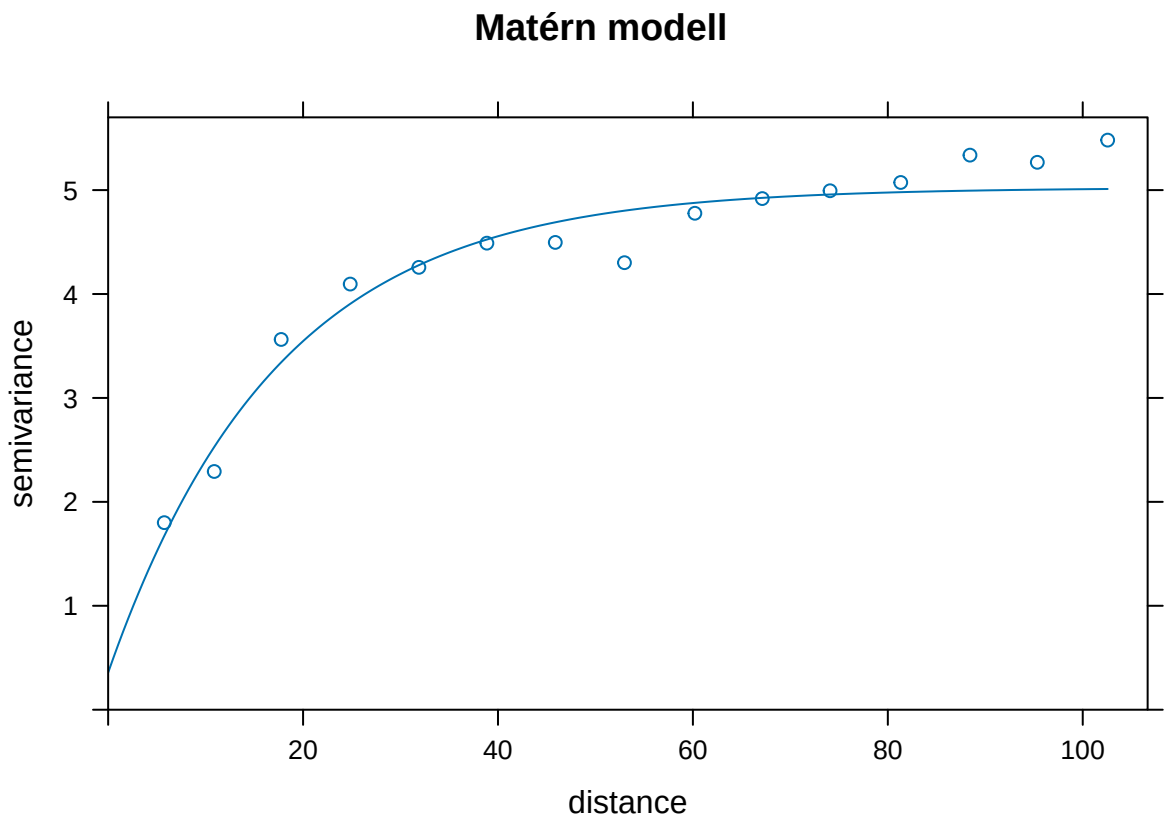


Figure 10: Samplevariogram för Matern modellen.

Table 5: Outputs för Matern modellen.

model	psill	range	kappa	ang1	ang2	ang3	anis1	anis2
Nug	0.3570204	0.00000	0.0	0	0	0	1	1
Mat	4.6662492	17.38368	0.5	0	0	0	1	1

Sammanfattningsvis ser man att den bästa modellen av de fyra är den sfäriska modellen. Detta är eftersom den ger den mest realitiska beskrivningen av beroendet i datamaterialet. Man märkte i den linjära modellen att sill, i princip, saknades. Den exponentiella modellen och Matérn verkar inte beskriva verkligheten på ett korrekt sätt. Detta är eftersom beroendet avtar gradvis men det upphör aldrig, och detta sätmmet inte överrens med det datamaterial som jag jobbat med. Därför är den bästa modellen utav dessa fyra den sfäriska.

