Areal Data Spatial Analysis

D G Rossiter

Cornell University

February 15, 2023

Outline

Outline

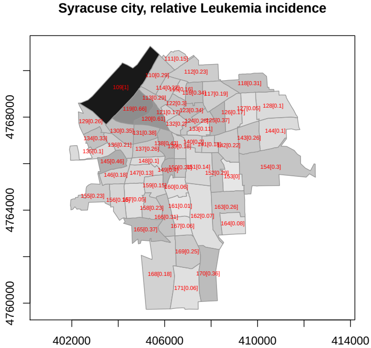
Topic: Areal data

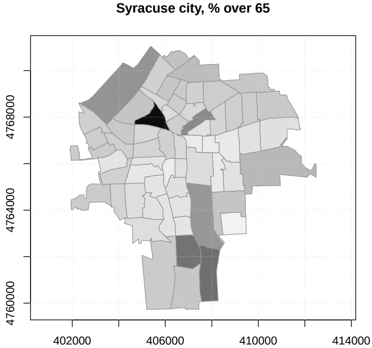
- Data are presented as attributes of **fixed polygonal areas**
  - generally irregularly-shaped, and/or not all same shape
  - examples: census blocks, voting districts, forest parcels ...
  - but methods can apply to regular grids
- Attributes can be analyzed in feature space (distribution, correlation, regression ...) but:
- Q: Is the data-generating **process**:
  - non-spatial (all in feature space),
  - spatial (all in geographic space), or
  - mixed?
- Q: If mixed, how does the spatial structure affect the **feature-space structure**?

Typical applications

- spatial econometrics [2]
- epidemiology [7, §11] [9]
- sociology / demographics [11]
- political science [20]
- natural resources, if data are presented as areal aggregates
  - forest management blocks, farms, ...

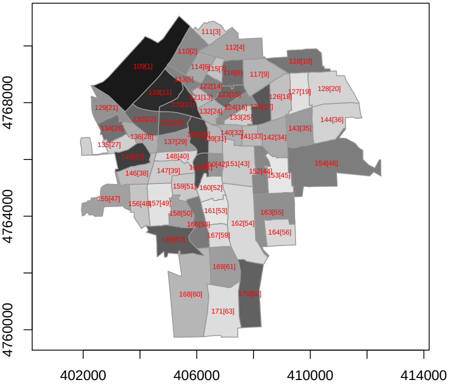
Example: Syracuse (NY) census and health





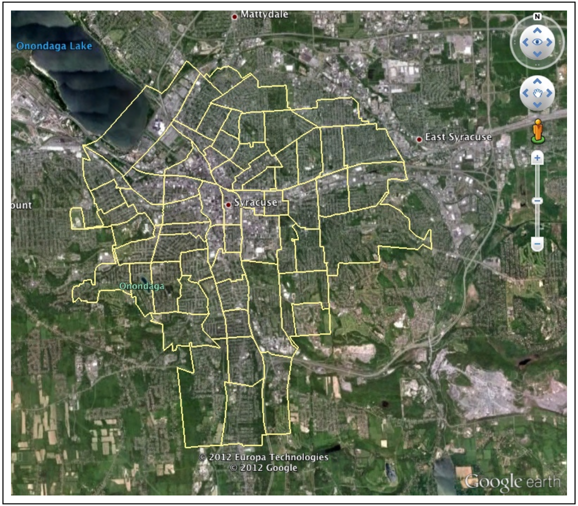
source: Bivand et al. [7, §9]
Q: Is leukemia incidence in a census tract correlated with mean age in the tract?
Or are there local "hot spots" that might have a point-source cause?

Syracuse city, rank of Leukemia incidence



This by `rank`, not **relative incidence**

Real world



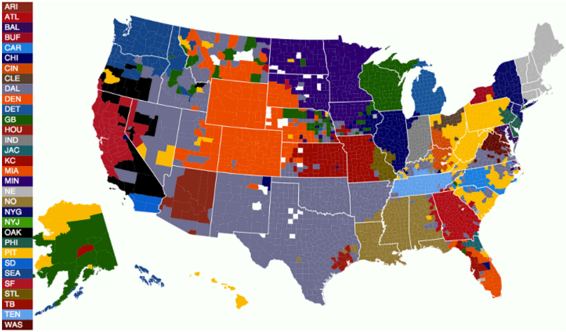
Google Earth view of census tract boundaries (KML file); can zoom and pan

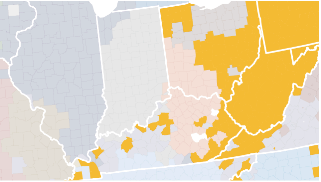
Characteristics of Syracuse leukemia data

Typical of most areal data:

- **Aggregated** by **reporting unit**
  - here, US census tracts; within City boundary
- Units were **not designed for our purpose**
  - here, study of the causes of leukemia
  - size, geographic and feature-space characteristics not what we would have designed
- **Uneven size and shape** of units
  - Different numbers of neighbours, lengths of common borders
- Units on edges have **unobserved neighbours**
- Uneven feature-space "size"
  - e.g., population, proportion residential vs. commercial
- "Points" (e.g., industry) assigned to the whole polygon

Example: Favourite NFL team by county (2012)





source: `https://www.facebook.com/notes/facebook-data-science/nfl-fans-on-facebook/10151298370823859`, 11-Feb-2013.
Source: 35M USA Facebook account holders who "liked" an NFL team in 2012; location is known

Question: What factor(s) determine this in **feature** & **geographic** spaces?

What factors control which team is the favourite?

- Team's success (over what period?)
- Team's games shown on local/national TV?
- Team plays in county's state?
- If no team in state, team plays in neighbouring state?
- Team plays in migrants' home state?
- Proximity/easy transportation of county to team's stadium?
- Proximity to team's training camp site?
- Demographic factors (occupation, ethnicity)?
- Popular players on team from locality/local college?
- Other factors binding a region?
- **Is there residual spatial correlation** after accounting for these factors? "Spillover effect".

Characteristics of areal data – attributes

- The **attributes** relate to the whole area of the polygon, and **can not be further localized**
  - Various methods of **dis-aggregation** using covariates with finer spatial resolution
  - e.g., satellite imagery to separate industrial and residential areas within one polygon
- Often the attributes are **aggregate** measures
  - e.g., population *count*, proportions
- The attributes may already be **normalized** to the area of the polygon
  - e.g., population *density*
- **Metadata** is vital for proper processing and interpretation
  - especially the **aggregation** method from individuals to areas

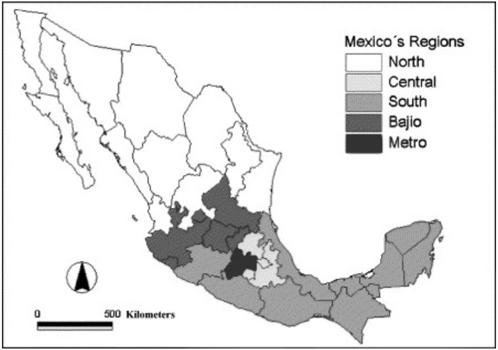
Characteristics of areal data – choice of tessellation

