|  |
| --- |
| **Spatial data analysis (in R)**  **prof.ucz.dr hab. Katarzyna Kopczewska**  **Class 02b –** spatial statistics to detect spatial clustering, spatial regimes and autocorrelation |

1. **Spatial statistics**

First step: a map of phenomenon - is there visible any spatial autocorrelation and spatial heterogeneity?

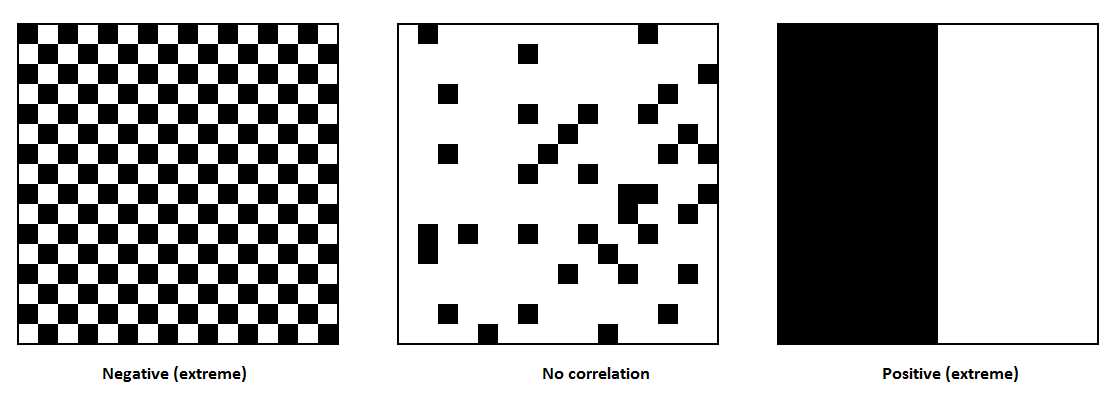
|  |  |
| --- | --- |
| **# spatial distribution – unemployment rate**  # unemployment rate in 2012  x<-data$XA21[data$year==2012]  summary(x)  cols<-rev(heat.colors(8))  brks<-(0:8)\*5  plot(pov, col=cols[findInterval(x, brks)])  plot(voi, add=TRUE, lwd=2)  legend("bottomleft", legend=brks, pt.bg=cols, bty="n", pch=22)  title(main="Unemployment rate in 2012", sub="In legend intervals from….%") |  |

**Tobler’s first law of geography, “Everything is related to everything else, but near things are more related than distant things.”**

**a) global statistics**

Global statistics generate a global index, a measure of spatial dependence. It allows us to study the spatial autocorrelation - positive or negative. Positive autocorrelation means that similar observations (regions where the values of the selected variable are similar) are in contact more often than would result from a random distribution (spreading of characteristics between the fixed locations). Autocorrelation negative means that the related contact is less than random.

**Spatial correlation**



**Moran's I -** Basic global statistics, a measure of global spatial autocorrelation, measurs process / pattern that appears on throughout the territory. Depends on the difference between the test and the average value - this is a test based on the covariance matrix, similar to the Pearson test.

 I = \frac{N} {\sum_{i} \sum_{j} w_{ij}} \frac {\sum_{i} \sum_{j} w_{ij}(X_i-\bar X) (X_j-\bar X)} {\sum_{i} (X_i-\bar X)^2}  

Values between -1 <I <0 = negative correlation – similar values touch less often than random

Values between 0 <I <1 = positive correlation - similar values touch more often than random

More at <http://en.wikipedia.org/wiki/Moran's_I>

Other measures: Geary’s C, Global G (by Getis), join-count statistics

**b) local statistics**

Local Moran = LISA (Local Indicator of Spatial Association), answers the question: How unusual is the combination of observations zi and its neighbors Wzi; measure calculated for each area and its neighbors, thus possible mapping of the local statistics (autocorrelation and assessment of changes in space), and mapping the significance of statistics. Average of local statistics Ii gives global statistics I:  I= \sum_i \frac{I_i}{N} , where N is the number of areas. Local statistics are standardized weighted average of the values in the neighborhood,  I_i = \frac{Z_i}{m_2} \sum_j W_{ij} Z_j where, and Zi = (Xi-Xaverage).