Deluppgift 1f)

Används kriging för att interpolera lgcatch över samma gridsom i b). Hur skiljer sig reslutaten? Diskutera.

I denna deluppgift har jag använt den sfäriska modellen att genomföra kriging med. Reslutatet visas i figur 11. Åter igen tolkas interpoleringen genom att röd/orange färger är uppskattningar vars värden är lägre, medan gula områden är högre värden. Interpoleringen (figur 11) visar att uppskattningarna är högre i områden där många observationer är nära varandra, medan observationer som är mer avlägsna från varandra ger uppskattningar vars värden är mindre. Skillnaden mellan detta utförande (kriging) mot deluppgift b där man använde IDW, är att i kriging baseras vikterna på den skattade rumsliga korrelationen. Detta ger en någorlunda bättre interpolation. Ser man en skillnad? Ja, man ser en stor skillnad i interpolationen mellan IDW och kriging. Interpolationen verkar mer tydligt i områdena där observatioerna finns, men samtidigt utanför. Speciellt kan man notera detta i spetsen av punkterna (längs vad som ser ut som en horisontell linje av observationer), där man kan notera att IDW uppskattade att fångsten av musslor är många nära kusten, medan kringen är mer centrerad kring observationerna.

```
# Nu använder jag kriging för att genomföra interpolering
kriging <- krige(lgcatch ~1,
                 spdf,
                 dataGrid,
                 # Använder sfärsiska modellen
                 model = sfärisk_variogram,
                 )
```

```
## [using ordinary kriging]
```

```
# Nu är det dax att visualisera reslutaten


image(kriging, main ="Kriging med den sfärsiska modellen")
map(
  "usa",
  xlim = c(-76.2, -70.4),
  ylim = c(37.4, 42.1),
  fill = TRUE,
  col = "lightgray",
  add = TRUE)
# Lägger till obs
points(spdf, pch = 1, cex = 0.2)
```

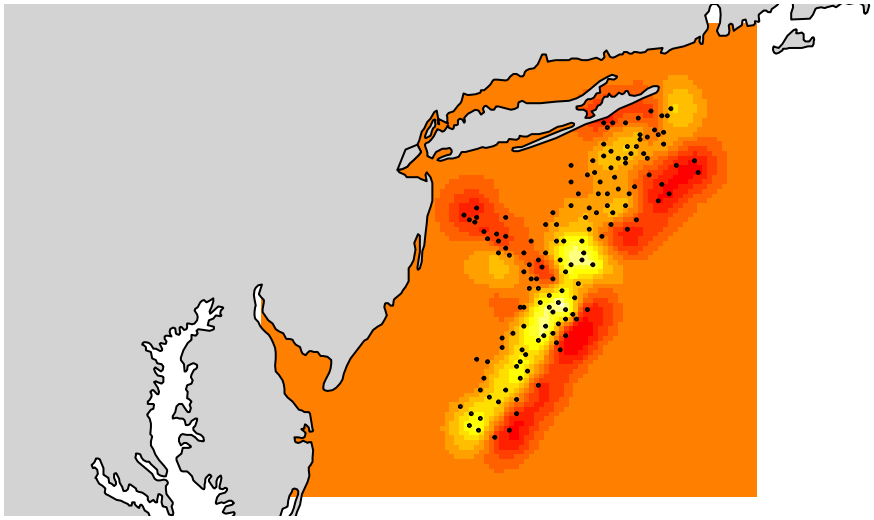


Figure 11: Interpolering med kriging för den sfäriska modellen.

Uppgift 2

I denna uppgift ska man analysera datamaterialet boston housing som innehåller medianförsäljningspriset av bostäder i 506 områden kring Boston.

Deluppgift 2a)

Visualisera datamaterialet över en karta för att informellt undersöka om det förekommer något spatialt beroende. Du behöver bara kolla på CMEDV här och inga kovariater.