- **tessellation** = division of the study area
  - The tessellation may have been done for a purpose not directly relevant to the analysis
  - E.g., crop yield statistics may be aggregated by political division, but the crop yield may be better modelled by agro-ecological zone
- changes to **boundaries** → ☹ longitudinal analyses
  - British county / authority boundaries
  - Chinese province/autononmous cities/region boundaries
  - Poland/Lithuania/Ukraine/Germany 20thcentury boundaries
  - area code zones, census tracts ...
- again, **Metadata** is vital for proper processing and interpretation

Characteristics of areal data – choice of scale

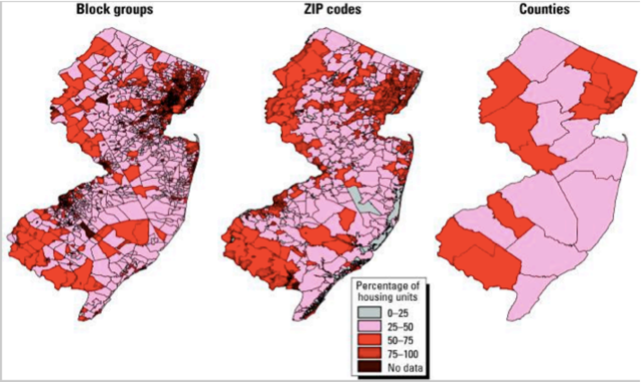
- The **scale** of the tessellation affects the analysis
  - a variation of the **bandwidth** problem for spatial fields
  - e.g., voting patterns by state vs. congressional district vs. county vs. ward; relation between e.g., family income and political preference
  - e.g., crop statistics by county may show strong spatial autocorrelation, which becomes much weaker at district or state level, although the underlying process is the same.
- Technical term: **modifiable areal unit problem** [13]

Example: Mexican electoral regions



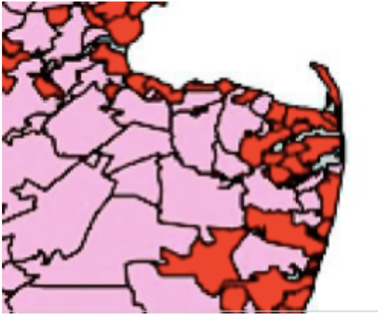
Source: [20, Fig. 1]
Two levels of aggregation: state, region
Question: what socio-economic factors determine voting patterns?

Example: New Jersey housing



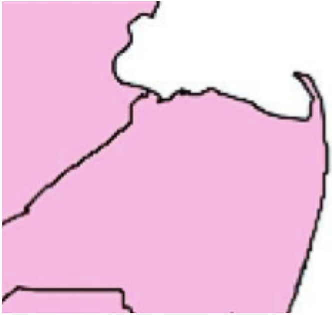
Source: [9, Fig. 1]
Percentage of homes built before 1950 (risk factor for Pb poisoning)
Aggregation level: USA census block group, ZIP code, county.

Detail



ZIP code

"hot spots" of 50–75% older houses



County

all 25–50% older houses
no hot spots

The "ecological fallacy" – non-spatial (1)

- **"Ecological"** = context of observations: "In an *ecological correlation* the statistical object is a *group* of persons"
- the **Fallacy**: inferences about a **fine-scale** grouping can be deduced from inferences for a **coarse-scale** grouping
  - E.g.: regression/correlation of voting preferences based on socio-economic factors at **state/province** level vs. same relations at **county** level.
  - The *aggregate* relation (at states level) can *not* be obtained by aggregating fine-scale regressions (50 per-state relations)
  - References: [18, 19];
    Ecological Fallacy In: Encyclopedia of Survey Research Methods
    `https://dx.doi.org/10.4135/9781412963947.n151`

The "ecological fallacy" – non-spatial (2)

- Fallacy: inferences about **individuals** ("individual-level correlations/regressions") can be deduced from inference for their **group**
  - E.g., Strong empirical-statistical relation between age of schoolchildren and height does *not* imply that a randomly-selected 5th grader is taller than a randomly-selected 4th grader.

The “ecological fallacy” – spatial

- Fallacy: inferences about aggregate data at **small area** can be deduced from inferences about aggregate data for an **enclosing larger area** *or* from inferences from all individual observations
  - E.g.: strong empirical-statistical relation between crime and size of police force (both normalized for population) at **state** level; does *not* imply that there is a strong relation at **city** level within a single state *or* overall.
- Message: *analyze at the level that you want to understand / make policy*.

The "ecological fallacy" and the MAUP

- Correlations at more general levels are generally **stronger** (higher $|r|$) than at finer levels.
- Regressions at more general levels are generally **stronger** (higher R^2) than at finer levels.
- This is because much noise has been averaged out.
- Factor for correlations: $\frac{1-\sigma_{XA}\sigma_{YA}}{\sqrt{1-\sigma_{XA}^2}\sqrt{1-\sigma_{YA}^2}}$
- σ_{XA}, σ_{YA}: variation of the two variables X and Y between strata;
- minimum possible value $= 1$ when there is no variation between strata

Topic: Neighbours

- We observe the results of some (partially?) **spatial process**
  - as opposed to the **non-spatial attributes** of the spatial unit
- Q: What part of the result is due to **spatial** factors?
- Q: How much is there influence on a spatial unit from its **neighbours**?
- To answer this, we need to define "neighbours" and "neighbourhood".

What is a ‘neighbour’?

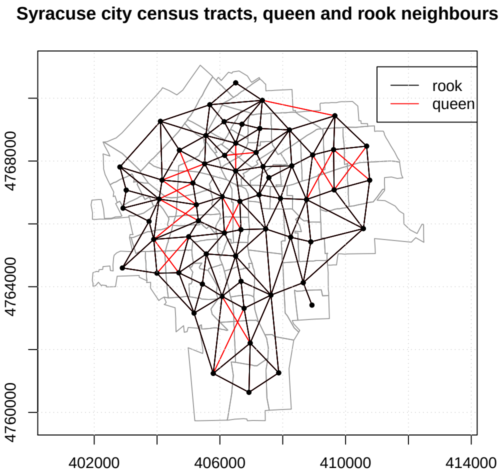
- Q: how do we quantify **“nearby”**?
- A1: **distance** between **centroids** of polygons
  - as with spatial fields; represents polygons as points
  - can use inverse distance, ID^2 …
- A2: common borders: **neighbours** (1st order)
  - “rook” (common line) vs. “queen” (common point) neighbours
  - terminology from legal chess moves
- A3: number of **steps** to reach a common border
  - 1st, 2nd, 3rd… **order** neighbours
- Distance or steps? depends on purpose of analysis
  - what is supposed to drive the **spatial process**?