Positive values​​ : the area surrounded by similar, low or high; area is part of a cluster,

Negative values​​ : the area surrounded by other values, is an outlier.

more at: <http://en.wikipedia.org/wiki/Indicators_of_spatial_association>

other measures: bivariate Moran’s I, Local G,

**c) mapping results - Moran Scatterplot**

Moran scatterplot, to evaluate the clusters and the relative values ​​of the variables in the locations surveyed and their neighbors. X-axis: test variable (preferably standardized), Y-axis: a spatial lag variable tested (preferably standardized).

Areas of positive spatial autocorrelation:

* Quadrant I (+, +) = HH (high-high) - high values surrounded by high
* Quadrant III (-, -) = LL (low-low) - low value surrounded by low

Areas where the spatial regimes change

* Quadrant II (+, -) = HL (high-low) – high values surrounded by low (island of wealth)
* Quadrant IV (-, +) = LH (Low-High) - low surrounded by high (the area of poverty, collapse).

By the points can be drawn regression line – its slope (wx~x) is the Moran.

1. **Basic functions in R for handling the spatial correlation measures**

moran Compute Moran's I

moran.mc Permutation test for Moran's I statistic

moran.plot Moran scatterplot

moran.test Moran's I test for spatial autocorrelation

geary Compute Geary's C

geary.mc Permutation test for Geary's C statistic

geary.test Geary's C test for spatial autocorrelation

globalG.test Global G test for spatial autocorrelation

joincount.mc Permutation test for same colour join count statistics

joincount.multi BB, BW and Jtot join count statistic for k-coloured factors

joincount.test BB join count statistic for k-coloured factors

localmoran Local Moran's I statistic

localmoran.exact Exact local Moran's Ii tests

localmoran.sad Saddlepoint approximation of local Moran's Ii tests

localG G and Gstar local spatial statistics

p.adjustSP Adjust local association measures' p-values

**6. Example codes in R**

**# Global Moran**

data12<-data[data$year==2012, ]

result01<-moran.test(data12$XA21, cont.listw)

result01

attributes(result01)

Moran's I test under randomisation

data: data12$XA21

weights: cont.listw

Moran I statistic standard deviate = 13.6153, p-value < 2.2e-16

alternative hypothesis: greater

sample estimates:

Moran I statistic Expectation Variance

0.462766094 -0.002666667 0.001168579

**# Moran scatterplot**

**# preparing the data**

x<-data12$XA21 # variable selection

zx<-as.data.frame(scale(x)) #standardization of variable

**# Moran scatterplot – automatic version**

moran.plot(zx$V1, cont.listw, pch=19, labels=as.character(data12$powiat\_name1))

**# Moran scatterplot – step by step version**

# chcecking the average and standard deviation

round(mean(zx$V1),0)

sd(zx$V1)

wzx<-lag.listw(cont.listw, zx$V1) # spatial lag of x

morlm<-lm(wzx~zx$V1) # linear regression

summary(morlm)

