Local Spatial Autocorrelation

Luc Anselin

THE CENTER FOR SPATIAL DATA SCIENCE
THE UNIVERSITY OF CHICAGO

http://spatial.uchicago.edu

LISA principle

Local Moran and Local Geary

Getis-Ord statistics

interpretation

extensions

LISA Principle

- Clustering vs Clusters

global spatial autocorrelation does NOT suggest the location of the clusters

cluster detection

identification of location

assessment of significance

many cluster detection methods

- Local Indicators of Spatial Association

LISA (Anselin 1995)

local spatial statistic - one for each location

sum of LISA proportional to a corresponding global statistic

- Local Spatial Autocorrelation Analysis

  assess significance of local statistic at each location

  identification of location of spatial clusters (hot spots, cold spots) and spatial outliers

  in absence of global S.A., or in presence of global S.A. (significance levels affected)

- LISA Forms of Global Statistics
  - every decomposable statistic

    if global = a. [Σ_i component(i)]

    then local = component(i)

Local Moran and Local Geary

Local Moran

- Local Form of Moran's I (Anselin 1995)

  for row-standardized weights
  (such that S_0 and N cancel out in Moran's I)

  variables as deviations from mean (z_i)

  $I_i = (z_i / m_2) \Sigma_j w_{ij} z_j$

  $m_2 = \Sigma_i z_i^2$ does not vary with i, thus constant

- $I_i = (1 / m_2) z_i \Sigma_j w_{ij} z_j = c. z_i \Sigma_j w_{ij} z_j$

- Link Local-Global

$\sum_i I_i = N.I$

or: $I = \sum_i I_i / N$

global Moran is average of local Moran statistics

- Inference

  analytical or computational

  analytical approximation is poor (do not use)

  computational based on conditional permutation

- Conditional Permutation

  conditional upon value observed at i

  hold value at i fixed, random permute remaining n-1 values and recompute local Moran

  repeat many times to obtain reference distribution

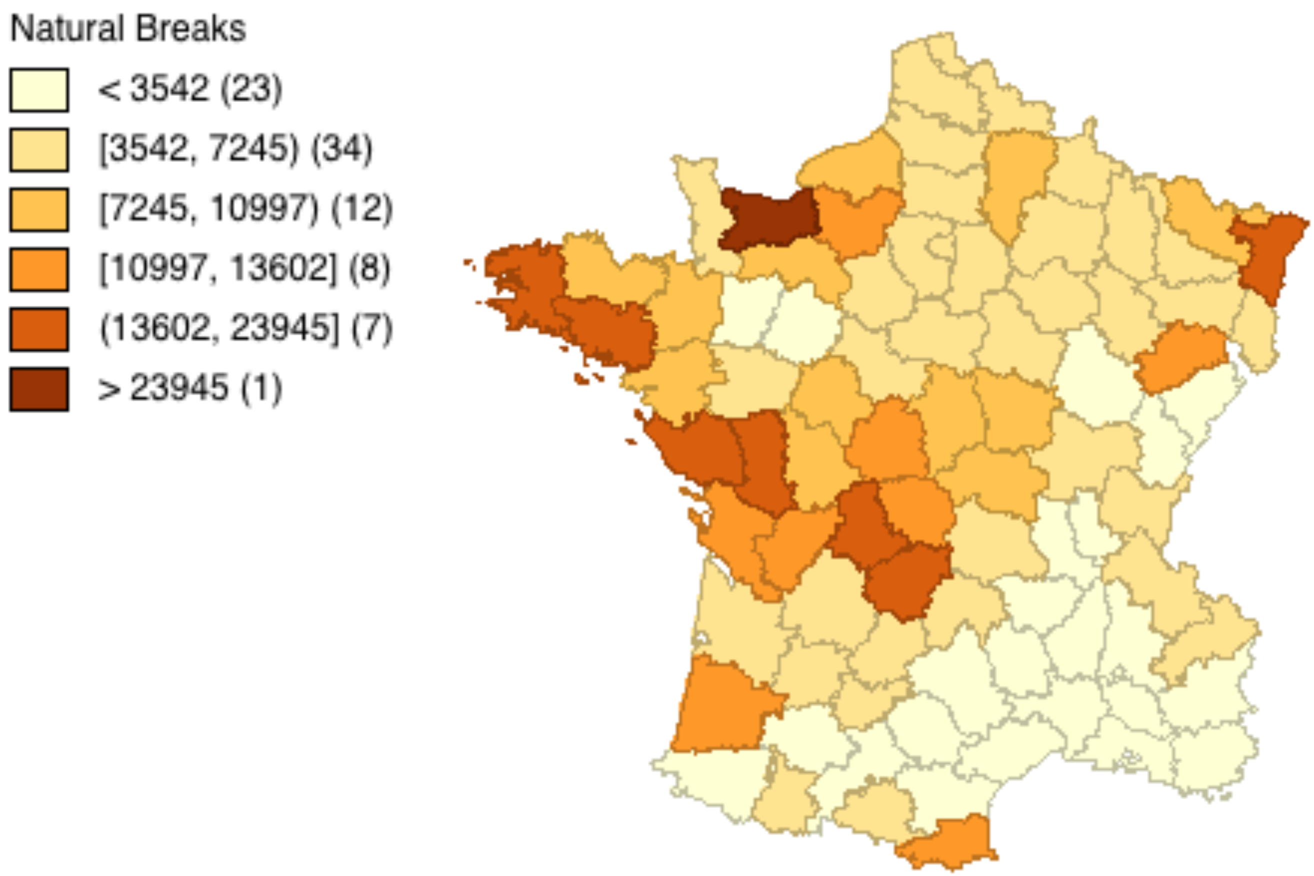
- conditional permutation for each location

- Local Significance Map

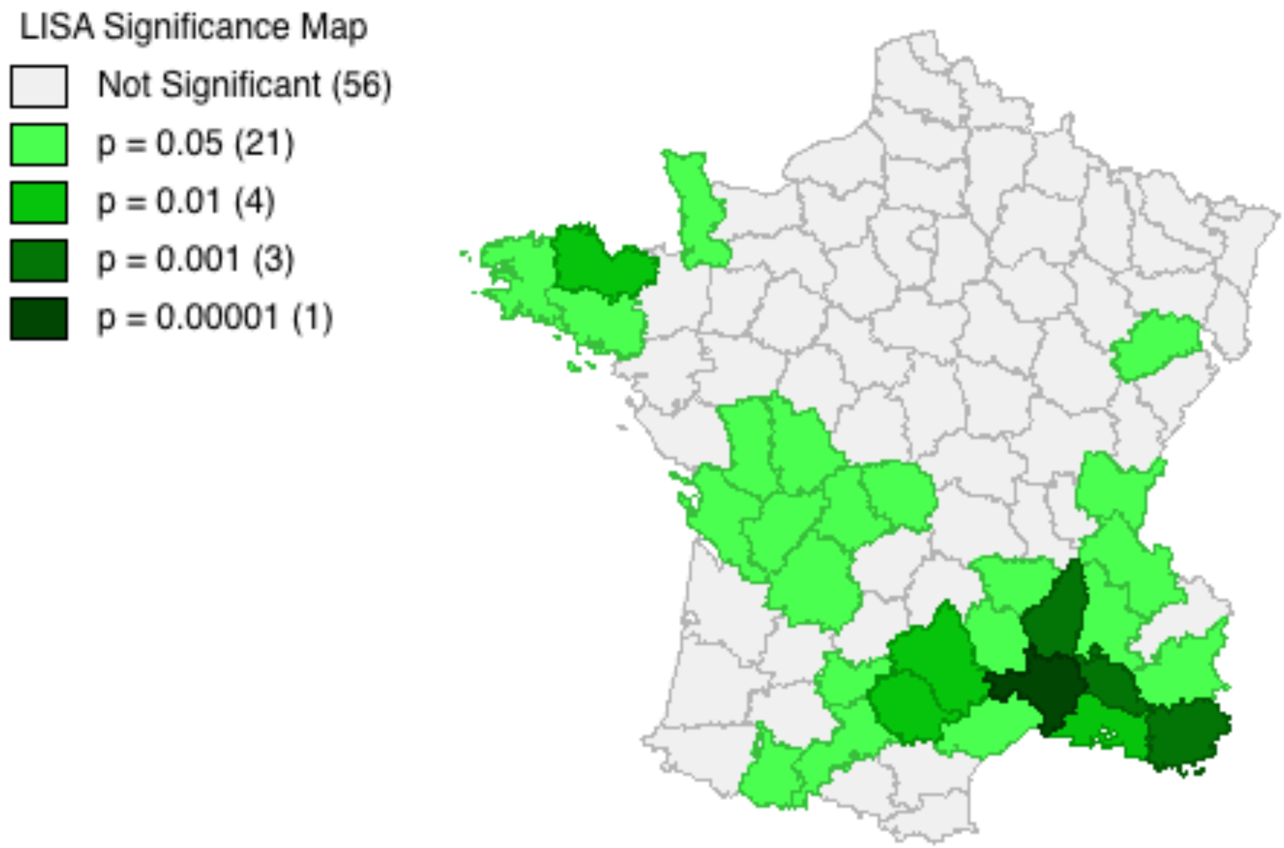
    shows locations with significant local statistic by level of significance

    not very useful for substantive interpretation

    diagnostic for sensitivity of results (for example, when only significant at 0.05)



Guerry data - Donations



local significance map
Guerry data - Donations (queen)

- Local Cluster Map

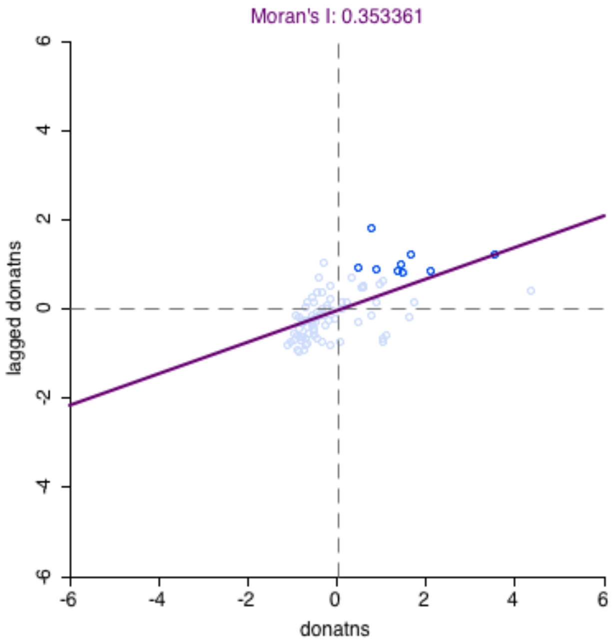
    shows locations with significant local spatial autocorrelation by type of association