I denna deluppgift används återigen samma färgkombination som i tidigare uppgifter. I figur 12 visas Boston Housing datamaterialet, där låga värden är gula/ljusare och högre värden är organgea/mörkare. Figuren visar att områden med höga värden tenderar att ligga nära andra områden med höga värden, och motsvarande kan ses för låga värden. Detta tyder på det finns spatialt beroende i datamaterialet, eftersom att likande värden bildar geografiska kluster.

```
# I denna deluppgift användерj jag Johan Alenlövs kod
# för att få grundstrukturen i datan. Sedan skapade
# jag själv visualiseringen, där jag utgick från hans kod.

boston <- read.table("/Users/hampusbeijer/Downloads/BostonHousing.txt", header = TRUE)
boston_coord <- cbind(boston$LON, boston$LAT) # Constructing a matrix with longitude and latitude as
llCRS <- CRS("+proj=longlat +ellps=WGS84") # WGS84 is the World Geodesic System from 1984
# Making a spatial data frame from a data frame and a matrix of coordinates
bostonPoints <- SpatialPointsDataFrame(boston_coord, boston, proj4string = llCRS, match.ID = TRUE)
# Create a neighborhood object based on Gabriel neighborhoods
boston_nb <- graph2nb(relativeneigh(boston_coord), row.names = row.names(boston), sym = TRUE)

# Nu under kommen min redigerade kod där jag utgår
# från Johans ursprungliga kod. Jag ändrar inte mycket
# men endast lite simpelt med visualiseringen

# Plottar upp kartan
map("county", region = "massachusetts",
    xlim = c(-71.3, -70.8),
    ylim = c(42, 42.4),
    fill = TRUE,
    col = "lightgray",
    bg = "white",
         main = "Datamaterial över Boston (CMEDV)")

# Sedan placerar jag ut observationerna
# men jag är intresserad av att se hur dessa
# relaterar till varandra därför delar jag in i
# intervall för att visa observationer som är snarlika
# alltså kan associeras som grannar.

färger <- heat.colors(100) # väljer att färga mina observationer.
färg <- färger[cut(boston$CMEDV, breaks = 100)]
plot(boston_nb, coords = boston_coord, add = TRUE, col = "black")
points(boston_coord, pch = 1, col = färg, cex = 0.6)
```

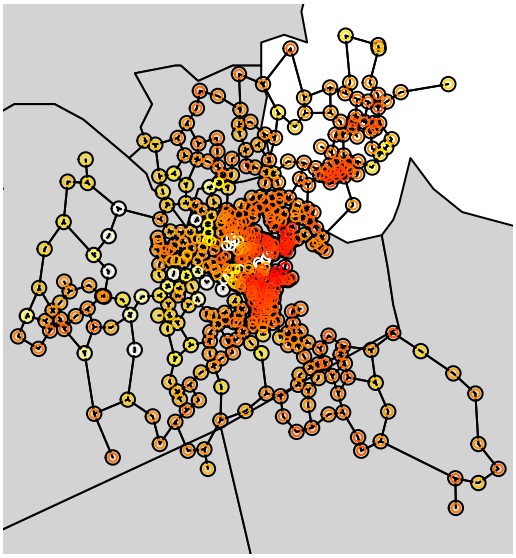


Figure 12: Kartdiagram över Boston datamaterialet

Deluppgift 2b)

Testa om residualerna från den linjära modellen ovan är spatialt oberoende genom Moran's I.

