Statistisk Analys av Komplexa Data
Spatiala Data

Mattias Villani

Statistik och Maskininlärning
Institutionen för Datavetenskap
Linköpings Universitet

Momentets innehåll

- Introduktion till **spatiala data**
- **Geostatistiska data**: Interpolation, Variogram och Kriging
- **Arealdata**: Spatiala autoregressionsmodeller.

Tre typer av spatiala data

- **Spatiala data**: **position** och **avstånd** har betydelse.
- Tre typer av spatiala data:
  - **Geostatistiska data** (Mätstationer)
  - **Arealdata** (Bildpixlar)
  - **Punktmönsterdata** (Fågelskådning)
- **Tempo-spatiala** data. Spatiala mätningar över **tid**. Ex: Meteorologiska mätningar.
- Inom delmomentet: Geostatistiska och lite areal data.

Geostatistiska data

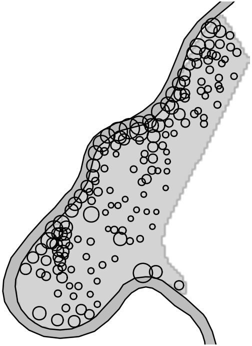
- $Y(s)$ är en slumpmässig vektor observerad vid positionen $s \in D$.
- **Positionerna** $s_1, s_2, ..., s_n$ är **fixa**.
- **Mätningarna** vid positionerna $Y(s_1), Y(s_2), ..., Y(s_n)$ är **slumpmässiga**.
- D är ofta en delmängd av $\mathbb{R}^2$. Longitud och latitud.

Geostatistiska data

- $Y(s)$ är en slumpmässig vektor observerad vid positionen $s \in D$.
- **Positionerna** $s_1, s_2, ..., s_n$ är **fixa**.
- **Mätningarna** vid positionerna $Y(s_1), Y(s_2), ..., Y(s_n)$ är **slumpmässiga**.
- D är ofta en delmängd av $\mathbb{R}^2$. Longitud och latitud.

- Exempel 1: Temperatur och nederbörd i min trädgård igår.
- Exempel 2: Mängd olja vid olika borrningsstationer.
- Exempel 3: Huspriser.

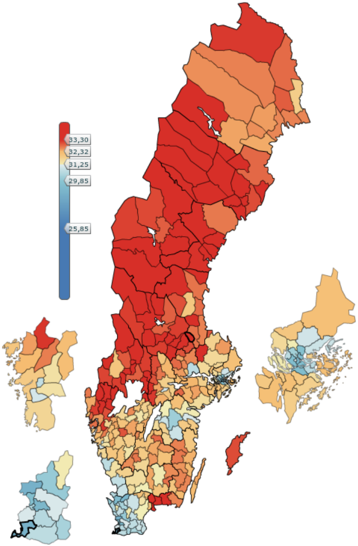
Meuse river data - Zinc concentration



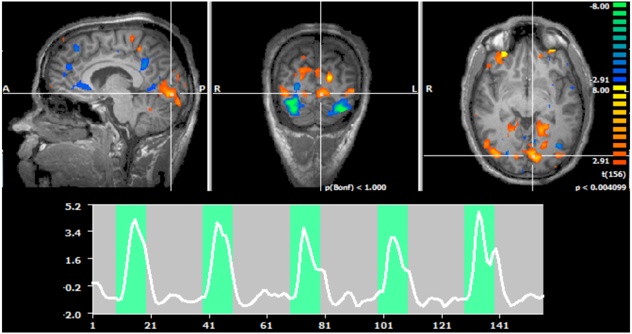
Areal data

- D är fortfarande en fix delmängd, men partitionerad i **arealenheter**. En mätning i varje area.
- Exempel 1: Jordbruksexperiment.
- Exempel 2: Bildanalys. Röntgen/Magnetkamera.
- Exempel 3: Huspriser på kommunnivå.
- Areal data kallas också **lattice** data.