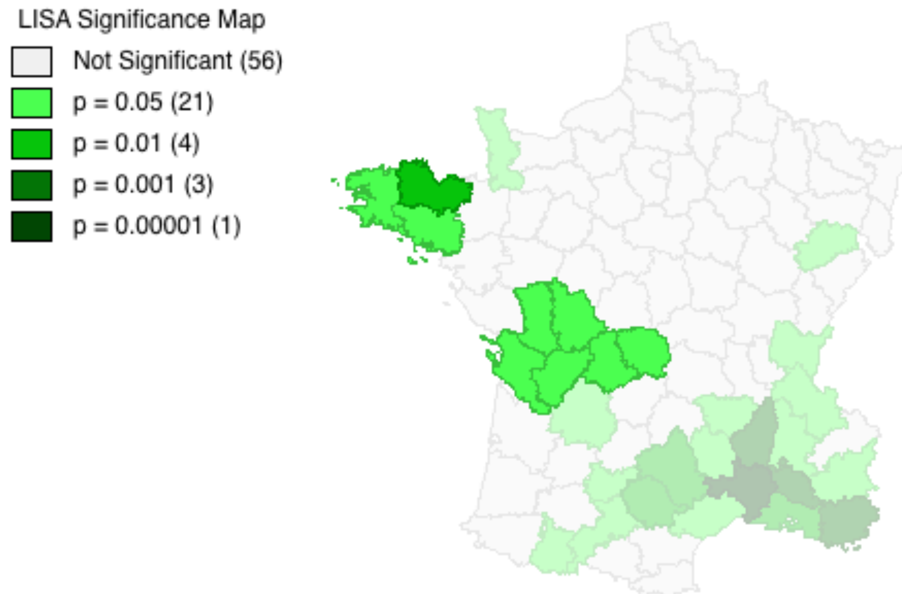
    four color scheme

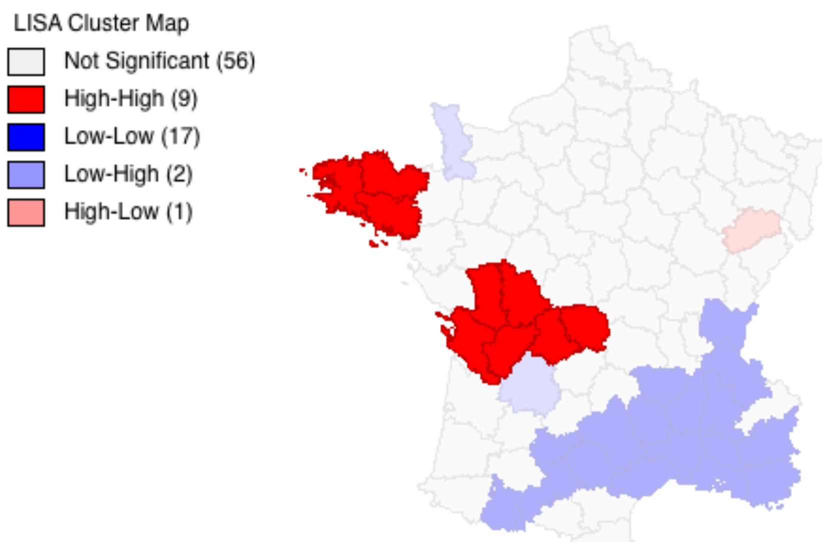
    spatial clusters: high-high and low-low

    spatial outliers: high-low and low high

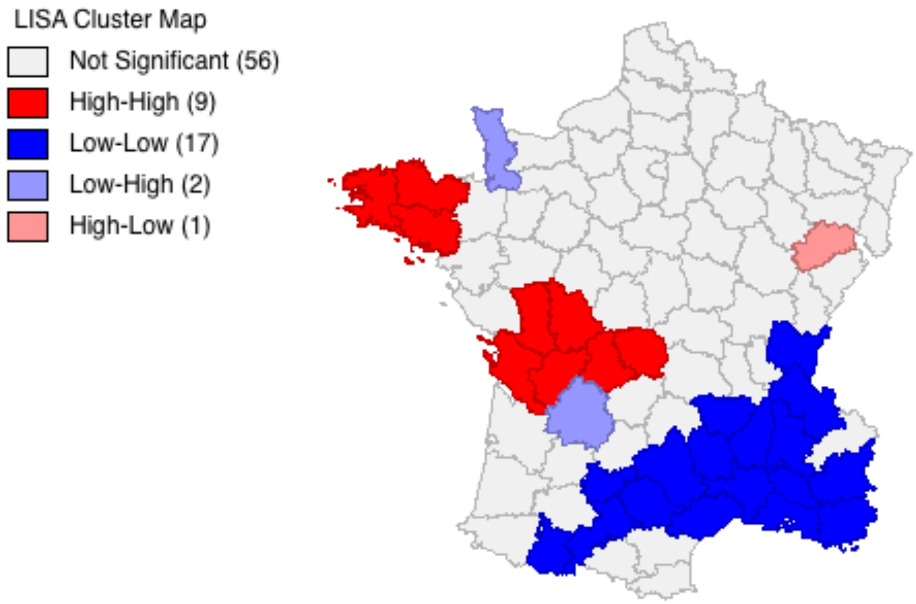
    shown for a given level of significance (sensitivity analysis)