Morans test undersöker om residualerna visar spatial autokorrelation. Följande hypoteser ställs upp:

$$H_0 : \text{Residualerna är spatialt beroende.} H_a : \text{Residualerna är inte spatialt beroende.}$$

I tabell 6 presenteras statistikan med tillhörande p-värde för Morans test. Statistikan är drygt 14.2 med ett p-värde som är mindre än 0.05. Detta innebär att med 5% signifikans kan nollhypotesen förkastas, och det konstanteras att residualerna är inte spatialt beroende. Modellen verkar alltså lämplig att gå vidare med.

```
# Följande kod för linjär modell är
# bifogad från Johan Alenlöv.

lmModel <- lm(log(CMEDV) ~ CRIM +
ZN + INDUS + CHAS + I(NOX^2) + I(RM^2) + AGE + DIS + RAD + TAX
+ PTRATIO + B + LAT + LON, data = boston )

# Extraherar residualerna mha residuals
residualer <- residuals(lmModel)

# Behöver viktmatrisen för detta så skapar den mha
# paketet spdep
W <- nb2listw(boston_nb, style ="W")

# Genomför nu Morrans I test
morre <- moran.test(residualer, W)

# Skapar tabell för presentering

tab <- data.frame(
  Statistiska = morre$statistic,
  pvalue = morre$p.value
)

kable(tab, caption ="Moran's I test på modellens residualer")
```

Table 6: Moran's I test på modellens residualer

	Statistiska	pvalue
Moran I statistic standard deviate	14.21471	0

Deluppgift 2c)

Anpassa en SAR modell och tolka resultaten. Finns det spatialt beroende i SAR modellen?

Modellen skattas nu med funktionen spautolm(). För att underöska om det finns spatialt beroende i modellen kollar man på λ statistikan. $\lambda = 0.71$ med tillhörande p-värde som är mindre än 0.05. Detta innebär att det finns spatialt beroende i SAR modellen. Rent teoretiskt innebär detta att responsvariabeln **CMEDV** påverkas av värden i det närliggande området.

Koefficienterna kan då tolkas som, t.ex. för *crim* ser man att denna är negativ och signifkant. Detta innebär att ju högre kriminalitet det är i området, desto lägre är bostadspriserna. RM^2 som visar antal rum, är positiv och signifkant. Detta innebär att desto större rummen är, desto högre är bostadspriset.

Sammanfattningsvis uppvisar den skattade SAR modellen ett signifkant spatialt beroende, där man kan konstantera att områdena i Boston påverkar varandra.

```
SAR <- spautolm(
  #Anger min grund modell
  log(CMEDV) ~ CRIM +
    ZN + INDUS + CHAS + I(NOX^2) + I(RM^2) + AGE + DIS + RAD + TAX
  + PTRATIO + B + LAT + LON,
  # Anger datamaterialet jag ska använda
  data = boston,
  # Anger viktmatris W
  listw = W,
  # Säger åt funktionen att jag vill använda SAR
  family ="SAR"
)
```

```
## Warning in sqrt(fdHess[1, 1]): NaNs produced
```

```
summary(SAR)
```

```
##
## Call: spautolm(formula = log(CMEDV) ~ CRIM + ZN + INDUS + CHAS + I(NOX^2) +
##     I(RM^2) + AGE + DIS + RAD + TAX + PTRATIO + B + LAT + LON,
##     data = boston, listw = W, family = "SAR")
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.7222291 -0.0663551 -0.0021348  0.0643053  0.7761908
##
## Coefficients:
##                Estimate  Std. Error z value  Pr(>|z|)
## (Intercept) -2.8656e+01  2.7927e+01 -1.0261 0.3048530
## CRIM        -9.5203e-03  1.0638e-03 -8.9496 < 2.2e-16
## ZN           1.6897e-03  5.9174e-04  2.8554 0.0042982
## INDUS        4.5235e-04  3.1442e-03  0.1439 0.8856040
## CHAS        -2.5762e-02  3.1132e-02 -0.8275 0.4079471
## I(NOX^2)    -6.0192e-01  1.6042e-01 -3.7522 0.0001753
## I(RM^2)      1.4860e-02  9.8949e-04 15.0181 < 2.2e-16
## AGE         -2.8093e-03  5.0761e-04 -5.5343 3.125e-08
## DIS         -3.3384e-02  1.2264e-02 -2.7220 0.0064882
## RAD          1.5919e-02  3.2917e-03  4.8360 1.325e-06
## TAX         -6.3605e-04  1.4801e-04 -4.2973 1.729e-05
## PTRATIO     -2.5343e-02  6.4155e-03 -3.9503 7.805e-05
## B            7.8695e-04  1.1934e-04  6.5940 4.281e-11
## LAT          3.3245e-01  3.6121e-01  0.9204 0.3573834
## LON         -2.5161e-01  3.0566e-01 -0.8232 0.4104114
##
## Lambda: 0.71252 LR test value: 270 p-value: < 2.22e-16
## Numerical Hessian standard error of lambda: NaN
```

```
##
## Log likelihood: 190.0231
## ML residual variance (sigma squared): 0.021829, (sigma: 0.14775)
## Number of observations: 506
## Number of parameters estimated: 17
## AIC: -346.05
```