R packages for areal data

- `sf` "Simple Features representation of Spatial Data" (Edzer Pebesma[1])
- `sp` "Classes and Methods for Spatial Data" (Edzer Pebesma, Roger Bivand[2])
- `spdep` "Spatial Dependence: Weighting Schemes, Statistics and Models" (Roger Bivand)
- `splm` "Econometric Models for Spatial Panel Data" (Giovanni Millo[3])

[1] University of Münster (D)
[2] NHH: Norwegian school of economics
[3] Generali insurance

Neighbours example

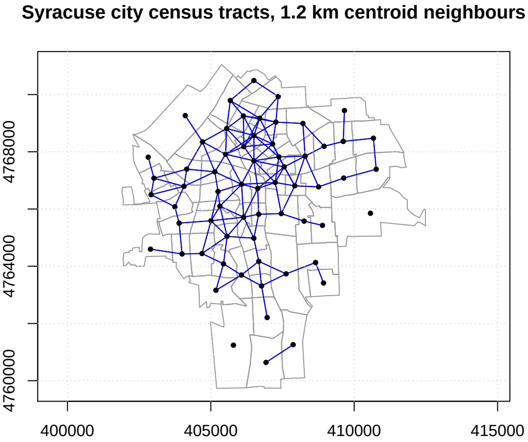


Finding neighbours

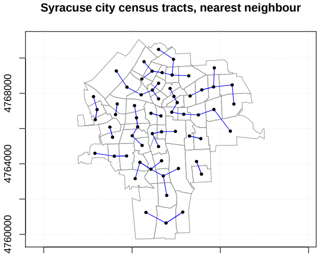
`spdep` functions:

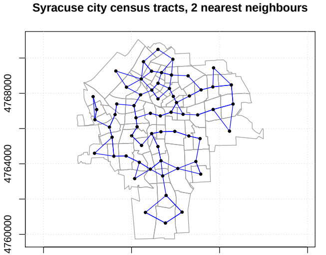
- `knearneigh` find *k* nearest neighbours for each polygon (class `knn`)
- `knn2nb` convert these to weights (class `nb` "neighbour list")
- `dnearneigh` identify neighbours within a given distance band (class `nb`)
- `nbdists` Distances along each link of a neighbour list.

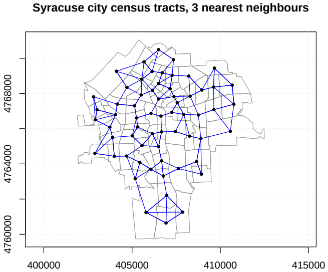
Nearest neighbours within a distance

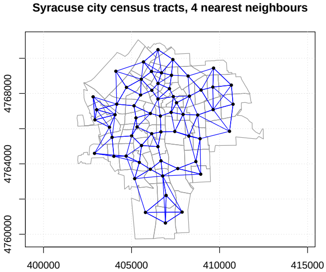


k nearest neighbours









Weighting neighbours

- In models of spatial processes an area's **neighbours** are presumed to have (some) **influence** on a target area.
  - See examples in spatial autocorrelation and spatial modelling.
- A neighbour can be more or less influential to a target polygon, depending on the **spatial process**.
- So, assign a **weight** to each **link** in the graph $\rightarrow$ (a)symmetric **weights matrix**.
- Weights **style** depends on presumed process (see next) – there is no "correct" weighting.

Weighting styles

- Style `B` (**binary**): weights of adjacent polygons affecting a target polygon are either 0 (*not* a neighbour of the target) or 1 (*is* a neighbour)
  - Implies process depends on the **number** of neighbours
  - Can also use with weighting based on **distances between centroids**: multiply 1's by some distance measure
- Style `W`: weights of adjacent polygons affecting a target polygon must sum to 1 (**row-standardized**)
  - All n neighbours equally influential $\rightarrow$ all weights $1/n$.
  - i.e., **total** influence to a target area is **constant**, influence from neighbours **divided among them**
  - Links originating at areas with few neighbours $\rightarrow$ larger weights (edge effect).

Assigning weights

`spdep` functions:

- `nb2listw` spatial weights for neighbours lists (class `listw`, `nb`); styles `W`, `B`, `C`, `U`, `S`

  W row-standardized
  B binary
  C globally-standardized: sum over all links to *n*
  U `C` divided by number of neighbours
  S variance-stabilizing

- `glist` argument to `nb2listw`: pass a list of vectors of weights corresponding to the neighbour relationships
  - example: pre-computed inverse-distance, ID^2W with `nbdists`; use style `B`, will modify "binary" weights
- `listw2mat` show weights matrix

Example weights matrix – style ‘B’

	109	110	111	112	113	114	115	116	117
109	0	1	0	0	1	0	0	0	0
110	1	0	1	1	1	1	0	0	0
111	0	1	0	1	0	0	0	0	0
112	0	1	1	0	0	1	1	1	1
113	1	1	0	0	0	1	0	0	0
114	0	1	0	1	1	0	1	0	0
115	0	0	0	1	0	1	0	1	0
116	0	0	0	1	0	0	1	0	1
117	0	0	0	1	0	0	0	1	0

$1 =$ is a neighbour; $0 =$ not; by definition **symmetric**
row and column headings are census tract identifiers

Example weights matrix – 'B' with Inverse-distance weighting

	109	110	111	112	113	114	115	116	117
109	0.0000	0.6035	0.0000	0.0000	0.6602	0.0000	0.0000	0.0000	0.0000
110	0.6035	0.0000	0.9265	0.5963	1.0111	1.4139	0.0000	0.0000	0.0000
111	0.0000	0.9265	0.0000	0.9858	0.0000	0.0000	0.0000	0.0000	0.0000
112	0.0000	0.5963	0.9858	0.0000	0.0000	0.7191	1.0020	1.1118	0.7829
113	0.6602	1.0111	0.0000	0.0000	0.0000	1.3676	0.0000	0.0000	0.0000
114	0.0000	1.4139	0.0000	0.7191	1.3676	0.0000	1.7476	0.0000	0.0000
115	0.0000	0.0000	0.0000	1.0020	0.0000	1.7476	0.0000	1.7162	0.0000
116	0.0000	0.0000	0.0000	1.1118	0.0000	0.0000	1.7162	0.0000	1.0592
117	0.0000	0.0000	0.0000	0.7829	0.0000	0.0000	0.0000	1.0592	0.0000

Neighbours weighted by inverse distance to centroids; e.g., (110, 111) closer pair then (110,109), so 109 will have less influence on 110 than will 111.

Example weights matrix – style ‘W’

	109	110	111	112	113	114	115	116	117
109	0.0000	0.2000	0.0000	0.0000	0.2	0.0000	0.0000	0.0000	0.0000
110	0.2000	0.0000	0.2000	0.2000	0.2	0.2000	0.0000	0.0000	0.0000
111	0.0000	0.5000	0.0000	0.5000	0.0	0.0000	0.0000	0.0000	0.0000
112	0.0000	0.1429	0.1429	0.0000	0.0	0.1429	0.1429	0.1429	0.1429
113	0.1429	0.1429	0.0000	0.0000	0.0	0.1429	0.0000	0.0000	0.0000
114	0.0000	0.2000	0.0000	0.2000	0.2	0.0000	0.2000	0.0000	0.0000
115	0.0000	0.0000	0.0000	0.2500	0.0	0.2500	0.0000	0.2500	0.0000
116	0.0000	0.0000	0.0000	0.2000	0.0	0.0000	0.2000	0.0000	0.2000
117	0.0000	0.0000	0.0000	0.1429	0.0	0.0000	0.0000	0.1429	0.0000

$0.2 = 1/5$ equal weight to the 5 neighbours of target polygon 109;
$0.14286 = 1/7$ equal weight to the 7 neighbours of target polygon 112 …

Rows sum to unity; `W` is not necessarily symmetric.

Outline

Topic: spatial autocorrelation

- Tobler's first law of geography (1970): "Everything is related to everything else, but near things are more related than distant things"
  - not always true!! It depends on the **process** that generated the spatial distribution of "everything"
- "Auto" = the same feature-space attribute
- Question 1: finding if this is true for a given attribute; quantifying the **range** and **degree** of autocorrelation.
- Question 2: finding out why – really "auto" or due to some other spatially-distributed (but non-spatial) attribute?