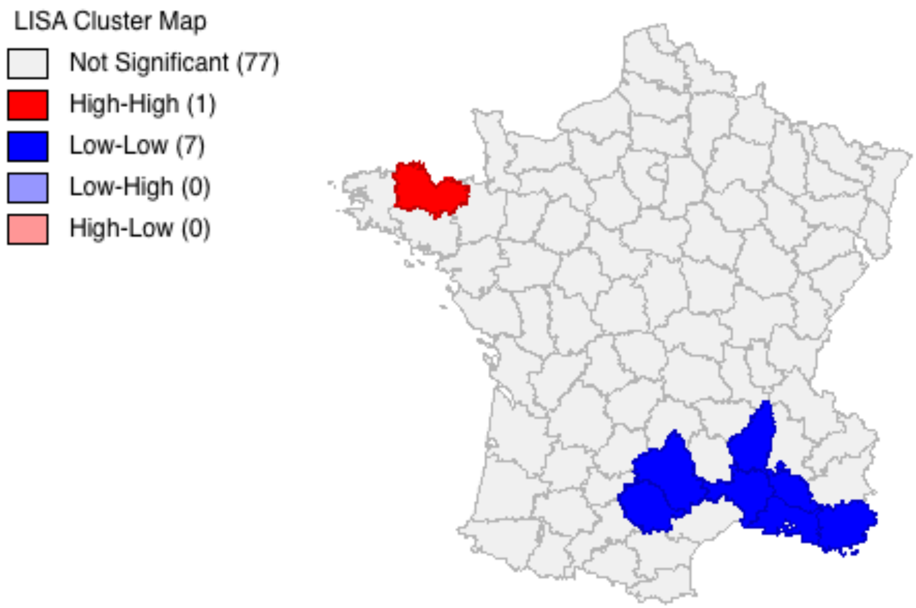




high-high significant locations

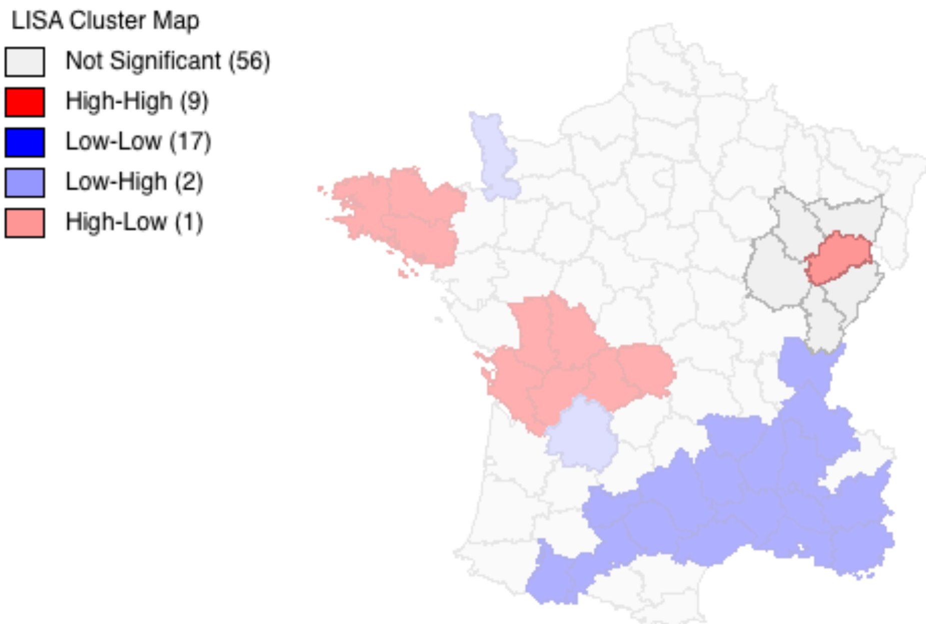


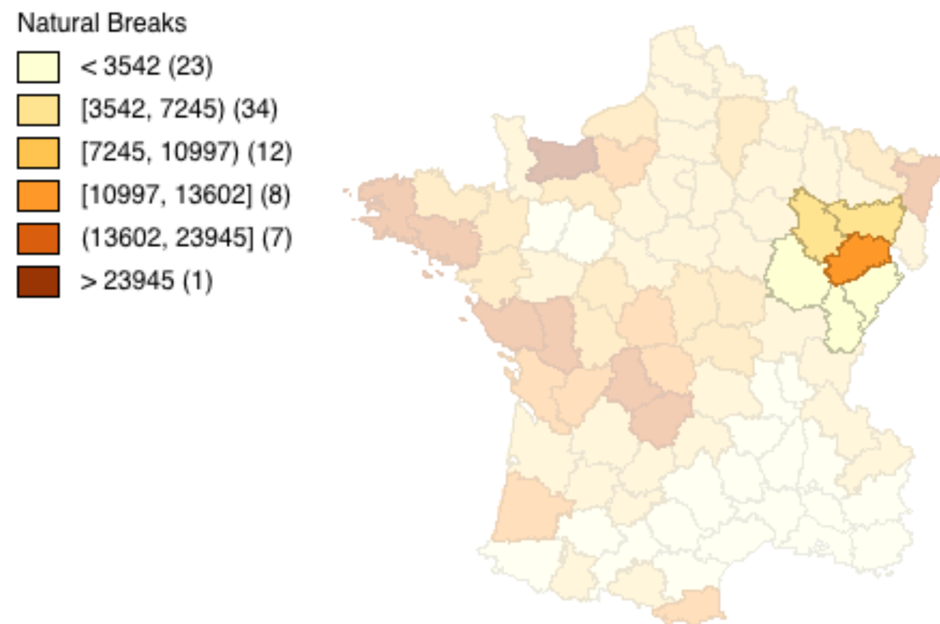
$p < 0.05$



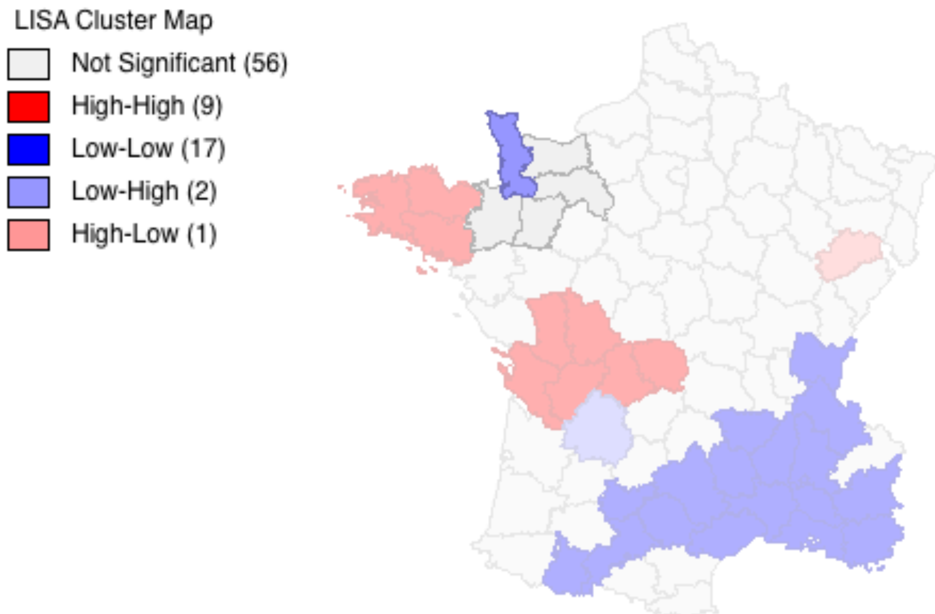
$p < 0.01$

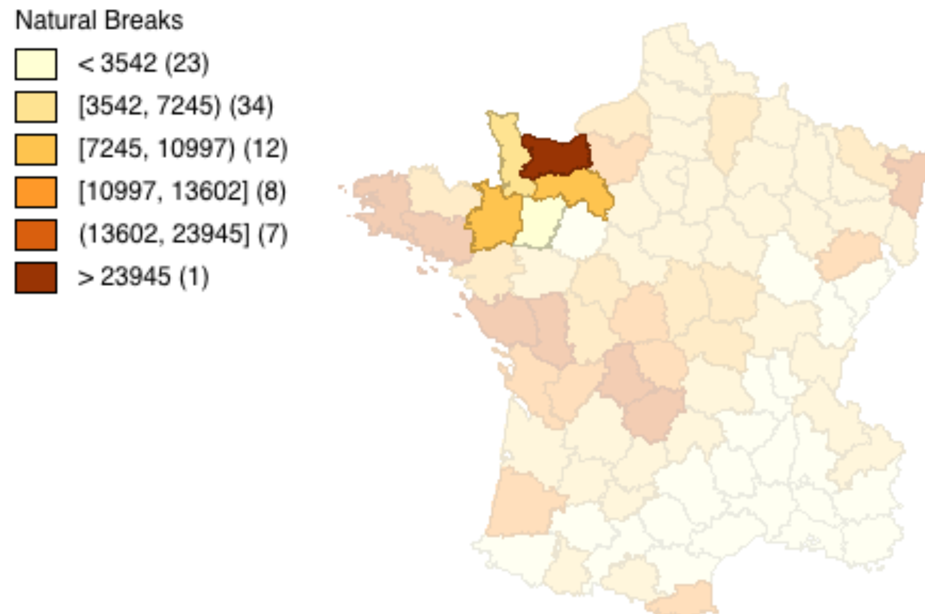
local cluster map for different p-values





high-low spatial outlier





low-high spatial outlier

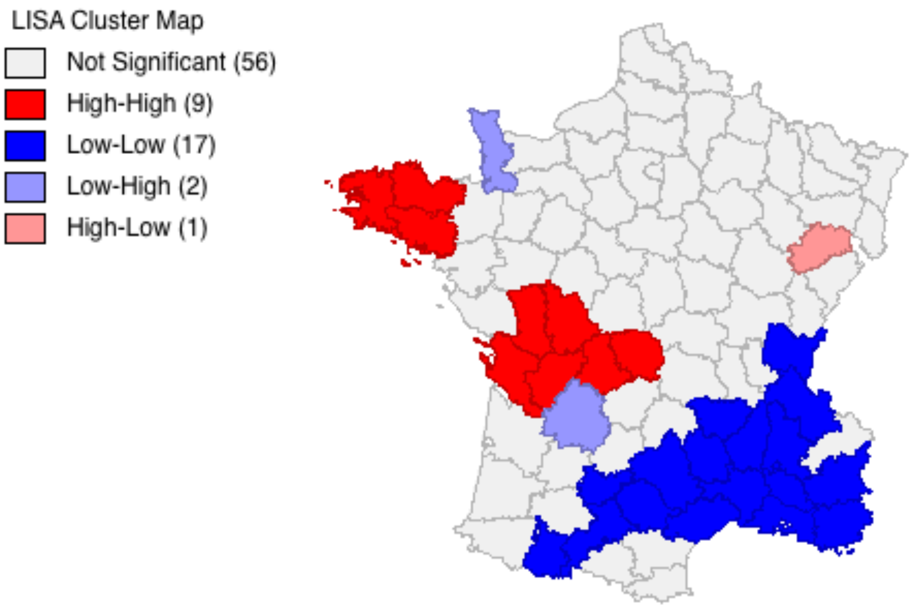
- What is a Cluster?

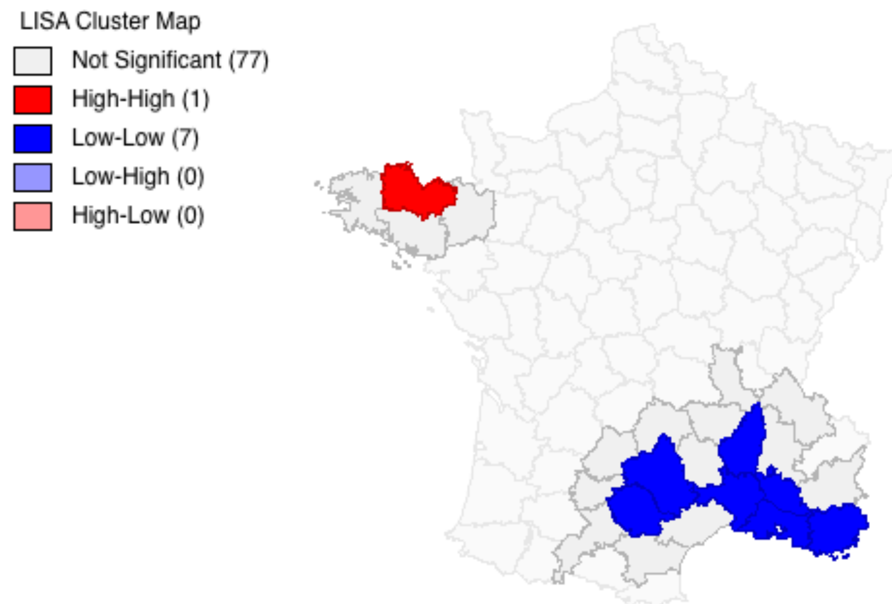
  locations with significant positive local spatial autocorrelation are the core of a cluster

- actual "cluster" includes neighbors as well as core

  regions of high/low values rather than individual locations

Cluster Cores





local cluster map for $p < 0.05$ compared to $p < 0.01$ with neighbors

Local Geary

- Geary's c

global version (Geary 1954)

$$c = \frac{\sum_i \sum_j w_{ij}(x_i - x_j)^2 / 2S_0}{\sum_i (x_i - \bar{x})^2/(n-1)}.$$

local version (Anselin 1995)

$$c_i = (1/m_2) \sum_j w_{ij}(x_i - x_j)^2, \qquad c_i = \sum_j w_{ij}(x_i - x_j)^2.$$

- Moments of Local Geary c

    randomization assumption (Sokal et al 1998)

    $E[c_i] = 2nw_i / (n - 1)$

    for row-standardized weights and standardized variable

    $E[c_i] = 2$

    practical inference: conditional permutation

- Interpretation

  $(x_i - x_j)^2$

  attribute dissimilarity

  distance in attribute space

  $\Sigma_j w_{ij} (x_i - x_j)^2$

  weighted average of distances in attribute space to neighbors in geographic space