slope<-morlm$coefficients[2] # coefficient from regression

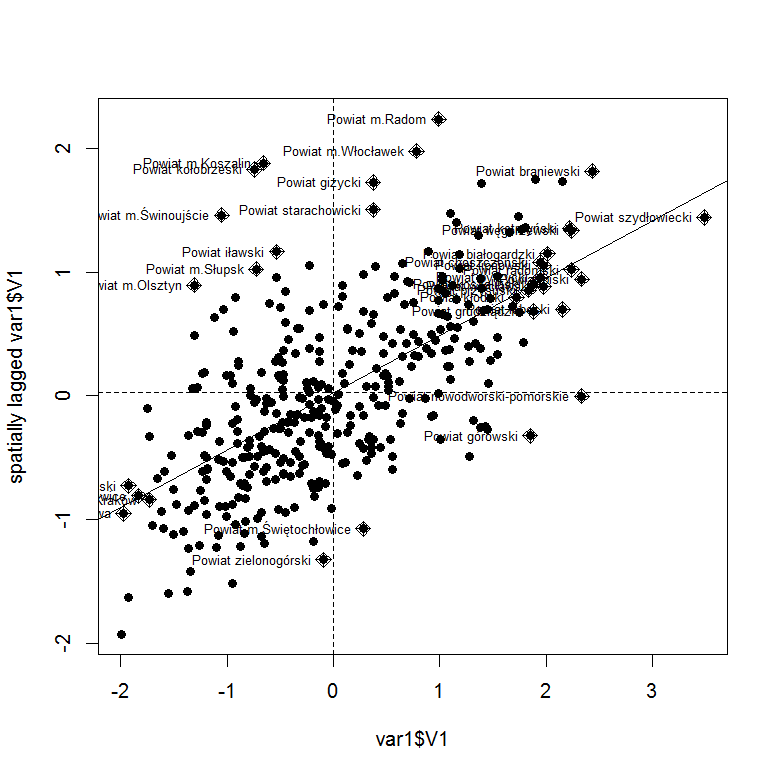
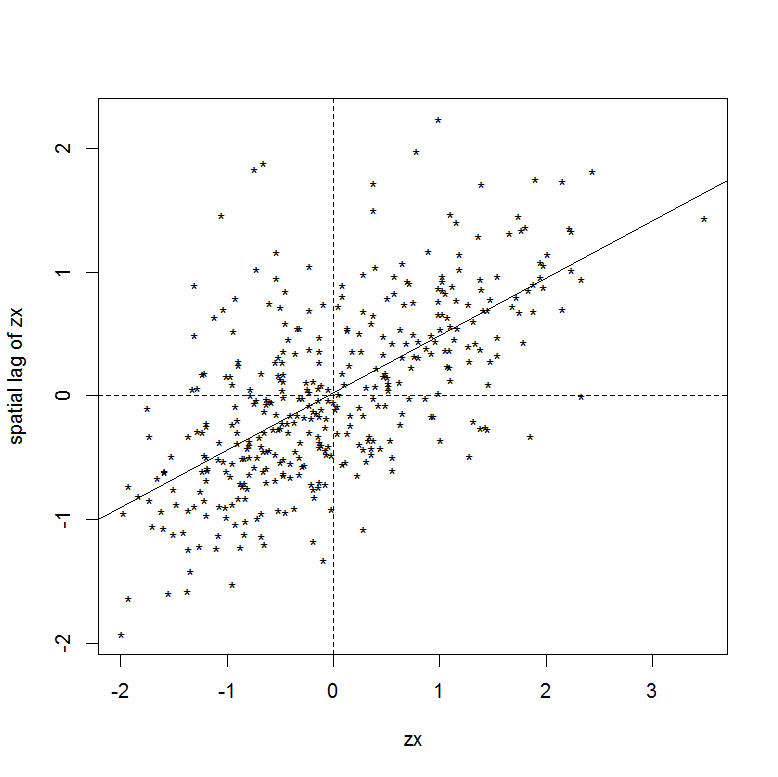
intercept<-morlm$coefficients[1] # constant term in regression

plot(zx$V1, wzx, xlab="zx",ylab="spatial lag of zx", pch="\*")

abline(intercept, slope) # regression line

abline(h=0, lty=2) # supplementary horizontal line y=0

abline(v=0, lty=2) # supplementary vertical line x=0

**# Mapping of moranscatterplots quarts**

**# we have x, wx & wzx from above**

cond1<-ifelse(zx>=0 & wzx>=0, 1,0) # I quarter

cond2<-ifelse(zx>=0 & wzx<0, 2,0) # II quarter

cond3<-ifelse(zx<0 & wzx<0, 3,0) # III quarter

cond4<-ifelse(zx<0 & wzx>=0, 4,0) # IV quarter

cond.all<-cond1+cond2+cond3+cond4 # all quarters in one

cond.all

cond<-as.data.frame(cond.all)

is.data.frame(cond)