Moran's I – motivation

- Q: are attribute values in neighbouring polygons (suitably weighted) more similar than is expected by chance?
  - A: using centroids and inverse distance as weights: variograms or correlograms
  - A: considering (weighted) neighbours: **Moran's I**
- **Assumption**: no spatial patterning due to some underlying spatial factor
  - i.e., apparent spatial correlation in this variable is *not* due to actual spatial correlation of another variable
  - This can be tested in simultaneous autoregressive model (SAR), see below.
- **Assumption**: the assigned neighbour **weights** are appropriate to the process

Moran's I – formula

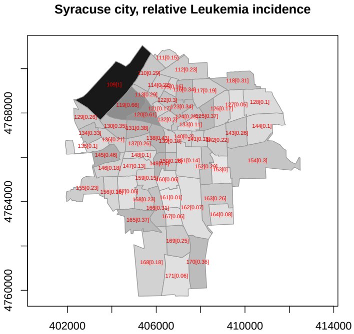
$$I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_i (y_i - \bar{y})^2} \tag{1}$$

- Variables:
  - y_i is the value of the variable in the ith of n polygons
  - $\bar{y}$ is the global mean of the variable
  - w_{ij} is the spatial **weight** of the link between polygons i and j
- The second term numerator is the weighted covariance; the denominator normalizes by the variance
- The first term normalizes by the sum of all weights $\rightarrow$ the test is comparable among datasets with different numbers of polygons and using different weightings.

Global Moran's I test

- Compute for **all pairs** of polygons (i, j)
  - Test is about correlation across the **whole map** – is there any patterning anywhere?
- Assign weights according to hypothesis
  - 1st order: only immediate neighbours (rook? queen?) have non-zero weights
  - 2nd, 3rd...order: zero weights for immediate, 2nd...neighbours, then non-zero weights for the next "ring" (boundary crossing)
- Expected value if random placement of response variable $-1/(n-1)$; complicated formula for variance
- Transform observed I to a normal Z score, compute probability it is by chance that different from the value expected if random allocation of the attribute value to polygons

Example: Syracuse leukemia



Equally-weighted first (rook) neighbours:

```
Moran's I Expectation Variance
0.2075836 -0.0161290 0.0050781

Moran's I test under randomisation
alternative hypothesis: greater
Moran I standard deviate = 3.1394
p-value = 0.000846
```

Conclusion: reject null hypothesis, there is **positive** spatial autocorrelation of leukemia incidence *across the map*. **Note** we have made no attempt (yet) to explain *why*.

Effect of weights

These represent different **hypotheses** about the relative importance of neighbours in the spatial process.

- **W** inversely proportional to the number of neighbours;
  - more weight to areas with few neighbours
- **B** binary: 1 for a neighbour, 0 otherwise;
- **C** globally standardized: inversely proportional to the total number of links;
- **IDW** inverse-distance to centroids

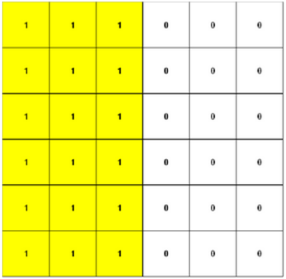
Syracuse leukemia Moran's I with different weights:

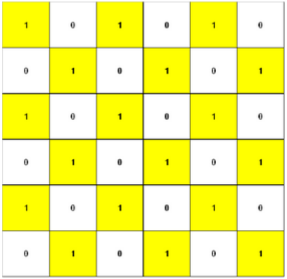
Style	Moran's I	p-value
W	0.207	0.0008
B	0.224	0.0002
IDW	0.195	0.0018

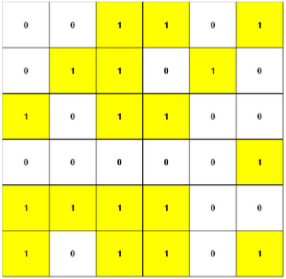
Autocorrelation of categorical variables

- “BB join count”
- Analogous to Moran’s I for continuous attributes
- “BB” = “black/black” vs. “BW” = “black/white” etc., but can have more “colours” (categories)
- Similar to a contingency table for non-spatial attributes
- Tests whether same “colour” joins (mergers) occur more frequently than would be expected by chance (i.e., if the colours were randomly assigned to areas)
- Sensitive to the definition of neighbours and weights
- Sensitive to MAUP and aggregation method
  - mode (most common), nearest (centre)
- R package `spdep`, function `joincount.test` etc.

Example “BW” patterns







global Moran’s I with binary (0/1) weighting:

Separated Even Random

$I = +1$ $I = -1$ $I = -1/(36 - 1) = -0.028$

Source: `http://www.statsref.com/HTML/index.html?two_dimensional_spatial_autoco.html`

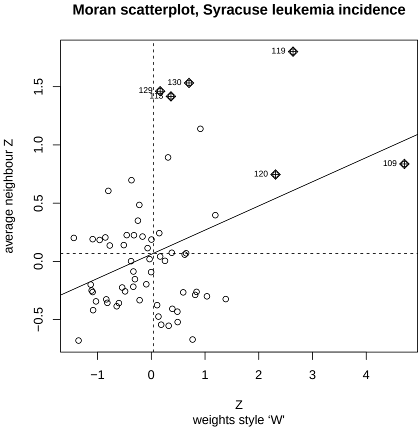
Local Moran's I

- Compute Moran's I **for each polygon separately**:

$$I_i = \frac{(y_i - \bar{y}) \cdot \sum_j (y_j - \bar{y})}{1/n \cdot \sum_i (y_i - \bar{y})^2} \tag{2}$$

- The denominator ensures that $\sum_i I_i = I$
- Show these on a **scatterplot** as Moran's I (x-axis) vs. the **average** Moran's I of all **neighbours** of the polygon
- The slope of the regression between these is global Moran's I!
- Identifies "hot" and "cold" spots of spatial correlation that contribute most to the global Moran's I

Example: Syracuse leukemia (1)



Slope of regression is global Moran's I

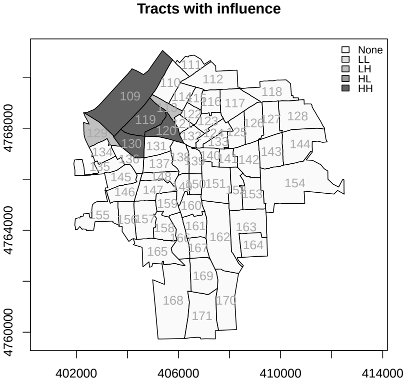
Point numbers are polygon IDs.

x-axis: Leukemia in a district

y-axis: Leukemia weighted-averaged in neighbour districts

Marked points have high leverage (influence on global Moran's I)

Example: Syracuse leukemia (2)

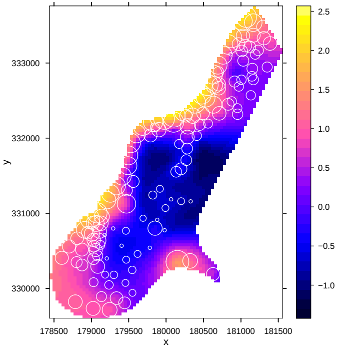


HH: the tract has high incidence, so do its neighbours; etc.