- Interpretation (2)

    significant and less than mean

        similarity

    significant and greater than mean

        dissimilarity

- Categories of Association

    combine with Moran scatterplot

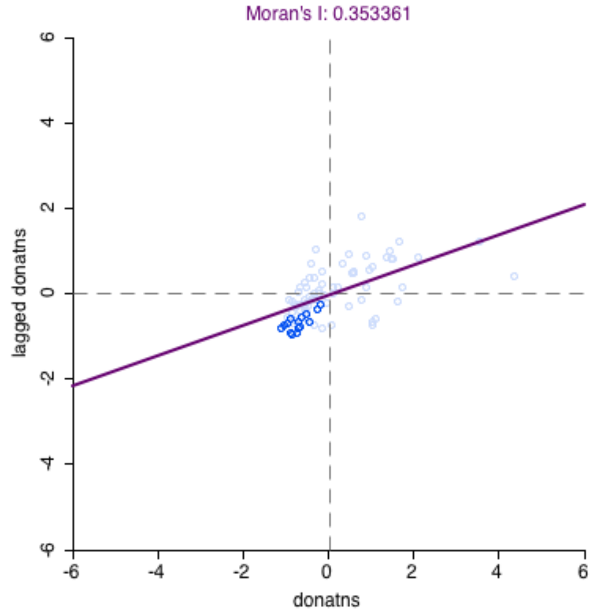
        but: scatter plot is for cross-product association, not squared difference

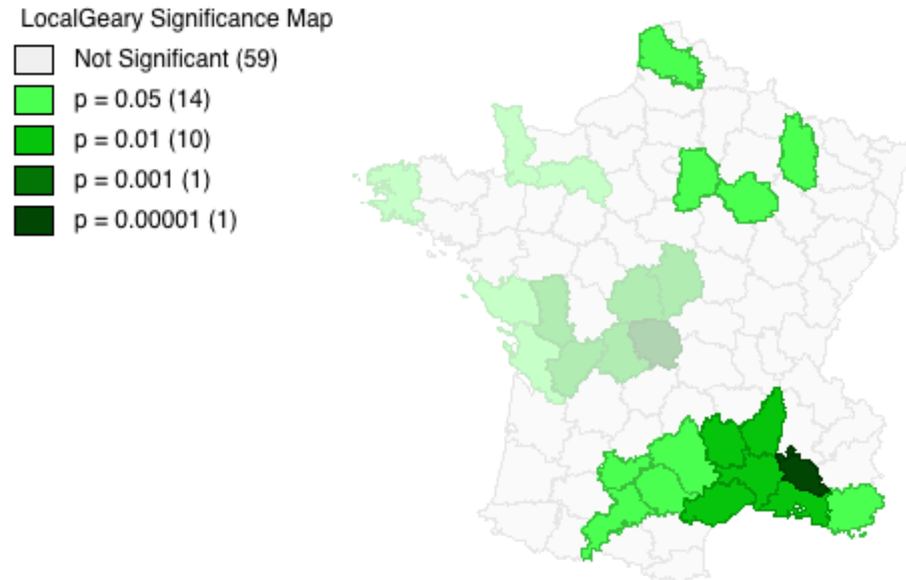
    positive - similarity

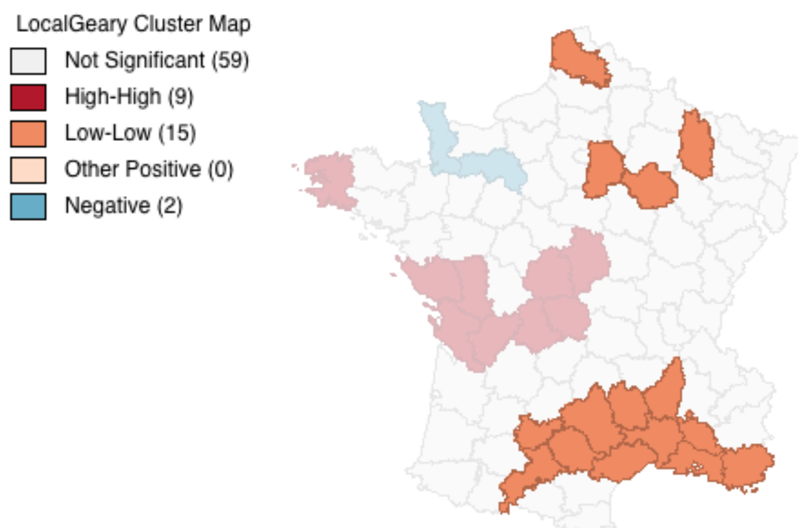
        high-high, low-low and other

    negative - dissimilarity

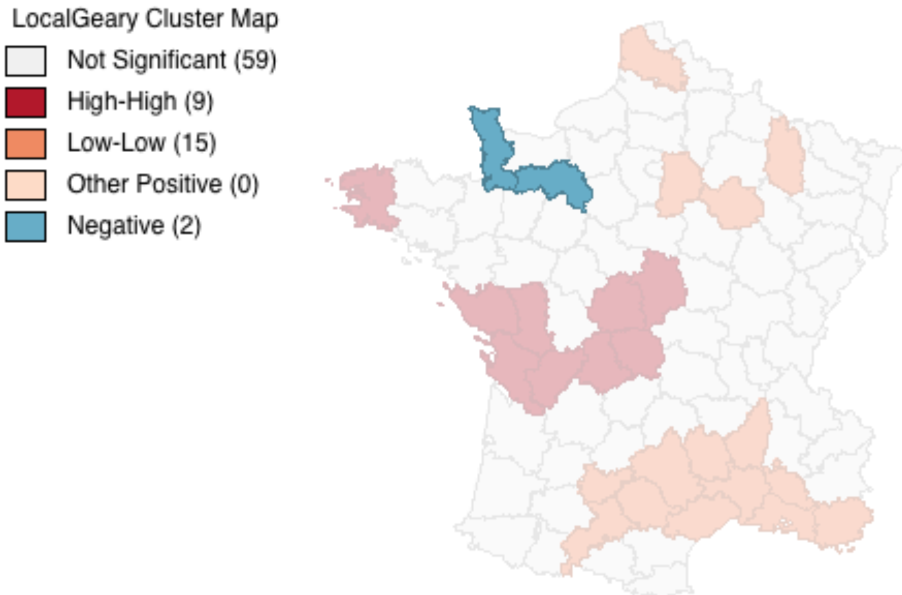
        no distinction between high-low and low-high