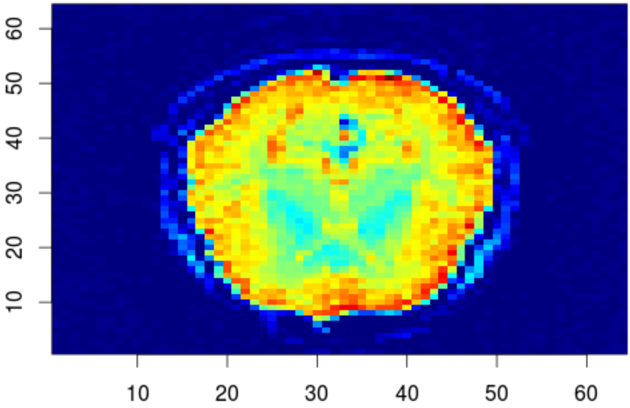
Kommunalskatt 2012



fMRI bilder av hjärnaktivitet



fMRI bilder av hjärnaktivitet



Punktmönsterdata

- Positionerna $s_1, s_2, ..., s_n$ är slumpmässiga. **Punktprocesser**.
- $Y(s_1), Y(s_2), ..., Y(s_n)$ är fixa som indikerar att en viss händelse inträffat vid s_i.
- Exempel 1: Positioner av viss trädsort.
- Exempel 2: Sjukdomsutbrott.
- Exempel 3: Brottsplatser.
- Klustring ofta viktigt. Tenderar vissa djurarter att befinna sig inom samma område? Olika områden (revir)? Attraction/repulsion.
- Kovariatinformation vid de slumpmässiga positionerna: **marked point pattern**. Ex: brottsstyp, fågelstorlek, trädomkrets.

Spatiala data i R

- Paketet **sp** är baspaketet med spatiala datatyper.
- Objekt för att rita upp spatiala data: **punkt**, **linje**, **polygon**, **grid** och **pixel**.
- `coordinates(dataMatris) <- dinaKoordinater` [`dinaKoordinater` är en matris med två kolumner innehållande longitud och latitud]
- dataMatris blir ett objekt av typen `SpatialPointsDataFrame`.
- Alternativ: funktionen `SpatialPointsDataFrame()`.
- `llCRS <- CRS("+proj=longlat +ellps=WGS84")` definerar det traditionella longitud-latitude koordinatsystemet.
- `library(maps); map(‘‘world’’, ‘‘sweden’’)` ritar upp en Sverigekarta.
- `gridObjekt <- as(SpatialDataMatris, "SpatialPixels")` Gör om matrisen `SpatialDataMatris` med punkter till en grid/pixlar.

Interpolation av spatiala data

- Vi har observerat $Z(s_1), Z(s_2), ..., Z(s_n)$ vid n fixa positioner.
- Vi vill beräkna anpassningen $\hat{Z}(s_0)$ i en ny punkt s_0.
- Interpolation viktat med inversa avståndet:

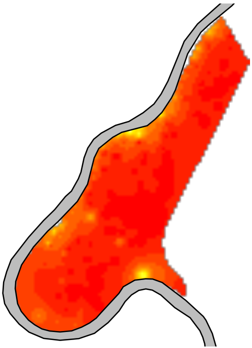
$$\hat{Z}(s_0) = \frac{\sum_{i=1}^{n} w(s_i, s_0) Z(s_i)}{\sum_{i=1}^{n} w(s_i, s_0)}$$

  där

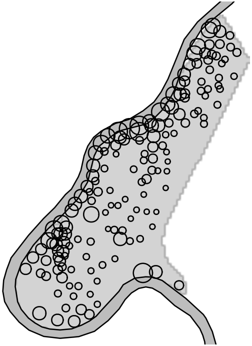
$$w(s_i, s_0) = \|s_i - s_0\|^{-p}$$

- `library(gstat); idwOut <- idw(zinc ~ 1, meuse, meuse.grid, idp = 2)`
- `image(idwOut)` ritar upp den interpolerade ytan.

Interpolation med inversa avstånd



Meuse river data - Zinc concentration

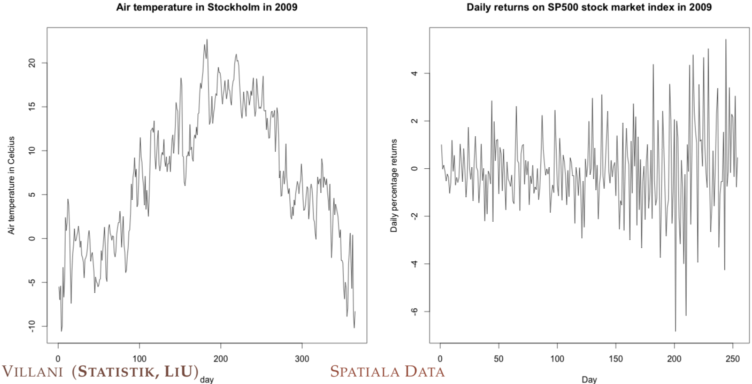


Tidserier

- **Tidserier**: data observerat över **tid**. $x_1, x_2, ..., x_T$.
- Autoregressiv model av ordning ett:

$$x_t = \phi x_{t-1} + \varepsilon_t$$

  där ε_t~N(0,σ^2) är oberoende slumpfel.
- **Beroende över tid**. Värdet idag beror på gårdagens värde.



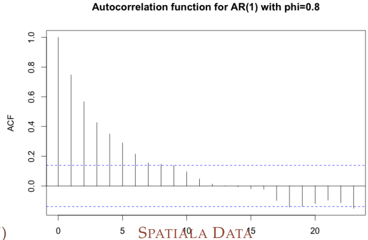
Autokorrelation

- **Stationär** tidsserie:
  - Konstant väntevärde över tiden
  - Konstant varians över tiden
  - Autokorrelationen mellan X_t och X_{t+h} beror bara på h.
- **Autokorrelationsfunktionen** visar beroendet över tid:

$$C(h) = Corr(X_t, X_{t+h})$$

- **Variogrammet** innehåller samma info som $C(h)$:

$$\mathrm{E}\left[(X_t - X_{t+h})^2\right]$$



Variogram

- Spatiala data är som tidserier där tiden t byts ut mot **rumsliga koordinater $\mathbf{s} = (s_1, s_2)$**.
- Process
$$Z(\mathbf{s}) = m + e(\mathbf{s})$$
där m är väntevärdet och $e(\mathbf{s})$ är spatialt brus.
- Alternativt
$$Z(\mathbf{s}) = X\beta + e(\mathbf{s})$$
- **Semivariogram**
$$\gamma(\mathbf{h}) = \frac{1}{2}\mathrm{E}\left[Z(\mathbf{s}) - Z(\mathbf{s}+\mathbf{h})\right]^2$$
där $\mathbf{h}$ är en vektor.
- **Variogram** $= 2\gamma(\mathbf{h})$.
- Relation till kovariansfunktionen $C(\mathbf{h})$=Cov$[Y(\mathbf{s}+\mathbf{h}), Y(\mathbf{s})]$:
$$\gamma(\mathbf{h})=2[C(\mathbf{0}) - C(\mathbf{h})]$$

Variogram, forts

- **Stationäritet över rummet.**
- Spatial korrelation beror endast på avståndsvektorn $\mathbf{h} = \mathbf{s}_i - \mathbf{s}_j$ mellan två punkter och inte på punkternas positioner.
- **Isotropisk korrelation**: vektorn $\mathbf{h}$ kan ersättas med dess längd $h = \|\mathbf{h}\|$.
- $\gamma(-\mathbf{h}) = \gamma(\mathbf{h})$ och $\gamma(\mathbf{0}) = 0$.
- **Nugget**: ofta antas att $\gamma(\mathbf{0}^+) = \lim_{h \to 0^+} \gamma(h) = \tau^2 > 0$. Går dock inte att skatta utan upprepade observationer vid samma position.

Sampelvariogrammet

- Problematiskt att skatta variogrammet eftersom vi kan ha inga eller få datapunkter som just har avståndet h. Jfr ACF i tidsserieanalys.
- Låt $I_1 = [0, h_1), I_2 = [h_1, h_2), ..., I_m = [h_{m-1}, h_m)$ vara en partitionering av intervallet $[0, h_m)$ där h_m är en övre gräns.
- Momentskattning av (semi)variogrammet:

$$\hat{\gamma}(I_j) = \frac{1}{2N_h} \sum_{i=1}^{N_h} [Z(s_i) - Z(s_i + h)]^2$$

  för alla $h \in I_j$.
- `library(gstat); plot(variogram(log(zinc) ~ 1, meuse))`
- `library(gstat); plot(variogram(log(zinc) ~ sqrt(dist), meuse))`
- `library(gstat); plot(variogram(log(zinc) ~ 1, meuse, alpha = c(0,45,90,135)))`