Topic: Hot-spot analysis

Question: are there portions of the study area with consistently higher (“hot”) or lower (“cold”) attribute values than average?

- **Point** data: interpolate from point values over a ‘fine” grid
  - kriging is a smoothing interpolator and will by construction show clusters
- **Area** data: compare areas to average
  - local Moran’s I
  - **Getis-Ord local** G



Getis-Ord local G statistics

- Symbolized as $G_i i$ and $G_i i^*$; the subscript i emphasizes that they are computed separately for each area.
- No attempt to characterize *overall* spatial dependency.
- They identify *local* areas where there may be dependency.

  "These statistics are especially useful in cases where global statistics may fail to alert the researcher to significant pockets of clustering." – Ord and Getis [17]

- Two variants: G_i and G_i^*, where the 'starred' variant includes the self-weights w_{ii} of each target polygon
  - G_i shows whether an area is within a surrounding hot spot
  - G_i^* shows whether the area itself is part of such a hot spot.

Getis-Ord local G – original formulation

A simple concept [10]:

$$G_i(d) = \frac{\sum_j w_{ij}(d) \cdot x_j}{\sum_j x_j} \tag{3}$$

- x the values of the target attribute
- i index of the local area
- j index running over all local areas,*not* including area i
- d buffer distance, selected by analyst
- w symmetric $0/1$ matrix: $1 \rightarrow$ area j is within distance d of area i; but $w_{ii} \doteq 0$

$G_i(d)$ is the proportion of the total of an attribute within distance d of target area i.
$G_i^*(d)$ includes the target area in the index j.

Getis-Ord local G – revised formulation

- Generalize [17] to any weighting, not just 0/1 and not just based on distances
- So it can use the same weighting styles as for Moran's I
- Define as a standard (normal) variate
  - original G_i less its expectation $W_i = \sum_{j \neq i} w_{ij}/(n-1)$ ...
  - ...divided by the square root of the variance:

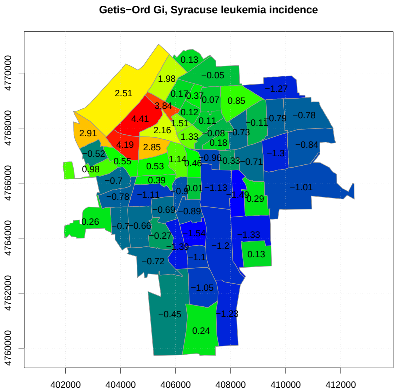
$$\mathrm{Var}(G_i) = \frac{W_i(n-1-W_i)}{(n-1)^2(n-2)} \cdot \left[\frac{s(i)}{\bar{x}(i)}\right]^2$$

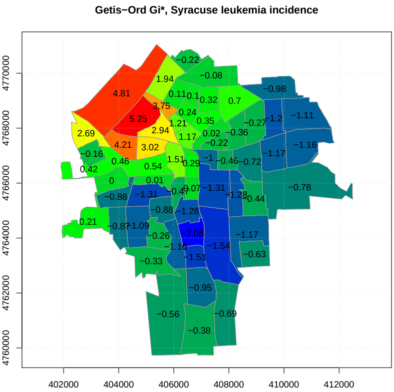
Getis-Ord local G – revised formula

$$G_i^* = \frac{\sum_{j=1}^{n} w_{i,j} x_j - \bar{x} \sum_{j=1}^{n} w_{i,j}}{\mathrm{Var}(G_i^*)^{1/2}} \tag{4}$$

- $\bar{x}, s$ are sample mean and standard deviation of the target variable; n areas
- the $w_{i,j}$ are the neighbour weights
- the numerator shows the difference between area j's weighted average of the target and the overall weighted average
- the denominator standardizes the index
- interpret as Z-score: $+ \rightarrow$ clustering of high values, $- \rightarrow$ clustering of low values

Example: Syracuse leukemia





Outline

LISA

- Developed by Anselin [3] (1995)
- Expanded and implemented in the `GeoDa` **computer program**[4] "Exploratory Spatial Data Analysis & spatial regression"
- Attractive interface to these techniques
- The GeoDa program, documentation and sample data are freely available for download from the Geodata Center's GitHub[5]
- Expanded concept implemented in Naimi et al. [15]
- Experimental R package `rgeoda`[6], interface to GeoDa API

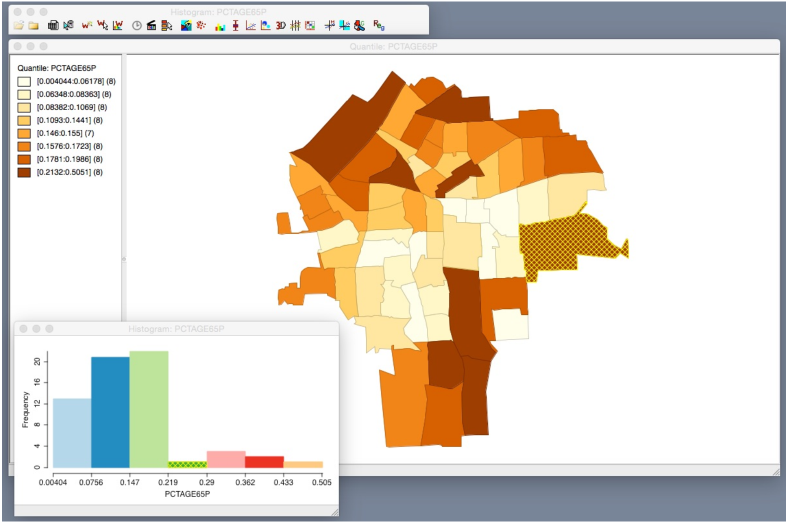
---

[4] `http://spatial.uchicago.edu/geoda`

[5] `http://geodacenter.github.io`

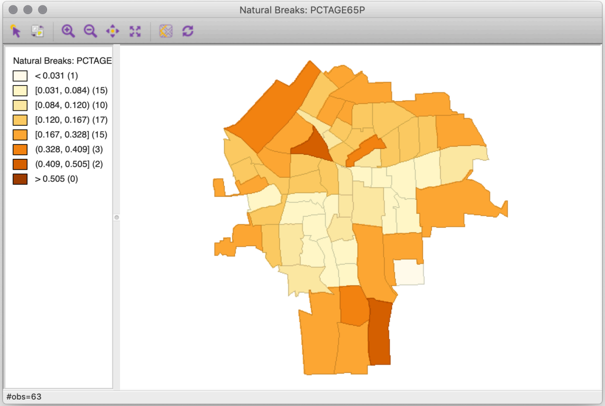
[6] `https://geodacenter.github.io/rgeoda/index.html`

Univariate exploratory graphics quantile plot



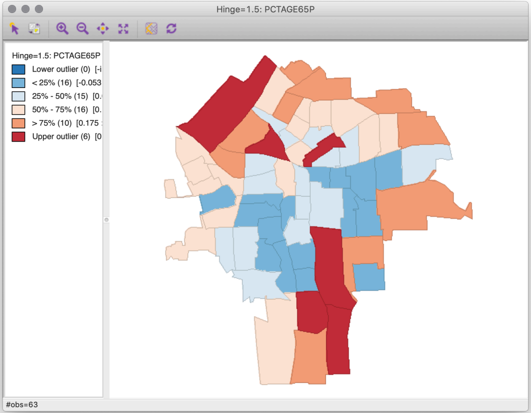
≈ equal numbers of observations in each quantile