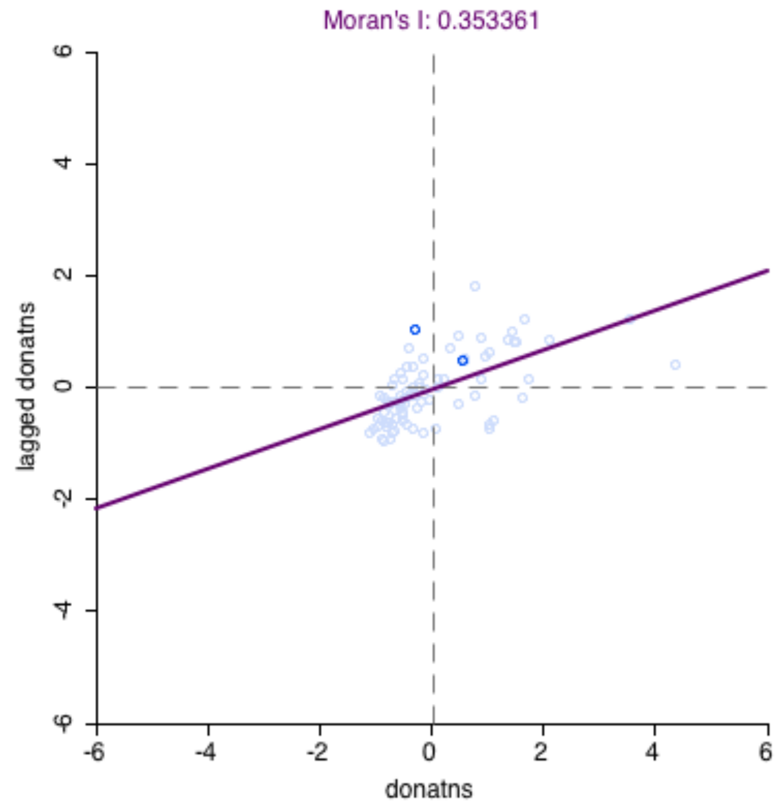




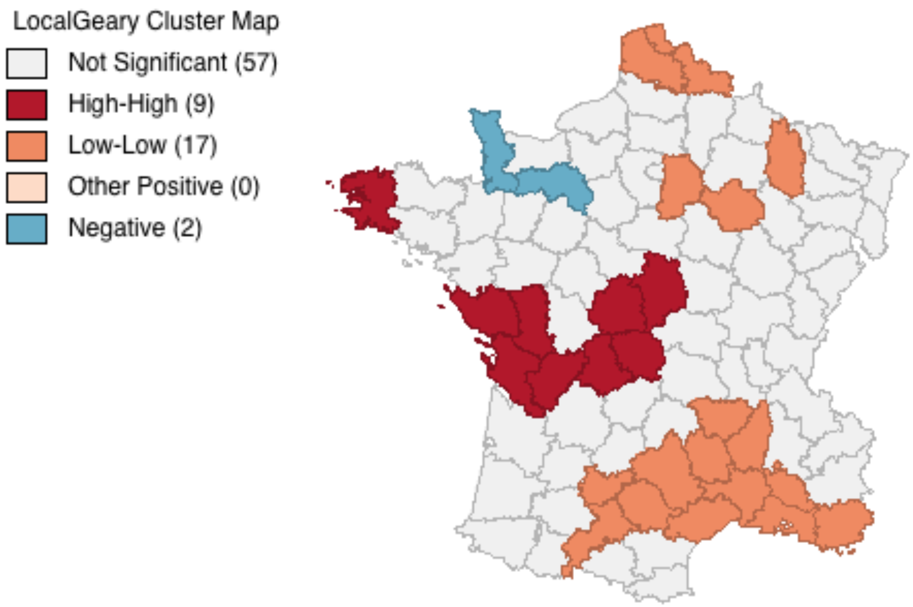


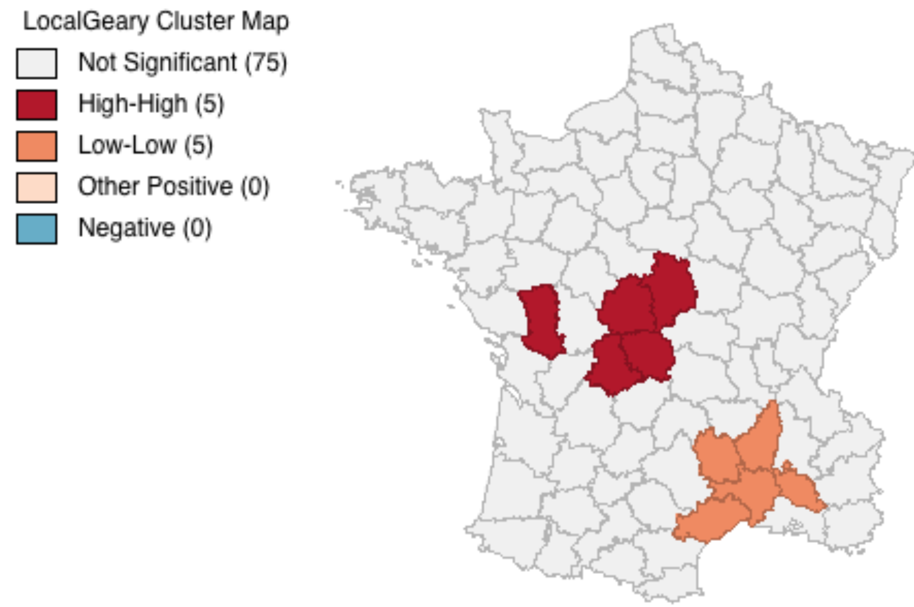
low-low significant locations





local Geary - negative spatial autocorrelation

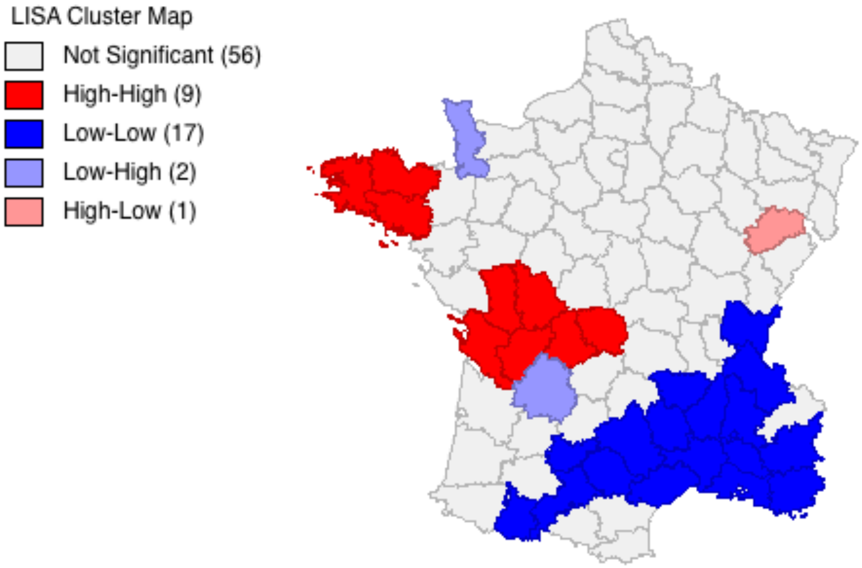


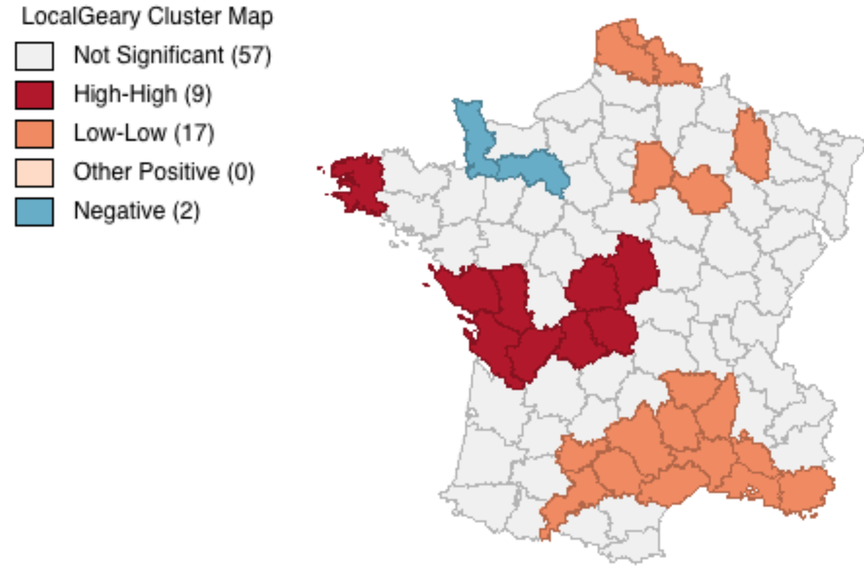


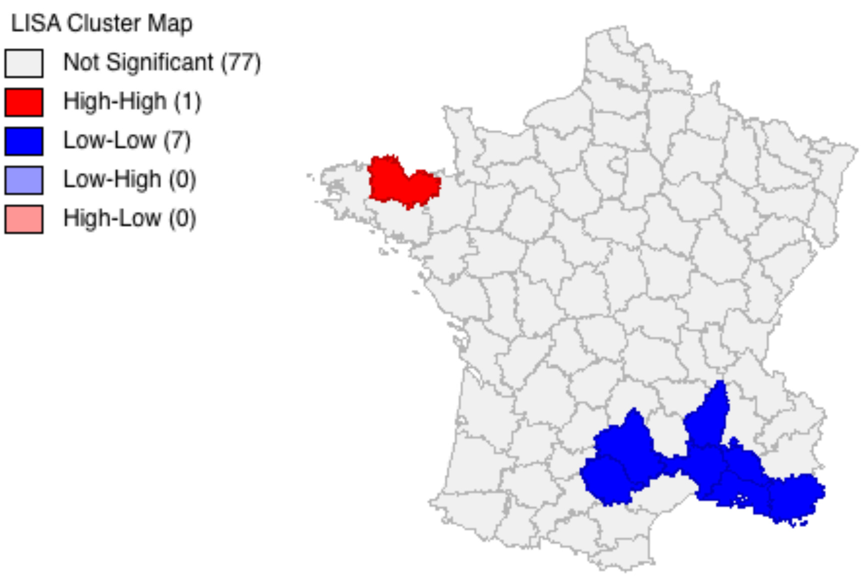
$p < 0.05$ $p < 0.01$

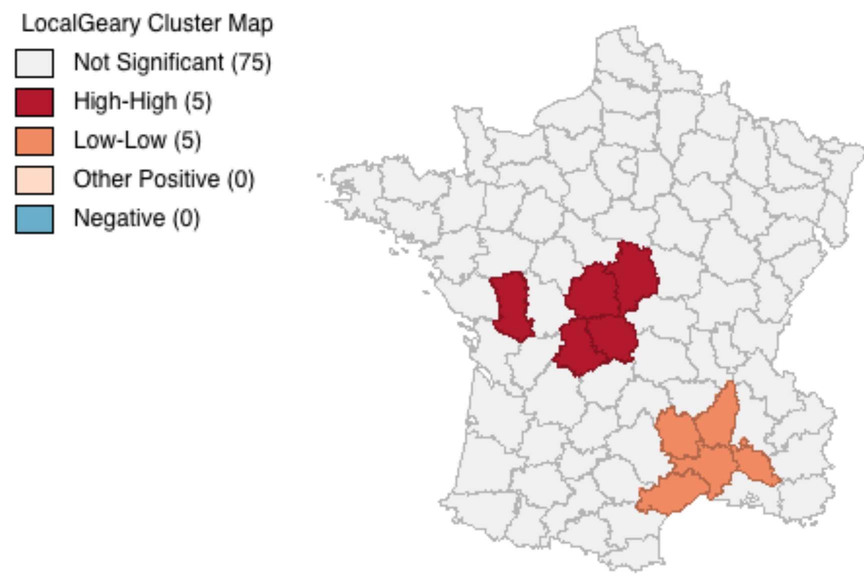
Local Geary cluster map (Guerry data - donations)

Local Moran and Local Geary Compared









Local Moran vs Local Geary - donations

- Comparison

different type of attribute similarity

cross-product (correlation) vs
squared difference (dissimilarity/semi-variogram)

power against different alternatives

but alternatives are unknown

Getis-Ord Statistics

- Local G Statistic

  Getis-Ord (1992) and Ord-Getis (1995)

  not a LISA in a strict sense (no local-global connection) but useful for detecting clusters

  based on point pattern analysis logic

  two versions: G_i and G_i^* (value at i included)

G_i Statistic

$G_i = \Sigma_j w_{ij}x_j / \Sigma_j x_j$ for j not equal i
i not included in either numerator or denominator

numerator is weighted average of neighbors (spatial lag)

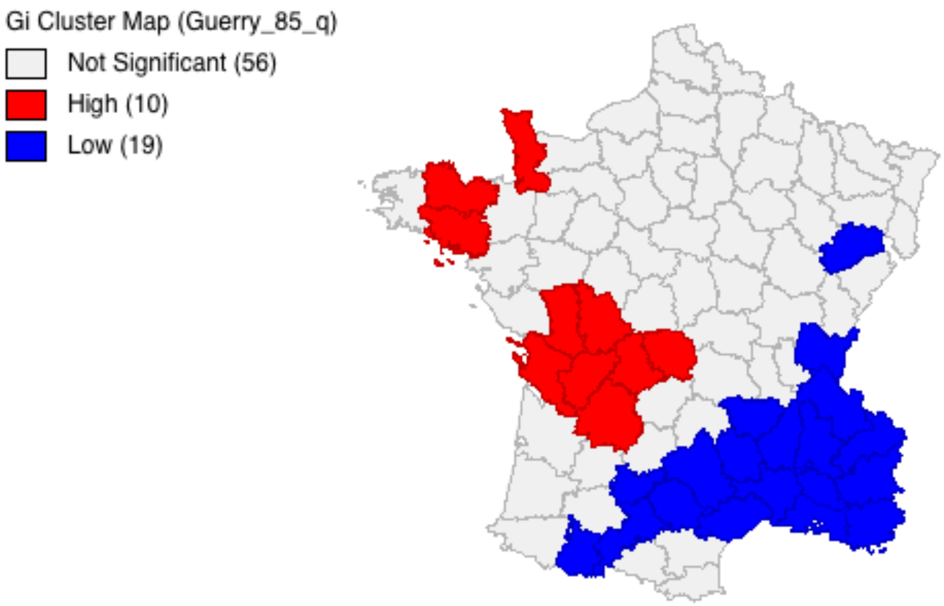
denominator is sum of all values, excluding the value of x at i