Modeller för isotropiska variogram

- **Linjär**

$$\gamma(h;\theta) = \begin{cases} \tau^2 + \sigma^2 h & \text{om} \quad h > 0 \\ 0 & \text{om} \quad h = 0 \end{cases}$$

- **Sfärisk**

$$\gamma(h;\theta) = \begin{cases} \tau^2 + \sigma^2 & \text{om} \quad h \geq 1/\phi \\ \tau^2 + \sigma^2 \left\{ \frac{3\phi h}{2} - \frac{1}{2}\phi^3 h^3 \right\} & \text{om} \quad 0 \leq h \leq 1/\phi \end{cases}$$

- **Powered exponential** $0 < p \leq 2$ **(Exponential** $p = 1$**, Gaussisk** $p = 2$**)**

$$\gamma(h;\theta) = \begin{cases} \tau^2 + \sigma^2 \left[1 - \exp(-\left|\phi h\right|^p)\right] & \text{om} \quad h > 0 \\ 0 & \text{om} \quad h = 0 \end{cases}$$

- **Matérn**

$$\gamma(h;\theta) = \begin{cases} \tau^2 + \sigma^2 \left[1 - \frac{(2\sqrt{\nu} h\phi)^\nu}{2^{\nu-1}\Gamma(\nu)} K_\nu(2\sqrt{\nu} h\phi)\right] & \text{om} \quad h > 0 \\ 0 & \text{om} \quad h = 0 \end{cases}$$

Nugget, Sill och Range

- Exempel: sfäriskt variogram

$$\gamma(h;\theta) = \begin{cases} \tau^2+\sigma^2 & \text{om} \quad h \geq 1/\phi \\ \tau^2+\sigma^2\left\{\frac{3\phi h}{2} - \frac{1}{2}\phi^3 h^3\right\} & \text{om} \quad 0 \leq h \leq 1/\phi \end{cases}$$

- **Nugget**: $\gamma(\mathbf{0}^+) = \lim_{h\to 0^+}\gamma(h) = \tau^2 > 0$
- **Sill**: $\lim_{h\to\infty}\gamma(h) = \tau^2+\sigma^2$
- **Partial sill**: Sill - Nugget $=\sigma^2$
- **Range**: h-värdet där $\gamma(h)$ först når sitt maximum: $1/\phi$.

Skattning av parametriska variogram

- Ickelinjär viktad minsta kvadrat (startvärden viktiga)

$$\sum_{j=1}^{p} w_j \left[\gamma(h) - \hat{\gamma}(h)\right]^2$$

- **Variogrammodeller i R**: `vgm(psill, model, range, nugget)`
- Exempel: `v <- vgm(1, ‘‘Sph’’, 800, 1)`
- `library(gstat); sampleVariogram <- variogram(log(zinc) ~ sqrt(dist), meuse)`
- `vFit <- fit.variogram(sampleVariogram, v); plot(sampleVariogram, vFit)`
- attributes(vFit) ger SSE.

Skattning av parametriska variogram - eyeballing

- Eyeball statistics: Prova olika parametervärdet tills dess parametriskt variogram anpassar sampelvariogrammet.
- `library(geoR); varioEye <- eyefit(variog(as.geodata(meuse["zinc"]), max.dist = 1500)); varioFit <- as.vgm.variomodel(varioEye[[1]])`
- Fungerar inte i RStudio. Kör "vanlig" R. Datamaterialet `meuse` `finns i paketet sp.`

Spatial prediktion - Kriging

- Vi har observationer $Z(\mathbf{s}_1), ..., Z(\mathbf{s}_n)$ och vill prediktera $Z(\mathbf{s}_0)$, Z-värdet vid en ny position $\mathbf{s}_0$.
- Vi har kovariater vid varje position: $X = (x(\mathbf{s}_1), ..., x(\mathbf{s}_n))'$ och $x(\mathbf{s}_0)$.
- Rimlig prediktor: $\hat{Z}(\mathbf{s}_0)$ är ett viktat medelvärde av värdena vid "närliggande" positioner.
- Bästa linjära väntevärdesriktiga prediktorn

$$\hat{Z}(\mathbf{s}_0) = x(\mathbf{s}_0)\hat{\beta} + v'V^{-1}\left(Z(\mathbf{s}) - X\hat{\beta}\right)$$

där $Z(\mathbf{s}) = (Z(\mathbf{s}_1), ..., Z(\mathbf{s}_n))'$, V är kovariansmatrisen för $Z(\mathbf{s})$ och v är kovariansvektorn innehållande kovarianserna mellan $Z(\mathbf{s}_0)$ och $Z(\mathbf{s})$, och

$$\hat{\beta} = \left(X'V^{-1}X\right)^{-1}X'V^{-1}Z(\mathbf{s})$$

är den vanliga GLS-skattningen.