Univariate exploratory graphics: natural breaks plot



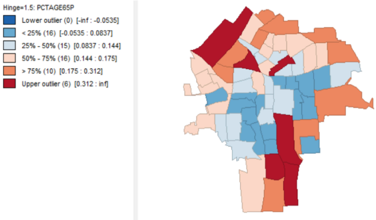
Algorithm to minimize the within-class/between-class variance – equivalent to univariate k-means

Univariate exploratory graphics: Box plot

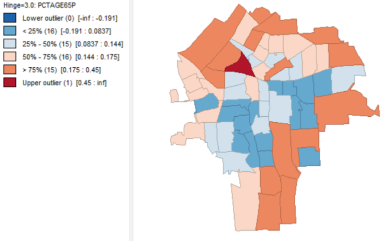


2nd and 3rd quartiles (half the observations); hinges = 1.5 x Interquartile range; outside this are "boxplot outliers"

Box plot map with two hinge limits



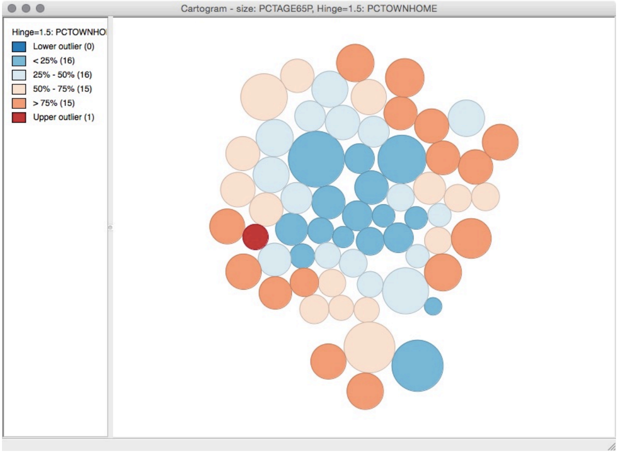
Hinge = 1.5



Hinge = 3.0

Hinge = 3.0 only shows the most extreme.

Bivariate exploratory graphics: cartogram

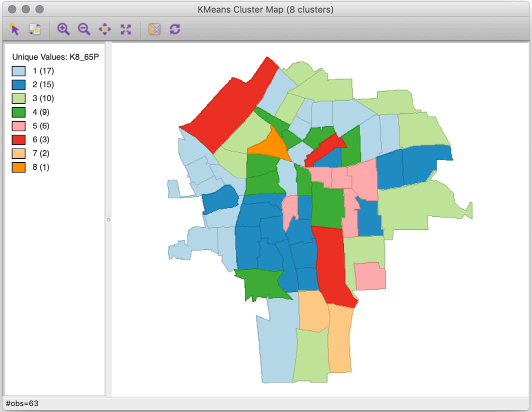


Shows one variable by size, the other by colour, space by the centroids

Clustering

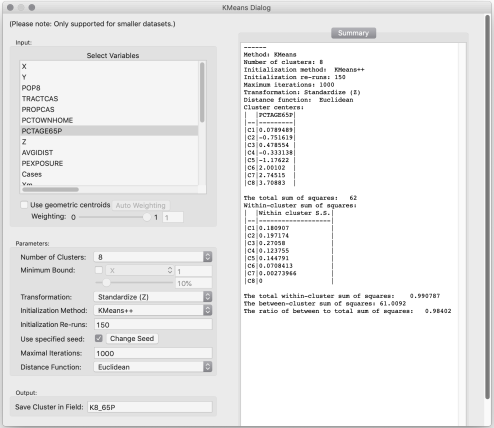
- Objective: group spatial units (e.g., census tracts) into "homogeneous" groups, according to their **feature-space** attributes
  - can also include **coördinates of centroids** as attributes to force **geographic compactness**
- Method: k-means
  - **one-step**: minimize within-class variance, maximize between-class variance; analyst fixes number of classes (k)
- Method: hierarchical clustering
  - **bottom-up** grouping to form increasingly-larger groups
  - each grouping has a "distance" between its members
  - can "cut" the dendrogram (graph) at any level to form any number of groups.

Clustering: univariate k-means (1)



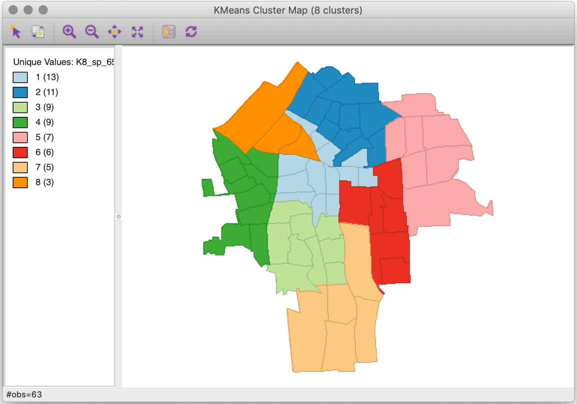
Algorithm to minimize the within-class/between-class variance

Clustering: univariate k-means (2)



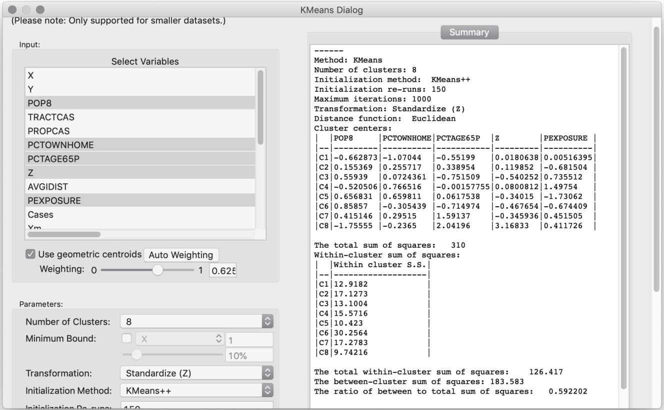
Algorithm to minimize the within-class/between-class variance

Clustering: multivariate geographic k-means (1)



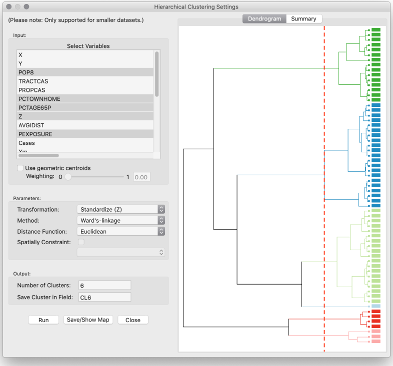
Algorithm to minimize the within-class/between-class variance, while forcing clusters to be **spatially-continguous**

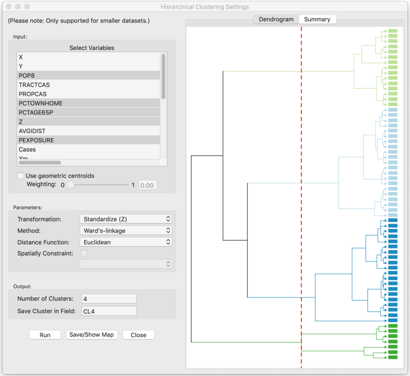
Clustering: multivariate geographic k-means (2)