**# map**

brks<-c(1,2,3,4)

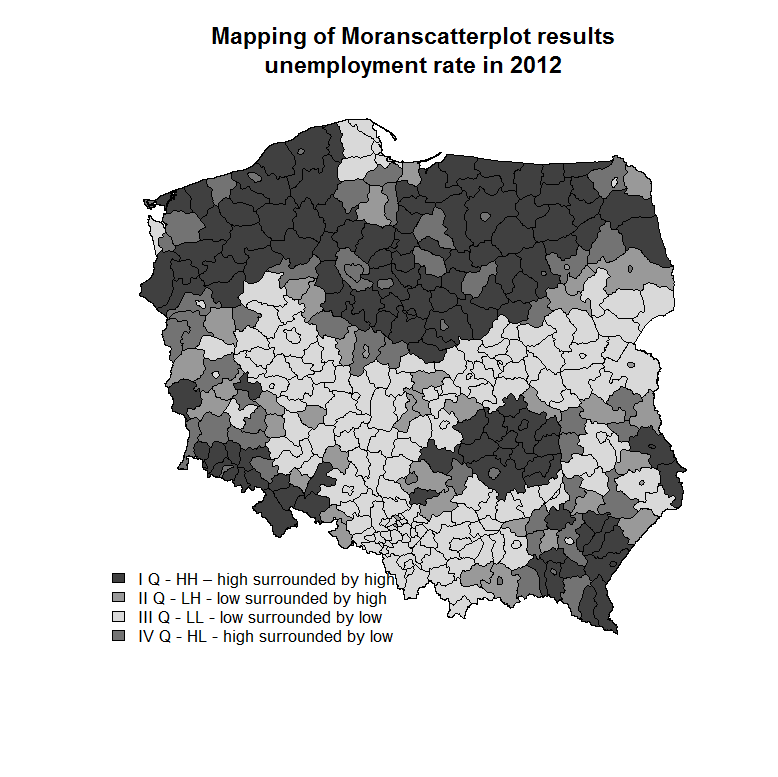
cols<-c("grey25", "grey60", "grey85", "grey45")

plot(pov, col=cols[findInterval(cond$V1, brks)])

legend("bottomleft", legend=c("I Q - HH – high surrounded by high", "II Q - LH - low surrounded by high", "III Q - LL - low surrounded by low", "IV Q - HL - high surrounded by low"), fill=cols, bty="n", cex=0.80)

title(main="Mapping of Moranscatterplot results

unemployment rate in 2012")



**# Calculations in the loop**

# calculate Moran’s I for all years in the dataset

# preparing object for writing the results

moran<-matrix(0, ncol=12, nrow=1)

colnames(moran)<-2006:2017

rownames(moran)<-"Moran’s I"

moran

for(i in 2006:2017){ # loop uses natural numbers

result01<-moran.test(data$XA21[data$year==i], cont.listw)

moran[1,i-2005]<-result01$estimate[1]}

moran

**# basic plot of the result**

plot(moran[1,])

**# improved plot of the result**

plot(moran[1,], type="l", axes=FALSE, ylab="", xlab="", ylim=c(0.3, 0.6))

axis(1, at=1:12, labels=2006:2017)

axis(2)

points(moran[1,], bg="red", pch=21, cex=1.5)