Deluppgift 2d)

Anpassa en CAR modell och tolka resultaten. Finns det spatialt beroende i CAR modellen?

Nu anpassas istället en CAR modell, tidigare kod återanvänds men family-argumentet ändras från SAR till CAR. För att undersöka det spatiala berondet i CAR-modellen kollar man åter igen på λ. Statistikan är $\lambda = 0.61$ men tillhörande p-värde som är mindre än 0.05. Det finns alltså et signifkant spatialt beronde i VAR modellen, vilket innebär att responsvariabeln, *CMEDV*, påverkas av närliggande områden. I sin tur innebär det att man tolkar koefficienterna på följande sätt: CRIM är negativ och signfikant, vilket innebär att ju högre krminalitet det finns i området desto lägre är bostadspriserna. För RM^2 som är positiv och signifkant, vilket innebär att ju större bostaden är desto högre är bostadspriset.

Sammanfattningvis ser man att den skattade CAR modellen har ett signifkant spatialt beroende, där bostadspriset påverkas av områden runt omrking.

```
# Kopierar koden från uppgift 2c)
CAR <- spautolm(

  log(CMEDV) ~ CRIM +
    ZN + INDUS + CHAS + I(NOX^2) + I(RM^2) + AGE + DIS + RAD + TAX
  + PTRATIO + B + LAT + LON,

  data = boston,

  listw = W,
  # OBS: det enda som förändras är detta argument
  family ="CAR"
)
```

```
## Warning in spautolm(log(CMEDV) ~ CRIM + ZN + INDUS + CHAS + I(NOX^2) + I(RM^2)
## + : Non-symmetric spatial weights in CAR model
```

```
## Warning in sqrt(fdHess[1, 1]): NaNs produced
```

```
summary(CAR)
```

```
##
## Call: spautolm(formula = log(CMEDV) ~ CRIM + ZN + INDUS + CHAS + I(NOX^2) +
##     I(RM^2) + AGE + DIS + RAD + TAX + PTRATIO + B + LAT + LON,
##     data = boston, listw = W, family = "CAR")
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.673266 -0.078866 -0.012626  0.076251  0.832937
##
## Coefficients:
```

```
##                Estimate  Std. Error  z value  Pr(>|z|)
## (Intercept) -4.5161e+01  1.7869e+01  -2.5274 0.0114926
## CRIM        -1.3736e-02  1.3662e-03 -10.0538 < 2.2e-16
## ZN          -2.1487e-03  6.5408e-04  -3.2850 0.0010197
## INDUS       -7.9603e-04  3.1417e-03  -0.2534 0.7999776
## CHAS        -5.4378e-02  3.8659e-02  -1.4066 0.1595426
## I(NOX^2)    -6.2205e-01  1.5363e-01  -4.0489 5.145e-05
## I(RM^2)      1.6311e-02  1.2287e-03  13.2753 < 2.2e-16
## AGE         -3.9421e-05  5.9275e-04  -0.0665 0.9469753
## DIS         -1.4792e-02  1.0781e-02  -1.3721 0.1700458
## RAD          1.0625e-02  3.3103e-03   3.2096 0.0013294
## TAX         -5.9528e-04  1.7298e-04  -3.4413 0.0005790
## PTRATIO     -3.1086e-02  6.8115e-03  -4.5637 5.026e-06
## B            6.3041e-04  1.3101e-04   4.8120 1.494e-06
## LAT         -3.1332e-02  2.2539e-01  -0.1390 0.8894410
## LON         -6.9950e-01  1.9701e-01  -3.5507 0.0003843
##
## Lambda: 0.61092 LR test value: 112.8 p-value: < 2.22e-16
## Numerical Hessian standard error of lambda: NaN
##
## Log likelihood: 111.4191
## ML residual variance (sigma squared): 0.034749, (sigma: 0.18641)
## Number of observations: 506
## Number of parameters estimated: 17
## AIC: -188.84
```