Algorithm to minimize the within-class/between-class variance, while forcing clusters to be **spatially-continguous**

Clustering: multivariate hierarchical: specification and dendrogram



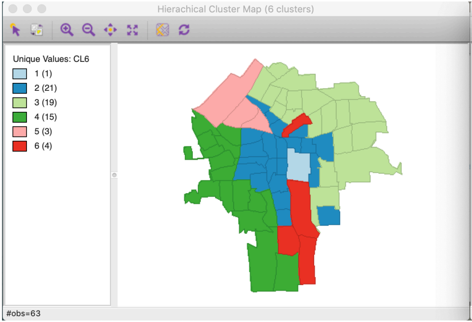


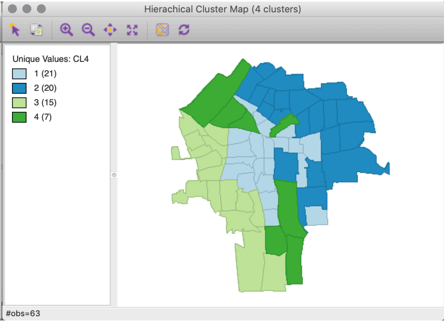
Group at any level of detail; see “distance” between groups in multivariate attribute space

Clustering: cluster statistics

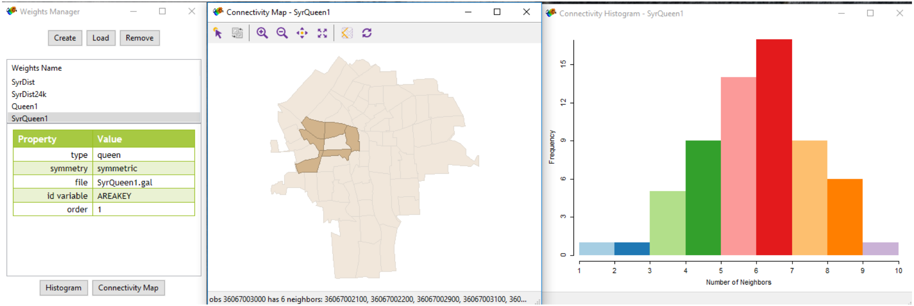
```
Number of clusters: 6
Transformation: Standardize (Z)
Method: Ward's-linkage
Distance function:  Euclidean
Cluster centers:
|  |POP8     |PCTOWNHOME|PCTAGE65P|Z         |PEXPOSURE  |
|--|---------|----------|---------|----------|-----------|
|C1|4.49078  |-1.46223  |-0.430876|-0.637778 |-0.309463  |
|C2|-0.205116|-0.867063 |-0.709461|-0.2613   |0.0199737  |
|C3|0.345393 |0.56543   |0.0892871|-0.13715  |-1.08593   |
|C4|-0.173746|0.847444  |-0.137512|-0.0483164|1.28624    |
|C5|-1.75555 |-0.2365   |2.04196  |3.16833   |0.411726   |
|C6|0.281756 |-0.768694 |2.39247  |-0.012324 |-0.00152647|

The total sum of squares:   310
Within-cluster sum of squares:
|  |Within cluster S.S.|
|--|-------------------|
|C1|0                  |
|C2|36.8377            |
|C3|33.667             |
|C4|27.6518            |
|C5|9.74216            |
|C6|4.68514            |

The total within-cluster sum of squares:    112.584
The between-cluster sum of squares: 197.416
The ratio of between to total sum of squares:   0.636827
```

Clustering: multivariate hierarchical: maps



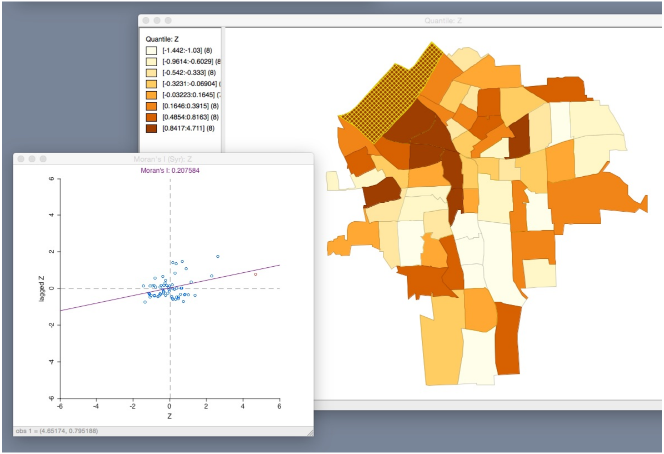


Weights and neighbours



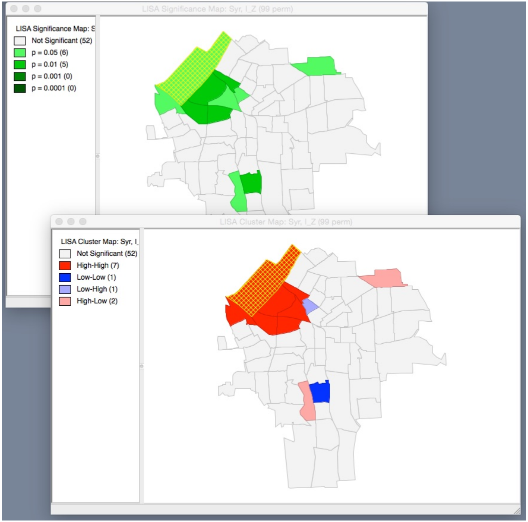
Must operationalize the concept of **“neighbour”** and give each one a **weight** for tests of spatial correlation, and to use in spatial regression models.

Moran's I



Click on point in local Moran graph, highlights the polygon on the map; slope of line is global Moran's I

Influential and clustered polygons for Moran's I



Spatial regression model

```
SUMMARY OF OUTPUT: SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION
Data set            : Syr
Spatial Weight      : Syr.gal
Dependent Variable  :           Z  Number of Observations:   63
Mean dependent var  :   0.0377522  Number of Variables   :    5
S.D. dependent var  :    0.996518  Degrees of Freedom    :   58
Lag coeff.   (Rho)  :    0.210796

R-squared           :    0.385150  Log likelihood        :    -74.1247
Sq. Correlation     : -            Akaike info criterion :     158.249
Sigma-square        :    0.610576  Schwarz criterion     :     168.965
S.E of regression   :    0.781394

-----------------------------------------------------------------------
    Variable     Coefficient      Std.Error        z-value   Probability
-----------------------------------------------------------------------
         W_Z      0.2107962      0.1584045        1.330747     0.1832725
    CONSTANT     0.05371322       1.964857      0.02733696     0.9781909
        POP8   -0.000272356   6.987756e-05       -3.897617     0.0000972
   PCTAGE65P       3.609485       1.053162        3.427284     0.0006097
   PEXPOSURE      0.1670564        1.86764      0.08944783     0.9287259
-----------------------------------------------------------------------

REGRESSION DIAGNOSTICS
DIAGNOSTICS FOR HETEROSKEDASTICITY
RANDOM COEFFICIENTS
TEST                                    DF      VALUE          PROB
Breusch-Pagan test                      3        6.185305      0.1029346

DIAGNOSTICS FOR SPATIAL DEPENDENCE
SPATIAL LAG DEPENDENCE FOR WEIGHT MATRIX : Syr.gal
TEST                                    DF       VALUE         PROB
Likelihood Ratio Test                   1        1.651197      0.1987961
========================= END OF REPORT===============================
```

Note "diagnostics for spatial dependence", this is the next topic here

Outline

Topic: spatial modelling

- Our aim is to **understand** some spatial process – *what explains the spatial distribution of a target variable*?
  - We feel we've understood it if we can build a "successful" **model**
  - A model can be used for **prediction** or **policy decisions**
- Special problems in **spatial** models:
  - How much of the process is **local** (**endogenous** to an area)?
  - How much of the process controlled by **other spatially-distributed attributes** (**exogenous**)?
  - Is there a **spillover effect** by which exogenous factors in neighbouring areas affect the outcome?
  - What is the proper **representation of space**? (distance, neighbours, weighting ...)