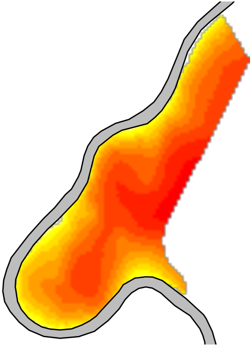
Spatial prediktion - Kriging, forts.

- Prediktionsvarians

$$Var\left(\hat{Z}(s_0)\right) = Var\left(Z(s_0)\right) - v'V^{-1}v' + \left(x(s_0) - v'V^{-1}X\right)\left(X'V^{-1}X\right)^{-1}\left(x(s_0) - v'V^{-1}X\right)$$

- **Universal kriging**.
- **Ordinary kriging**: endast intercept i regressionsytan.
- `krige(log(zinc) ~ sqrt(dist), meuse, meuse.grid, fittedSphVariogram)`

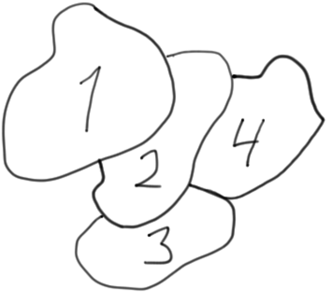
Kriging - Meuse data



Areal data

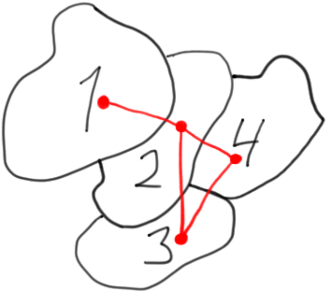
- Datapunkter är definierade över arealenheter (kommuner, pixlar)
- Arealernas **närhet** till varandra äv viktig. Motsvarar avstånd för geostatistiska data.
- **Spatiala grannskap** (neighbourhoods):
  - Vilka regioner gränsar till region i?
  - Vilka länder handlar med varandra? Hur mycket?
  - Vilka delar av hjärnan är sammankopplade med fibrer?
  - Vem är du vän med på Facebook?

Spatiala grannskap - ett exempel



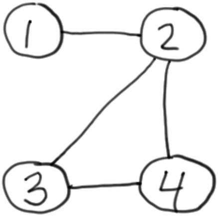
$$W = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix} \qquad \tilde{W} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0.33 & 0 & 0.33 & 0.33 \\ 0 & 0.5 & 0 & 0.5 \\ 0 & 0.5 & 0.5 & 0 \end{pmatrix}$$

Spatiala grannskap - ett exempel



$$W = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix} \qquad \tilde{W} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0.33 & 0 & 0.33 & 0.33 \\ 0 & 0.5 & 0 & 0.5 \\ 0 & 0.5 & 0.5 & 0 \end{pmatrix}$$

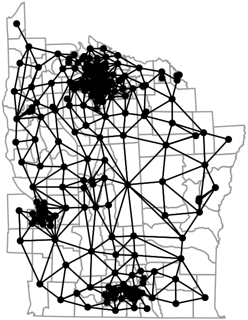
Spatiala grannskap - ett exempel



$$W = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \end{pmatrix} \qquad \tilde{W} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0.33 & 0 & 0.33 & 0.33 \\ 0 & 0.5 & 0 & 0.5 \\ 0 & 0.5 & 0.5 & 0 \end{pmatrix}$$

US Census tract data - New York counties

US Census tract data - New York counties



Moran's I

- Näthetsmatris $W = (w_{ij})$[objekt av klassen `nb` i R]
- Test för spatialt beroende, **Moran's I**

$$I = \frac{n}{\sum_{i=1}^{n}\sum_{i=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- **Moran's I** i vektorform

$$I = \frac{(\mathbf{y} - \bar{y})'\tilde{\mathbf{W}}(\mathbf{y} - \bar{y})}{(\mathbf{y} - \bar{y})'(\mathbf{y} - \bar{y})}$$

- $-1 \leq I \leq 1$ och $E(I) = -\frac{1}{n-1}$ vid spatiellt oberoende.
- Globalt vs Lokalt beroende.
- Geary's C mer lokalt alternativ till Moran's I.