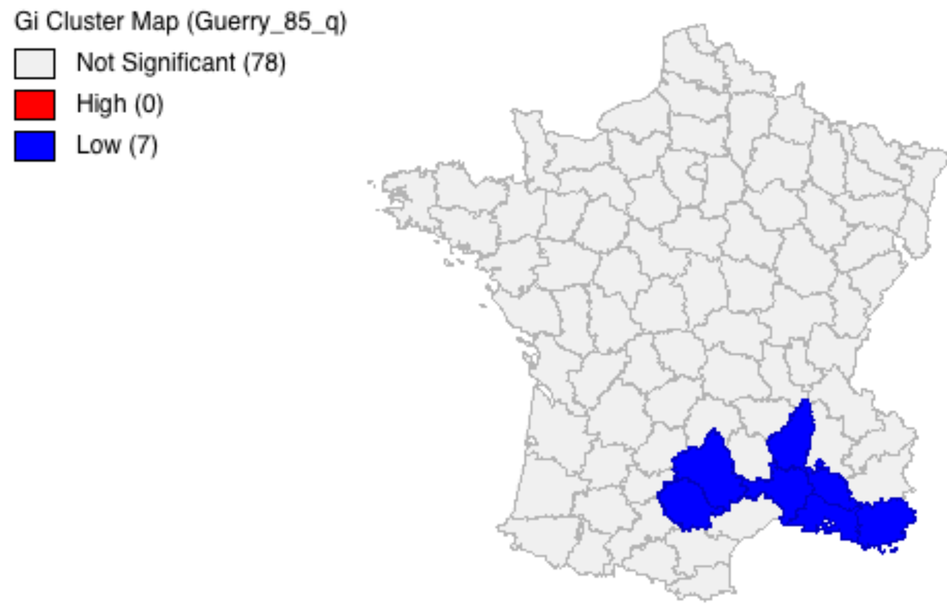
- Inference

  analytical: based on an approximation

  not very reliable

  conditional permutation inference: same principle as for local Moran





$p < 0.05$

$p < 0.01$

Gi statistic cluster map

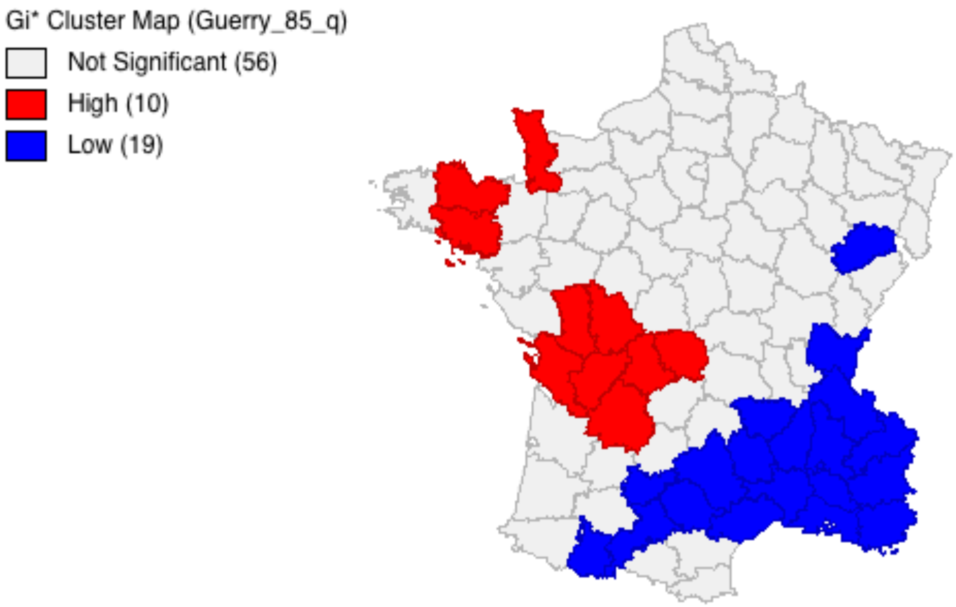
G_i^* Statistic

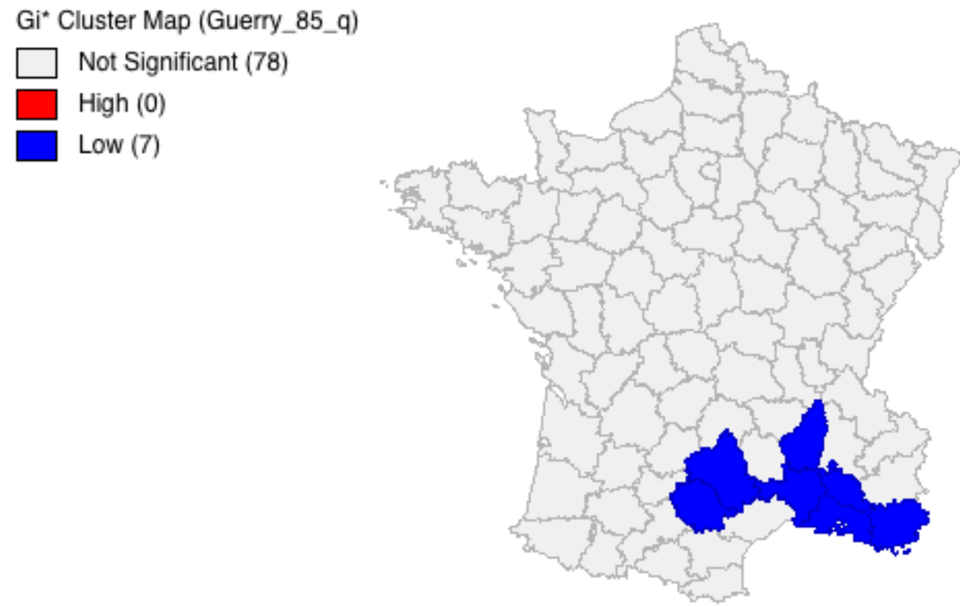
$G_i^* = \Sigma_j \, w_{ij} x_j \, / \, \Sigma_j \, x_j$ for all j

i included in both numerator and denominator

numerator is weighted average of neighbors and value at i (need to define w_{ii})

denominator is sum of all values, thus constant





$p < 0.05$

$p < 0.01$

G*i statistic cluster map

- Interpretation

  significant values only - ignore others

  positive G_i (G_i^*) = local clustering of high values
  hot spot

  negative G_i (G_i^*) = local clustering of low values
  cold spot

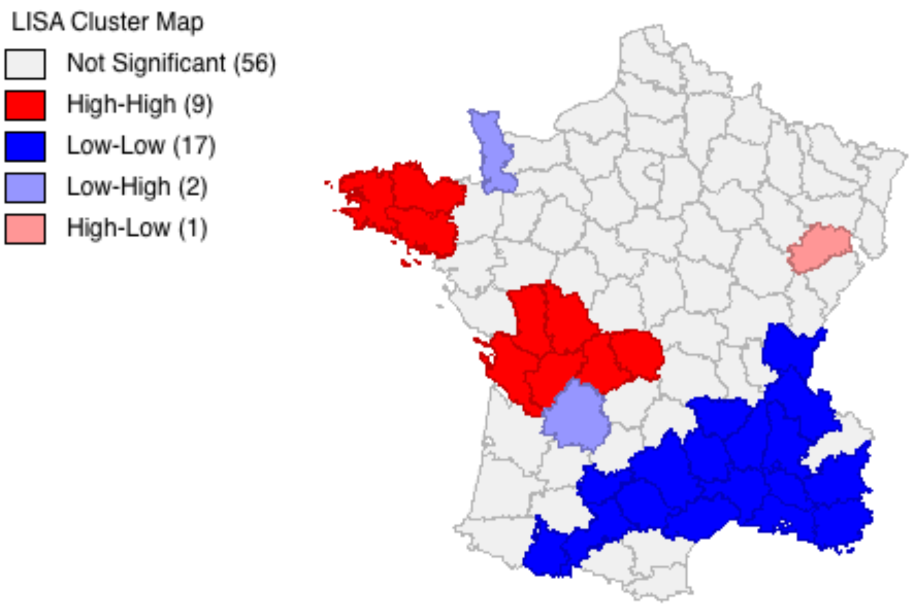
  does NOT detect spatial outliers

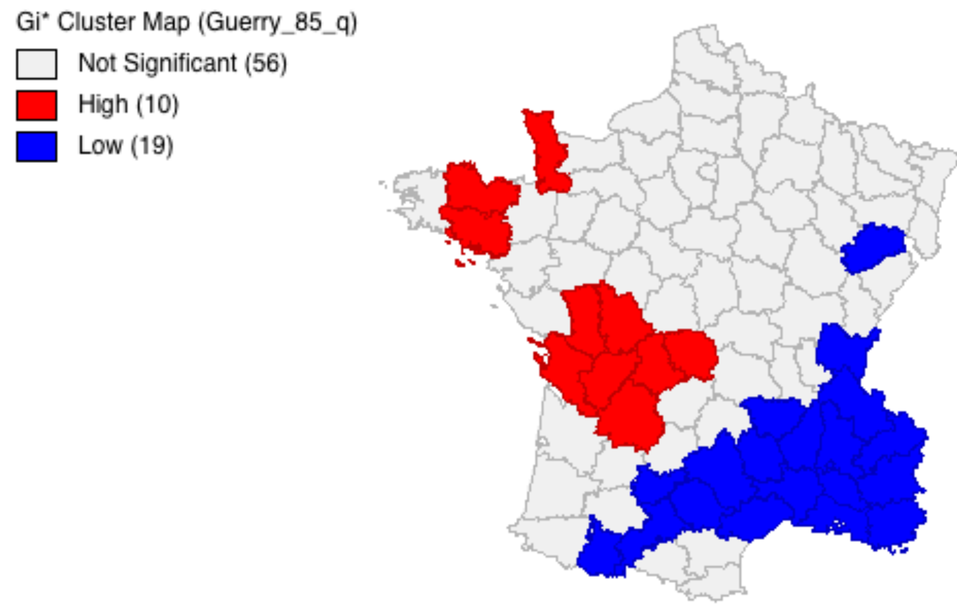
- Local Moran vs. G Statistics

  G statistics useful when negative spatial autocorrelation is negligible (then hot spots and cold spots)