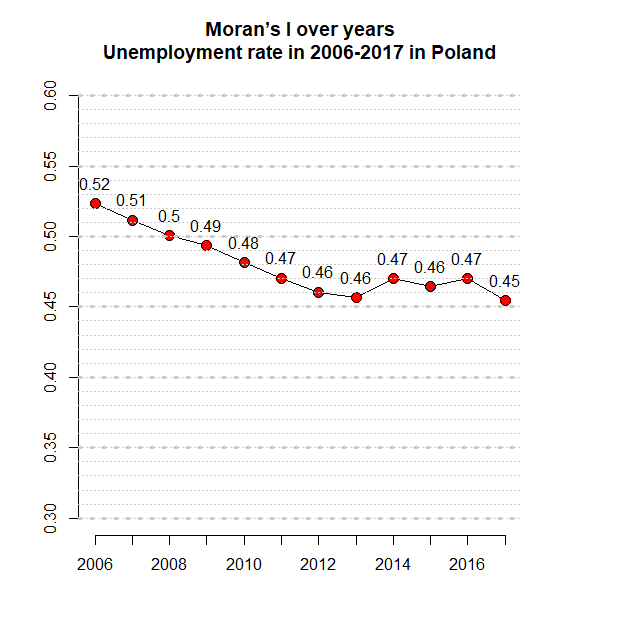
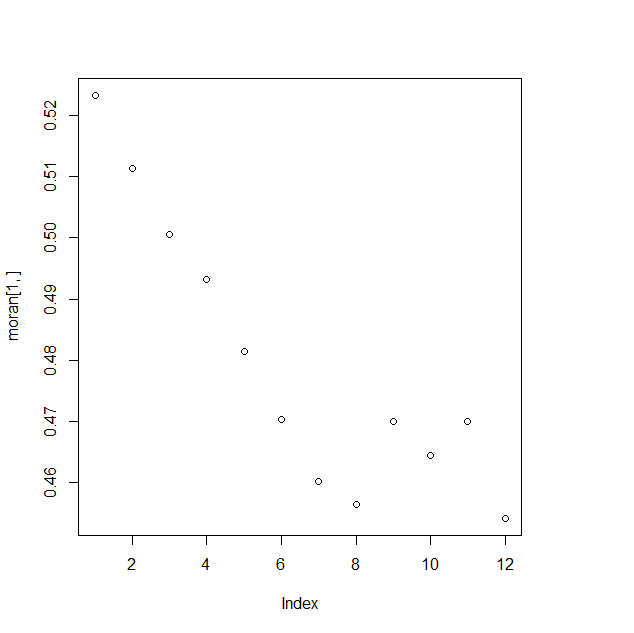
abline(h=c(0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6), lty=3, lwd=3, col="grey80")

abline(h=0.30+(0:30)\*0.01, lty=3, lwd=1, col="grey80")

text(1:12, moran[1,]+0.015, labels=round(moran[1,],2))

title(main="Moran’s I over years

Unemployment rate in 2006-2017 in Poland")



**# join-count test (based on colours)**

One of the types of statistics testing spatial relationships are join-count statistics. The idea of the test is to divide the variable value into groups, assign a colour to them and apply colours to the map. The number of contacts (connections) of the same colours relative to the total number of contacts is tested. Comparison of the observed frequency with the expected values is the essence of the spatial autocorrelation test. As spatial autocorrelation is considered more frequent than expected in probability contact of objects of the same colour. Tests are carried out for each colour group. The basis is a two-colour, black and white system, which is used in the test terminology (BW, Black-White).

The highlighted colour (e.g. black) is the examined feature, e.g. high variable level or presence of a phenomenon. Joint-count statistics determine the contact frequencies of black-black (BB), white-black (BW) and white-white (WW) areas. Hence, BB and WW statistics examine positive spatial autocorrelation, while BW statistics examine negative spatial autocorrelation.

summary(data12$XA21)

var.factor<-factor(cut(data12$XA21, breaks=c(0,10, 20, 40), labels=c("low", "medium", "high")))

head(var.factor)

# parameters of graphics

brks1<-c(0, 10, 20, 40)

cols<-c("green", "blue", "red")

# scatterplot of values

plot(data12$XA21, bg=cols[findInterval(data12$XA21, brks1)], pch=21)

abline(h=c(10,20,40), lty=3)