Deluppgift 2e)

Undersök om resultaten från SAR och CAR modellerna förändras om man ändrar viktmatrisen. Så att vikten mellan områden är proportionellt mot det inversa avståndet. Använda nbdists och argumentet glist i funktionen nb2listw().

I denna deluppgift ska nu CAR och SAR modellerna skattas om, men viktmatrisen ska användas så att vikten mellan områden är proportionella mot det inversa avståndet. Detta görs på följande sätt:

```
# Nu undersöks det om reslutaten från tidigare modeller
# SAR och CAR ändras om man ändrar viktmatrisen.

# Jag börjar med att beräkna avståndet mellan grannarna
grann_avstånd <- nbdists(boston_nb,boston_coord)

# För att kunna skapa de inversa avstånds vikterna
# behöver man skapa en funktion som kan ändra om de.
# detta gör jag genom att använda lapply och lägg in
# grannavståndet tillsammans med en funktion

inversa_avståndet <- lapply(
  grann_avstånd,
  function(x)1/x
  )

# Nu när inversa avståndet har skapats kan man gå vidare
# till att skapa den nya viktmatrisen. Detta måste göras
# med funktionen nb2listw() där man i funktionen plaserar
# det inversa avståndet lika med glist!

W_2 <- nb2listw(
  # vanliga datan
  boston_nb,
  # Här anges inversa avståndet
  glist = inversa_avståndet,
  # Här anges att man vill ha den
  # vanliga viktmatrisen som är rad
  # -standardiserad.
  style = "W"
)
```

Nu när den nya viktmatrisen W_2 är skapad kan de nya modellerna skattas om.

```
#SAR
SAR2 <- spautolm(
  log(CMEDV) ~ CRIM + ZN + INDUS + CHAS + I(NOX^2) + I(RM^2) +
    AGE + DIS + RAD + TAX + PTRATIO + B + LAT + LON,
  data = boston,
  listw = W_2,
  family = "SAR"
)
```

```
## Warning in sqrt(fdHess[1, 1]): NaNs produced
```

```r
#CAR

CAR2 <- spautolm(
  log(CMEDV) ~ CRIM + ZN + INDUS + CHAS + I(NOX^2) + I(RM^2) +
    AGE + DIS + RAD + TAX + PTRATIO + B + LAT + LON,
  data = boston,
  listw =W_2,
  family = "CAR"
)
```

```
## Warning in spautolm(log(CMEDV) ~ CRIM + ZN + INDUS + CHAS + I(NOX^2) + I(RM^2)
## + : Non-symmetric spatial weights in CAR model
## Warning in spautolm(log(CMEDV) ~ CRIM + ZN + INDUS + CHAS + I(NOX^2) + I(RM^2)
## + : NaNs produced
```