  G statistics do not consider spatial outliers, local Moran does

  Local Moran needs to be combined with classification of type of spatial autocorrelation





local Moran

local Gi*

local Moran's I compared to local Gi*

Interpretation

- Inference

  exact and asymptotic inference

  Sokal et al (1998), Tiefelsdorf (2002)

  the multiple comparison problem

  de Castro and Singer (2006) [false discovery rate]

  local inference in the presence of global autocorrelation

  Ord and Getis (2001), Rogerson (2015)

  power calculations

  Bivand et al (2009)

- Multiple Comparisons

significance level for a given location assumes only that location is being analyzed

because all locations are analyzed, individual p-value is incorrect (too low)

various corrections (e.g., Bonferroni bounds, false discovery rate) but none fully satisfactory

in practice: cautious interpretation

- Multiple Comparisons in Practice

  target significance level α

    FWER: family wide error rate

    the probability of making one false rejection out of k comparisons

  what should α be?

    new recommendation, NOT 0.05 but 0.005?

- Bounds on Type I Error

    Bonferroni bounds

        use α/k

        in local statistics k = n, too large

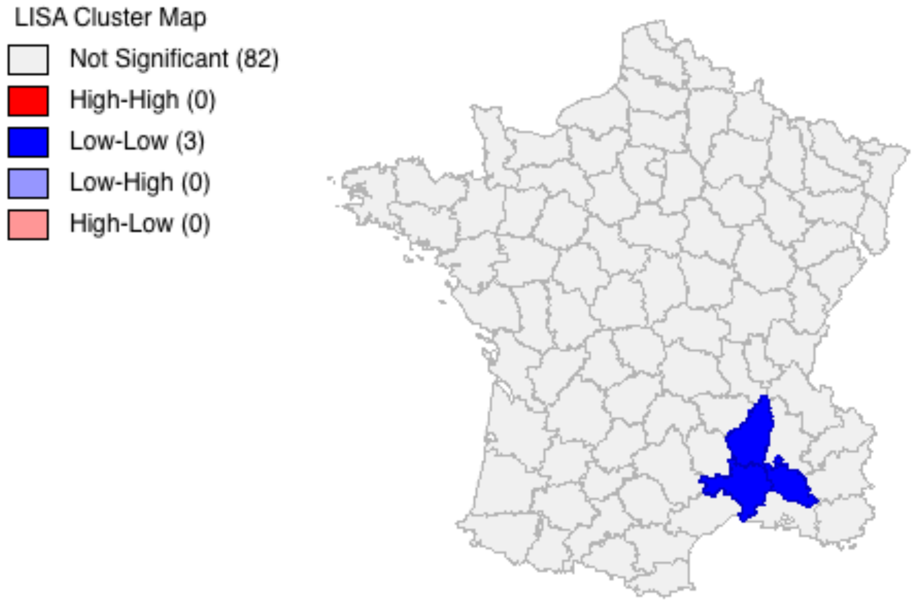
    Sidak bounds

        use $1 - (1 - \alpha)^{1/k}$

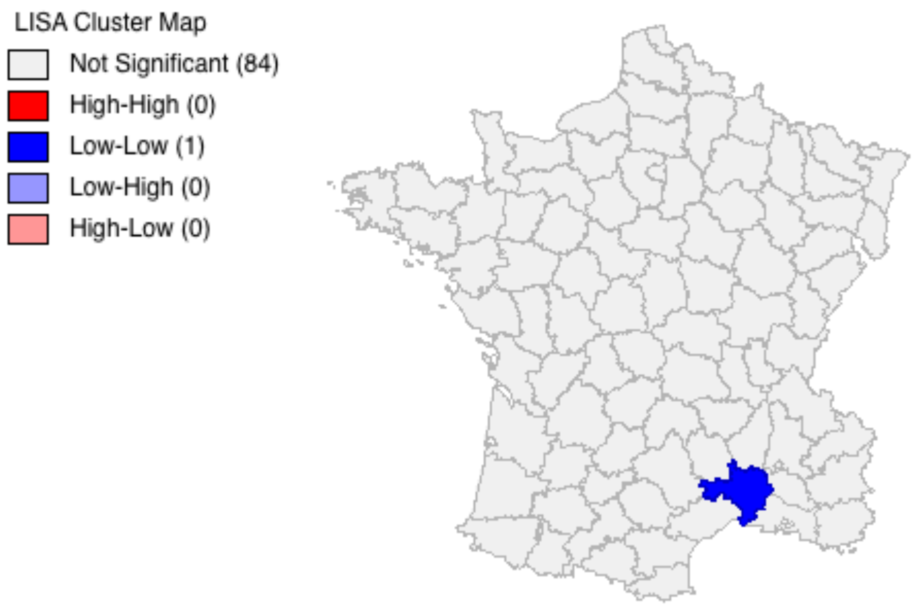
    practical issue

        what should k be?

        adjustment for overlap



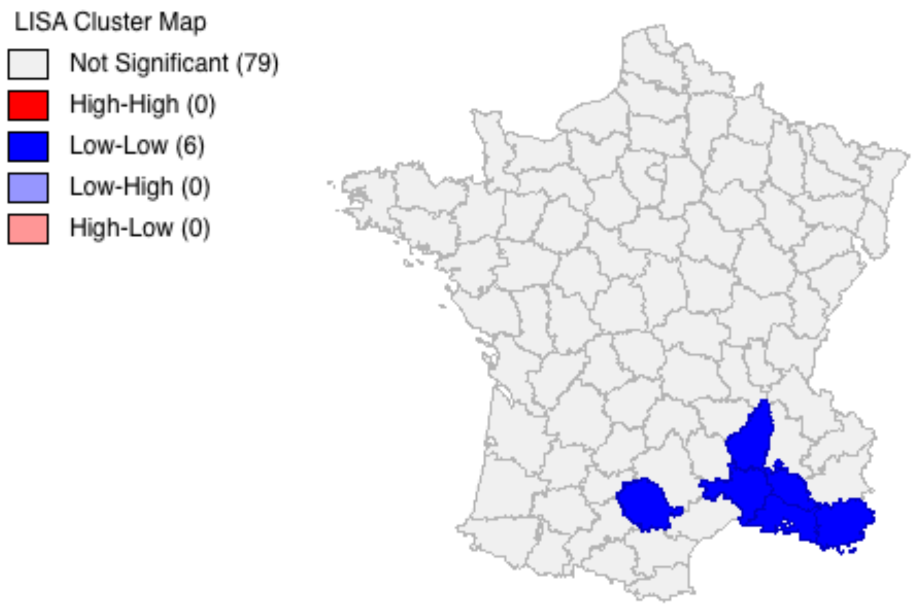
$\alpha = 0.05 \; p < 0.0006$



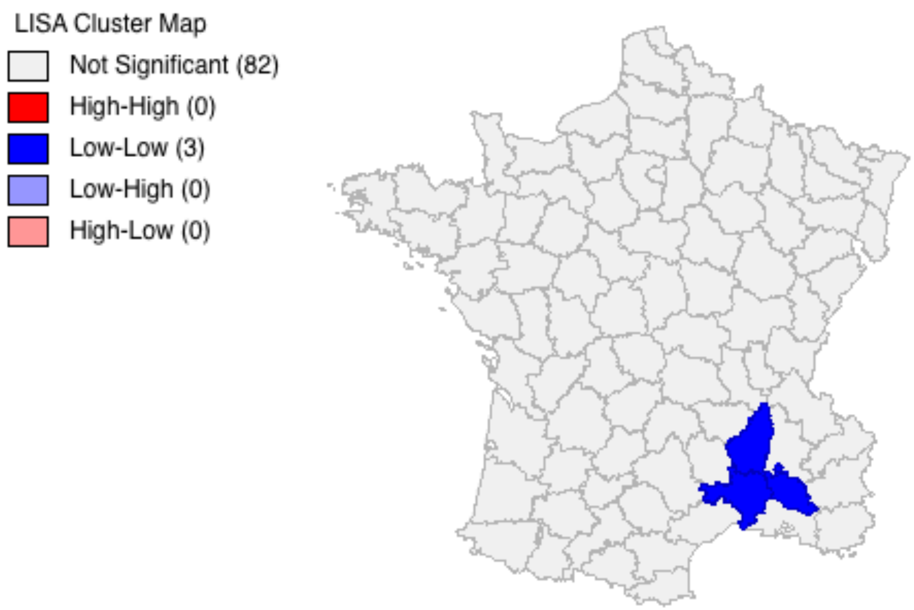
$\alpha = 0.01 \; p < 0.0001$

Bonferroni bounds

- False Discovery Rate
  - Benjamini and Hochberg (1995)
  - three step process
    - sort p-values for each observation, p(i) from smallest to largest
    - select imax as i such that $p(i) \leq (i/N)\alpha$
    - all observations with $i \leq$ imax are "significant"
  - Efron and Hastie (2016)
    - not significant but interesting



$\alpha = 0.05$ $p < 0.0035$



$\alpha = 0.01$ $p < 0.00035$

False Discovery Rates (FDR)

- Exploratory Only

  LISA clusters and outliers are identified, but not explained

  suggests interesting locations

  multiple processes can yield the same pattern

    inverse problem

- Univariate Only

  univariate spatial autocorrelation can be due to other covariates

  univariate analysis ignores multivariate interactions

  scale mismatch can create impression of clusters without a meaningful process interpretation