# spatial distribution with three colours

plot(pov, col=cols[findInterval(data12$XA21, brks1)])

plot(voi, add=TRUE, lwd=2)

title(main="Unemployment in 2012")

legend("bottomleft", legend=c("low", "medium", "high"), leglabs(brks1), fill=cols, bty="n")

joincount.test(var.factor, cont.listw)

Join count test under nonfree sampling

data: var.factor

weights: cont.listw

Std. deviate for low = 6.1872, p-value = 3.062e-10

alternative hypothesis: greater

sample estimates:

Same colour statistic Expectation Variance

8.9004509 3.7757256 0.6860419

Join count test under nonfree sampling

data: var.factor

weights: cont.listw

Std. deviate for medium = 4.5951, p-value = 2.163e-06

alternative hypothesis: greater

sample estimates:

Same colour statistic Expectation Variance

74.22410 64.14248 4.81364

Join count test under nonfree sampling

data: var.factor

weights: cont.listw

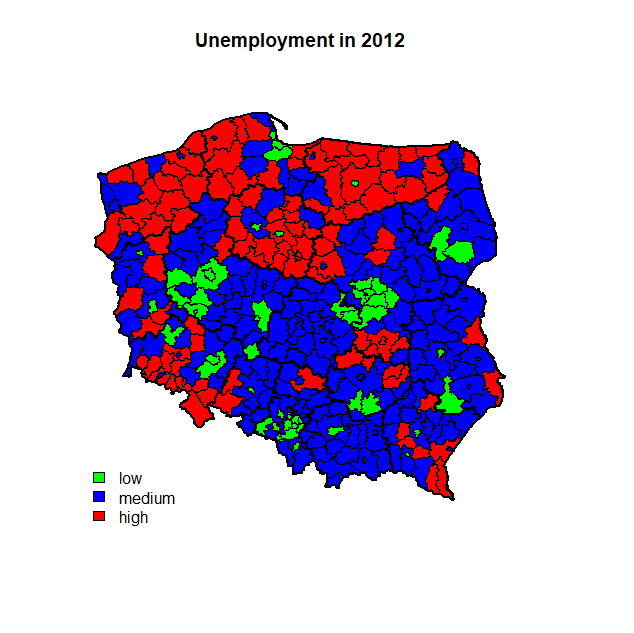
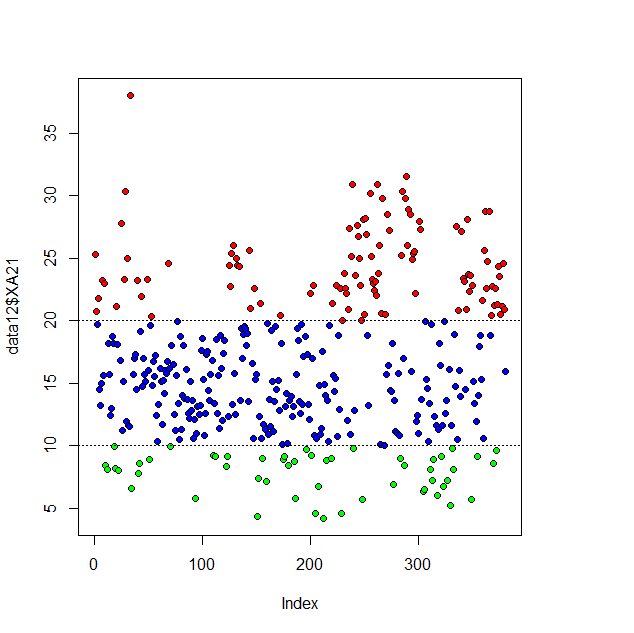
Std. deviate for high = 9.5638, p-value < 2.2e-16

alternative hypothesis: greater

sample estimates:

Same colour statistic Expectation Variance

28.281548 14.406332 2.104815



**Local Moran I statistics (LISA)**

Local Moran I statistics measure whether a region is surrounded by neighbouring regions with similar or different values of the examined variable in relation to the random distribution of these values in space. II is a smoothed index for individual observations, so it can be used to find the so-called hot spots and local clusters. Local Moran I statistics are proportional to Moran's global statistics. It is expressed by the formula:

|  |  |
| --- | --- |
|  |  |

where the wij elements come from a spatial first-row row-weighted (row-standardised) matrix. Statistical significance tests are based on the distribution resulting from conditional randomisation or permutation (Anselin, 1995), and hypotheses are verified on the basis of a pseudo level of significance. Standardised Local Moran statistics assume significantly negative values when the object i is surrounded by relatively low values in neighbouring objects, and significantly positive values when the object i is surrounded by relatively high values. A low p-value (p <0.05) means that the region i is surrounded by relatively high values, while a high p-value (p> 0.95) means that the region i is surrounded by relatively low values (Bao, 2000).

Significant local Moran statistics are plotted on the map, i.e. for which the p-value is below 0.05 or above 0.95. To facilitate data operations, the variable header has been changed from Pr .z…> 0) to Prob with the **names()** command.

locM<-localmoran(spNamedVec("XA21", data12), cont.listw)

oid1<-order(data12$ID\_MAP)

locMorMat<-printCoefmat(data.frame(locM[oid1,], row.names=data12$poviat\_name1[oid1]), check.names = FALSE)

# map of the significance of Moran's local statistics

names(locMorMat)[5]<-"Prob"

brks<-c(min(locMorMat[,5]), 0.05000, 0.95000, max(locMorMat[,5]))

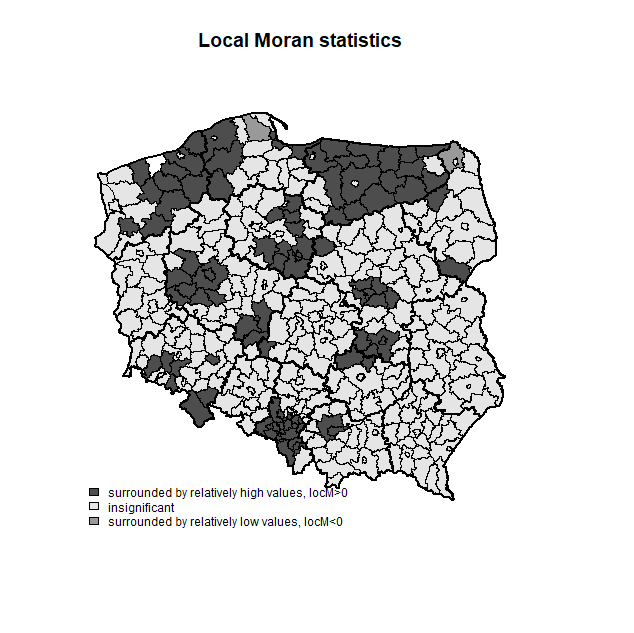
cols<-c("grey30", "grey90", "grey60")

plot(pov, col=cols[findInterval(locMorMat[,5], brks)])

legend("bottomleft", legend=c("surrounded by relatively high values, locM>0", "insignificant", "surrounded by relatively low values, locM<0"), fill=cols, bty="n", cex=0.75)

title(main=" Local Moran statistics ", cex=0.7)

plot(voi, add=TRUE, lwd=2)



**Task 1:** For the four spatial weights matrix (according to four criteria) and variable on salaries in Poland=100 (XA14):

- For each spatial weight matrix for years 2006:2017 find the Moran’s I – are there any significant changes? Does the choice of spatial weights matrix matter for the result? Plot it as four lines over time.

- Prepare for extreme years mapped Moran scatterplot. Are there any significant differences (if yes, where?

**Task 2:** Build a point dataset by assigning centroids (and their coordinates) to poviat data. Test Moran’s I for two selected variables using knn matrix with knn=1,2,3…..12. Are there any regularities?