Under presenteras utskriften för den ny-skattade CAR modellen. Man kan notera att när den nya viktmatrisen används på CAR modellen, så minskar λ fråm 0.61 till 0.23 och den är fortfarande signifikant. Dock kan man notera att AIC ökar med den nya viktmatrisen. Man kan tolka detta som att det finns ett spatialt beroende kvar, men det är inte lika starkt som tidigare modells. Slutsatsen kan dras att CAR modellen påverkas av viktmatrisen, där reslutaten förändras för viktmatrisen med de inversa avstånden

```r
summary(CAR2)
```

```
##
## Call: spautolm(formula = log(CMEDV) ~ CRIM + ZN + INDUS + CHAS + I(NOX^2) +
##     I(RM^2) + AGE + DIS + RAD + TAX + PTRATIO + B + LAT + LON,
##     data = boston, listw = W_2, family = "CAR")
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.736949 -0.096792 -0.011807  0.081651  1.000646
##
## Coefficients:
##                Estimate  Std. Error z value  Pr(>|z|)
## (Intercept) -3.0071e+01  1.4800e+01 -2.0318 0.0421736
## CRIM        -1.3680e-02  1.4591e-03 -9.3755 < 2.2e-16
## ZN          -1.1784e-03  6.4767e-04 -1.8194 0.0688509
## INDUS       -2.9145e-03  2.9577e-03 -0.9854 0.3244190
## CHAS         3.8302e-02  3.9785e-02  0.9627 0.3356798
## I(NOX^2)    -6.9311e-01  1.4213e-01 -4.8764 1.080e-06
## I(RM^2)      1.5607e-02  1.2647e-03 12.3410 < 2.2e-16
## AGE         -3.5918e-04  5.8751e-04 -0.6114 0.5409548
## DIS         -2.3523e-02  9.8718e-03 -2.3828 0.0171800
## RAD          1.1021e-02  3.1662e-03  3.4807 0.0005001
## TAX         -5.5224e-04  1.7450e-04 -3.1647 0.0015525
## PTRATIO     -3.7766e-02  6.5450e-03 -5.7701 7.920e-09
## B            6.6235e-04  1.2532e-04  5.2853 1.255e-07
## LAT         -5.5094e-02  1.8296e-01 -0.3011 0.7633246
## LON         -5.0410e-01  1.6475e-01 -3.0598 0.0022150
##
## Lambda: 0.23359 LR test value: 37.928 p-value: 7.3405e-10
## Numerical Hessian standard error of lambda: NaN
```

```
##
## Log likelihood: 73.98554
## ML residual variance (sigma squared): 0.043226, (sigma: 0.20791)
## Number of observations: 506
## Number of parameters estimated: 17
## AIC: -113.97
```

I utskriften nedan presenteras reslutatet för den nya SAR modellen. Denna visar att när viktmatrisen har bytts ut till en med inversa avstånd, så förblir λ nästan oförändrat fast sänks med 0.01 enhet. λ är fortfarande signifikant. Det som skiljer denna modell från den tidigare SAR, är att denna har lägre AIC värde. Det vill säga det finns fortfarande ett spatialt beroende kvar som är mycket stabilt. Slutsatsen kan dras att SAR modellen förändras då viktmatrisen byts ut mot en med inversa avstånd, där man kan notera en förbättring i modellen (enligt AIC).

```
summary(SAR2)
```

```
##
## Call: spautolm(formula = log(CMEDV) ~ CRIM + ZN + INDUS + CHAS + I(NOX^2) +
##     I(RM^2) + AGE + DIS + RAD + TAX + PTRATIO + B + LAT + LON,
##     data = boston, listw = W_2, family = "SAR")
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.7808560 -0.0634500 -0.0020534  0.0627788  0.8807874
##
## Coefficients:
##                Estimate  Std. Error z value  Pr(>|z|)
## (Intercept) -3.1232e+01  2.6554e+01 -1.1762 0.2395188
## CRIM        -8.4796e-03  1.0748e-03 -7.8896 3.109e-15
## ZN           1.4842e-03  6.0568e-04  2.4505 0.0142672
## INDUS       -7.1030e-04  3.2493e-03 -0.2186 0.8269618
## CHAS        -3.7735e-02  3.0405e-02 -1.2411 0.2145861
## I(NOX^2)    -4.4958e-01  1.7011e-01 -2.6429 0.0082190
## I(RM^2)      1.4510e-02  9.6686e-04 15.0068 < 2.2e-16
## AGE         -2.8321e-03  5.0944e-04 -5.5592 2.711e-08
## DIS         -2.8431e-02  1.2200e-02 -2.3305 0.0197809
## RAD          1.3027e-02  3.4098e-03  3.8205 0.0001332
## TAX         -5.6822e-04  1.4897e-04 -3.8142 0.0001366
## PTRATIO     -2.7189e-02  6.5770e-03 -4.1339 3.566e-05
## B            7.5709e-04  1.1628e-04  6.5107 7.481e-11
## LAT          3.1216e-01  3.4133e-01  0.9145 0.3604325
## LON         -3.0004e-01  2.9119e-01 -1.0304 0.3028184
##
## Lambda: 0.69627 LR test value: 277.84 p-value: < 2.22e-16
## Numerical Hessian standard error of lambda: NaN
##
## Log likelihood: 193.94
## ML residual variance (sigma squared): 0.021546, (sigma: 0.14679)
## Number of observations: 506
## Number of parameters estimated: 17
## AIC: -353.88
```