Extensions

- Extensions
  - categorical data
    - Boots (2003, 2006)
  - points on networks
    - Yamada and Thill (2007)
  - optimal spatial weights
    - Getis and Aldstadt (2004), Aldstadt and Getis (2006), Rogerson (2010), Rogerson and Kedron (2012)
  - space-time, income mobility
    - Rey (2016)

- Computation

    conditional permutation calculations

        Lee (2009), Hardisty and Klippel (2010)

    software implementations

        open source: GeoDa, PySAL, R

        commercial: ESRI ArcGIS, Carto

Multivariate Local Geary

- Multivariate Spatial Autocorrelation
    - difficulty combining multi-attribute similarity and locational similarity
        - confusion with in-place correlation
        - distance in multi-dimensional attribute space
    - extension of global Moran's I to principal components
        - Wartenberg (1985), Dray et al (2008), Dray and Jombart (2011)
    - bivariate case
        - distinction correlative and spatial correlation (Lee 2001)

- Two-Variable Case

attribute dissimilarity

distance in attribute space

$$d_{ij}^2 = (z_{1,i} - z_{1,j})^2 + (z_{2,i} - z_{2,j})^2$$

weighted attribute distance to neighbors

$$\begin{aligned}
\sum_j w_{ij} d_{ij}^2 &= \sum_j w_{ij} [(z_{1,i} - z_{1,j})^2 + (z_{2,i} - z_{2,j})^2] \\
&= \sum_j w_{ij} (z_{1,i} - z_{1,j})^2 + \sum_j w_{ij} (z_{2,i} - z_{2,j})^2 \\
&= c_{1,i} + c_{2,i}
\end{aligned}$$

- Multivariate Local Geary

in general, weighted attribute distance to geographic neighbors as sum of $c_{v,i}$

$$c_{k,i} = \sum_{v=1}^{k} c_{v,i},$$

standardized to keep similar scale to univariate local Geary

$$c_{k,i} = \sum_{v=1}^{k} c_{v,i}/k.$$

- Inference and Interpretation
  - conditional permutation
    - multiple comparison problem
    - correct for number of variables
  - similarity and dissimilarity only
    - multivariate interaction too complex for high-high and low-low interpretation

- Multivariate Local Geary Cluster Map

locations of those places that are similar or dissimilar in multivariate attribute space to their geographic neighbors

close/far points in multivariate attribute space are close in geographic space

tension between attribute similarity and locational similarity common to any multivariate spatial clustering

- Example: SIDS NC counties (n = 100)

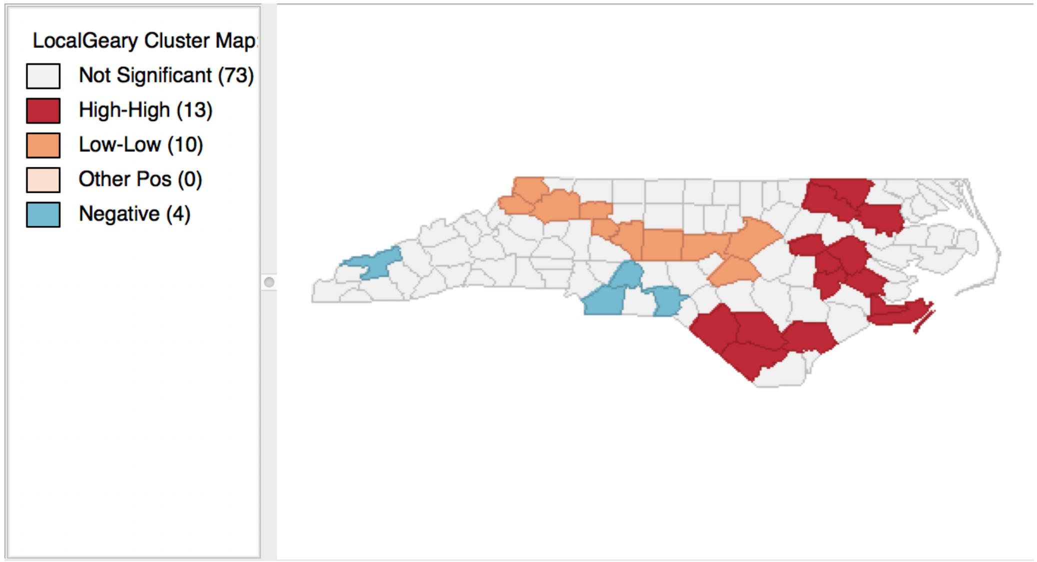
    - Cressie (1992) and many other application

        - illustration of Getis-Ord local statistics

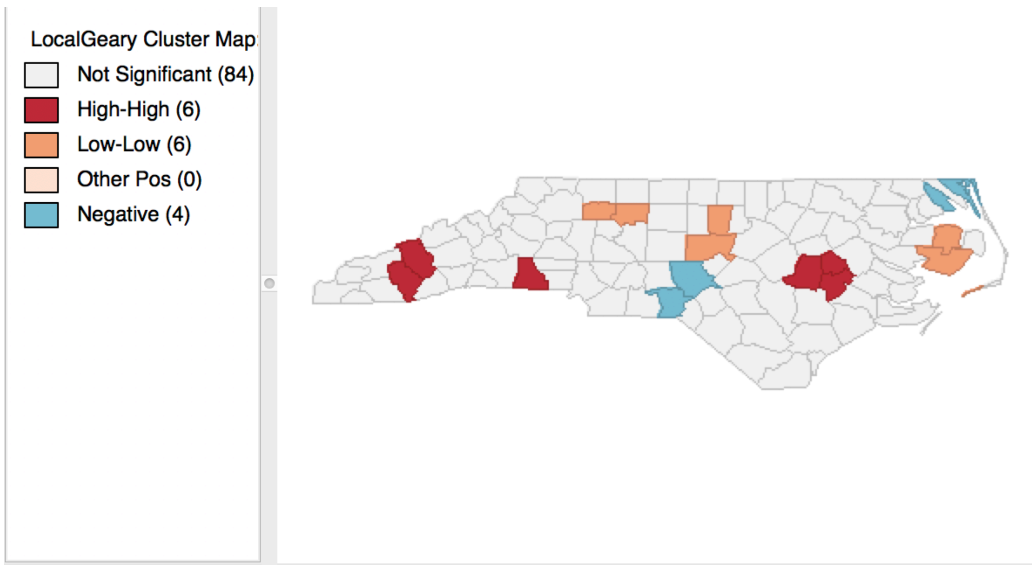
        - SIDS rates 1974 and 1979

    - very low global spatial autocorrelation

        - not significant for SIDS 70

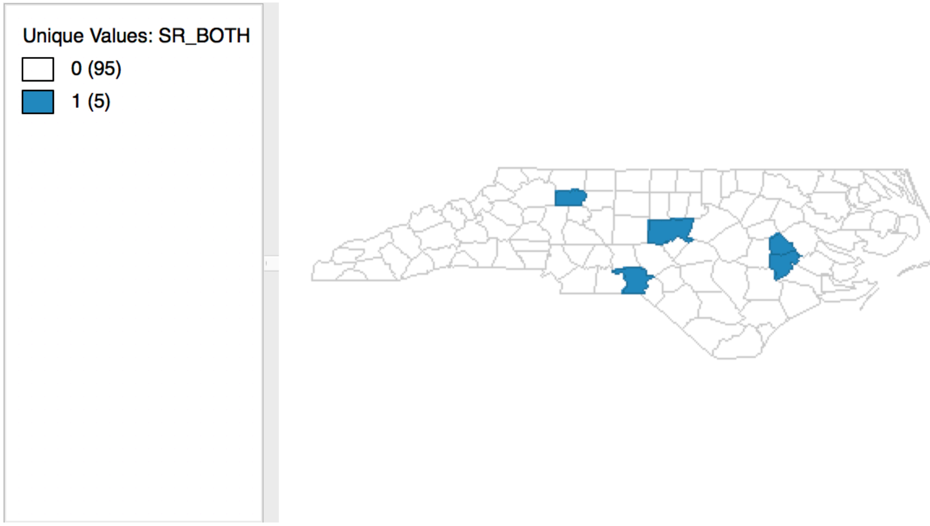


SIDS 74

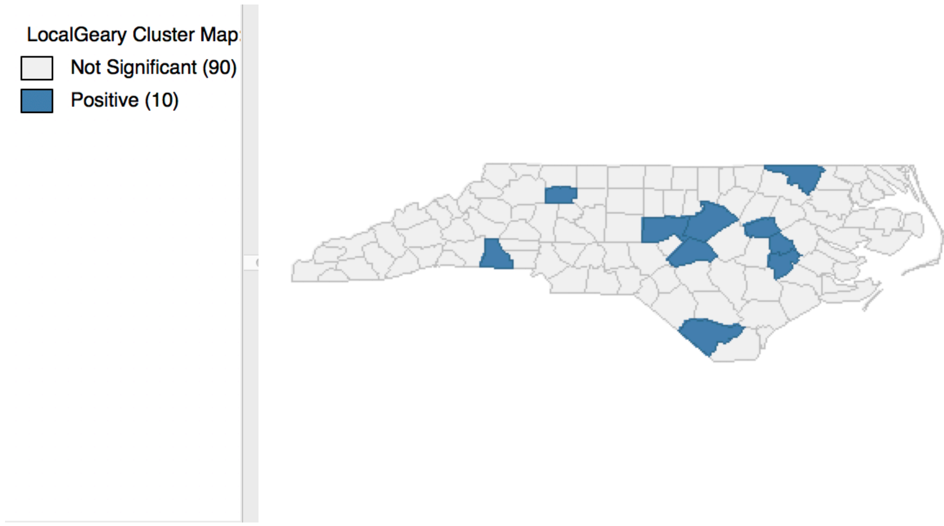


SIDS 79

uni-variate Local Geary cluster map



univariate local Geary significant for both variables



bivariate local Geary cluster map

- Multivariate Local Clusters

    offers alternative insight beyond linear association

    difficult to interpret as number of variables considered increases

    issues with p-value

        multiple comparisons

        multiple variables

    most useful when applied to principal components