R kommandon för arealdata

- Mål: vikter för varje arealenhet som beskriver dess beroende av enhetens grannar.
- Två steg:
  1. Definera närhetsmatris (vem är granne med vem?). `nb` objekt i R [**n**eigh**b**ourhood].
  2. Bestäm vikter mellan alla par av grannar. nb2listw [**n**eigh**b**ourhood **to** **list** of **w**eights]

R kommandon för arealdata, forts.

- Närhetsvikter: `listw` objekt.
- **NYlistw[[2]]** är en lista där det *i*:te listelementet innehåller information om **vilka grannar** den *i*:te arealenheten har.
- **NYlistw[[3]]** är en lista där det *i*:te listelementet innehåller information om **vilka vikter** den *i*:te arealenhetens grannar har.
- Exempel: NYlistw <- nb2listw(NY_nb, style = "B") [Binärt definierade grannar. Granne eller ej.]
- NYlistw[[2]][[3]] returnerar c(2,13,35), vilket säger att enheterna 2, 13 och 35 är grannar med enhet 3.
- NYlistw[[3]][[3]] returnerar c(1,1,1), vilket säger att enheterna 2, 13 och 35 har alla vikt 1 för enhet 3 (binär grannrelation).
- Exempel: NYlistw <- nb2listw(NY_nb, style = "W") [Binärt definierade grannar. Granne eller ej.]

Simultan autoregressiv modell (SAR)

- Tidsserieanalys: regression med AR(1) feltermer

$$y_t = \beta_0 + \beta_1 x_t + e_t$$
$$e_t = \rho e_{t-1} + \varepsilon_t$$

där $\varepsilon_t \overset{iid}{\sim} N(0, \sigma^2)$.

- Simultan autoregressiv modell (SAR)

$$y_i = \beta_0 + \beta_1 x_i + e_i$$
$$e_i = \sum_{j=1}^{m} b_{ij} e_j + \varepsilon_i$$

där $\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$.

- b_{ij} representerar spatiala beroenden mellan regioner.
- $b_{ii} = 0$ för alla i. Regionen beror inte på sig själv.

Simultan autoregressiv modell (SAR), forts

- Modellen kan skrivas som

$$E(\mathbf{y}) = \mathbf{X}\beta$$
$$Var(\mathbf{y}) = (\mathbf{I} - \mathbf{B})^{-1} \Sigma_{\varepsilon} \left(\mathbf{I} - \mathbf{B}'\right)^{-1}$$

  där $\mathbf{B} = (b_{ij})$ och $^{\circ}{}_{\varepsilon}$ är en diagonal matris, ofta $\Sigma_{\varepsilon} = \sigma^2 I$.
- Vanligt val: $\mathbf{B} = \lambda \mathbf{W}$. Spatial autokorrelationsparameter: λ.
- R funktionen: `nysar <- spautolm(Z ~ PEXPOSURE + PCTAGE65P + PCTOWNHOME, data = NY8, listw = NYlistw)`

Conditional Autoregressive Model (CAR)

- Beroendet för residualerna modelleras betingat på omgivningen (neighbourhood)

$$e_i | e_{j\sim i} \sim N\left(\sum_{j\sim i} \frac{c_{ij} e_j}{\sum_{j\sim i} c_{ij}}, \frac{\sigma^2_{e_i}}{\sum_{j\sim i} c_{ij}}\right)$$

- Om t ex $\mathbf{C} = \lambda \mathbf{W}$ och $\Sigma_\varepsilon = \sigma^2 I$ så gäller för CAR att

$$E(\mathbf{y}) = \mathbf{X}\beta$$
$$Var(\mathbf{y}) = \sigma^2 (I - \lambda \mathbf{W})^{-1}$$

  där $\mathbf{C} = (c_{ij})$.

- R: `nycar <- spautolm(Z ~ PEXPOSURE + PCTAGE65P + PCTOWNHOME, + data = NY8, family = "CAR", listw = NYlistw)`