Deluppgift 2f)

Jämför resultaten från alla modeller. Vilken modell är att föredra?

För att undersöka vilken av modellerna som är att föredra bör man kolla på om det finns spatialt beroende i modellen, AIC, och modellanpassningen. Modellerna har följande AIC värden:

```
AIC <- data.frame(
  Modell = c("SAR (första)", "SAR (andra)", "CAR (första)", "CAR (andra)", "OLS modell"),
  AIC = c(-346.1,-353.9,-188.8, -114.0, AIC(lmModel))
)
kable(AIC, caption ="Alla modellers AIC värden.")
```

Table 7: Alla modellers AIC värden.

Modell	AIC
SAR (första)	-346.10000
SAR (andra)	-353.90000
CAR (första)	-188.80000
CAR (andra)	-114.00000
OLS modell	-78.04308

Utifrån tabell 7, ser man att CAR modellen försämras då en ny viktmatris används. För SAR modellerna ser man motsatsen, där den andra modellen förbättras när viktmatrisen med det inversa avståndet används. Den modell med sämst AIC är den vanliga linjär regressionsmodellen som beräknades först, där AIC $= -78$. SAR modellen med den inversa avstånds viktmatrisen är den modell som presterar bäst av alla utifrån AIC.

När det kommer till att kolla på det spatiala beroende kan den vanliga linjär regression ignoreras. Detta är eftersom att i linjär regression antas residualerna vara oberoende, men Morans I säger emot detta. Nu presenteras följande modellers spatiala beroende i en tabell:

```
tab2 <- data.frame(
  Modell = c("SAR (första)", "SAR (andra)", "CAR (första)", "CAR (andra)"),
  lambda = c(0.71, 0.70, 0.61,0.23),
  Signifkant = c(rep("Ja",4))
)

kable(tab2, caption = "Styrkan i det spatiala beroendet för alla modeller.")
```

Table 8: Styrkan i det spatiala beroendet för alla modeller.

Modell	lambda	Signifkant
SAR (första)	0.71	Ja
SAR (andra)	0.70	Ja
CAR (första)	0.61	Ja
CAR (andra)	0.23	Ja

I tabell 8 presenteras styrkan av det spatiala beroendet. I tabellen kan man notera att för den första SAR modellen är värder 0.71 och sänks med en enhet när viktmatrisen ändras. För den första CAR modellen är $\lambda = 0.61$, men när viktmatrisen ändras sänks denna till 0.23. Man vill här ha ett så högt värde som möjligt, och förstås signifikant. Detta innebär ett starkare beroende. Reslutatet visar att den modell som är minst

stabil är CAR modellen med den inversa-avstånds viktmatrisen. Förstaplats blir delad, där den första och andra SAR modellen, i princip, är lika bra.

Med detta sagt kan man konstantera att den bästa modellen är SAR med den viktade inversa avståndsmatrisen. Detta är eftersom denna modell har lägst AIC och samtidigt ett högt, stabilt λ värde.