Finding a "correct" model

- How do we know a model is correct, even if it fits well?
- Problem is **model mis-specification**
- Typical case: **apparent** spatial autocorrelation, caused by an **underlying factor** that is itself spatially-correlated
  - e.g., spatially-correlated productivity of forest blocks; related to spatially-correlated soil conditions.
  - Should analyze according to a **hypothesis** and **assumptions** based on **theory**.
- Method: compare models by their **likelihood** (see below)

Reference: Bivand et al. [7, §9]

Spatial dependence vs. information (1)

- A **non-spatial** analysis (in feature space) assumes **independence** of model residuals.
- "Nearby" (in geographic space) areas may be similar because of some spatially-correlated **underlying factor** in geographic or feature space.
  - e.g., house prices in adjacent city wards all affected by similar proximity to city centre (**geographic** space)
  - e.g., crop yields in adjacent reporting districts all affected by the same climate and similar soils (**feature** space).
- Feature-space attributes of "nearby" areas may affect the target attribute ("**spillover** effect")
  - e.g., attractiveness of a ward for housing may depend not only on its own proportion of green space, but on the proportion in "nearby" wards

Spatial dependence vs. information (2)

- Question: **does the non-spatial** (feature-space) **model remove all the apparent spatial correlation**?
- If the **residuals** are spatially-correlated, the actual amount of information (roughly, "degrees of freedom") is reduced.
  - Spatial autocorrelation usually **reduces** the amount of information supplied by each observations
  - This is because once we know surrounding areas we know something about a target area
- The feature-space model may have incorrect **coefficients**

Zero-mean models

- **Definition**: model where the expected deviance in each polygon from the global mean of a variable is zero
- There may be spatial correlation but it is an attribute of the **spatial process** of the target variable only
  - Example: diffusion of a pollutant from point sources through a homogeneous soil
- Equivalent to **first-order stationarity** in random field theory (geostatistics)
- This is *not valid* if there is **another spatially-distributed variable** that, in feature space, (partially) determines the value of the target variable

Combining feature and geographic space

1. Build a **feature-space** model (e.g., linear model)
2. Check **residuals** for spatial autocorrelation
   - for areal data, use global Moran's I; for point data can also use variograms
3. *If no autocorrelation*, we are done, feature space explains everything
4. *If there is autocorrelation*, build a model accounting for spatial autocorrelation. Various forms (see below):
   - Simultaneous Autoregressive Models (SAR)
     - spatial error SAR
     - spatial lag SAR
     - spatial Durbin SAR
     - spatial Durbin error SAR
   - Conditional Autoregressive Models (CAR)
5. Verify that the spatial model is more correct than the non-spatial model (e.g., Likelihood Ratio test)

Linear model with independent residuals

This is the **non-spatial** formulation; response is explained by predictors **in attribute space only**:

$$Y = X\beta + \varepsilon \tag{5}$$

- X: design matrix of predictor values
- ε : independent and identically-distributed $\mathcal{N} \sim (0, 1)$ errors
- To estimate: β, the linear regression coefficients
- Solve by minimization of $\varepsilon^2 = (Y - X\beta)^2$
- BLUE is Ordinary Least Squares (OLS):

$$\beta = \left(X^T X\right)^{-1} X^T Y \tag{6}$$

Example: Central NY leukemia

Leukemia incidence based on likely feature-space predictors:

PEXPOSURE exposure to TCE (tricholoroethylene)[7] sources

- toxic chemical linked to cancer

PCTAGE65 % of residents $>$ 65 years old

- cancer incidence may increase with age

PCTOWNHOME % of homes owned

- wealthier =? better health care? less likely to have worked in a chemical plant?

These are **predictors** more-or-less linked to **presumed causes** ...but **correlation** $\neq$ **causation**!

281 census tracts in 8 Central NY counties: Cayuga, Onondaga (includes **Syracuse city**), Madison, Chenango, Broome, Tioga, **Tompkins**, Cortland

[7] https://wwwn.cdc.gov/TSP/substances/ToxSubstance.aspx?toxid=30

Linear model results

Build an additive linear model using these predictors.

```
lm(formula = Z ~ PEXPOSURE + PCTAGE65P + PCTOWNHOME,
            data = NY8@data)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.5173     0.1586   -3.26   0.0012 **
PEXPOSURE     0.0488     0.0351    1.39   0.1648
PCTAGE65P     3.9509     0.6055    6.53  3.2e-10 ***
PCTOWNHOME   -0.5600     0.1703   -3.29   0.0011 **

Adjusted R-squared:  0.184
```

%$>$ 65 years (+), % homeowners (-) "significant", TCE (+) not
But **does the model satisfy linear modelling assumptions?**

Spatial correlation of linear model residuals

```
^^IGlobal Moran I for regression residuals
model: lm(formula = Z ~ PEXPOSURE + PCTAGE65P + PCTOWNHOME,
                    data = NY8@data)
weights: NY8listwB

Moran I statistic standard deviate = 2.64, p-value = 0.0042
alternative hypothesis: greater
sample estimates:
Observed Moran I      Expectation         Variance
       0.0830903       -0.0098913        0.0012423
```

The residuals are (positively) spatially-correlated among neighbours, i.e., **similar residuals are clustered**; so the OLS solution to the linear model **is not correct**.

Simultaneous autoregressive models (SAR)

The solution is to use models that **simultaneously** solve for:

1. the **regression coefficients**, i.e., the effects of the predictors on the response;
2. the **autoregressive error structure**, i.e., the strength and nature of the spatial autocorrelation

Several forms of this, depending the cause of spatial autocorrelation:

- as a result of accounting for the *spatial distribution of the predictors*; a spatially-correlated residual effect: **'induced spatial dependence'** ("spatial error model")
- as a result of a spatial process *within the target variable itself*: **'inherent spatial autocorrelation'** ("spatial lag model")
- both ("mixed model")

SAR model selection (1)

What process do we think is producing the spatial correlation?

- Spatially-correlated residual effect due to a *spatially-correlated feature-space cause* not included in the model: **SAR error model**
  - maybe we don't suspect that it is a cause
  - maybe it has not been measured
  - leukemia example: occupation
  - ecology example: soil type (if not known or in model)
  - crime example: gun laws, sentencing guidelines
- A *diffusion* effect: **SAR lag model**
  - leukemia example: infection (e.g., feline leukemia, not known to occur in human leukemia)
  - ecology example: spread of an invasive species
  - social example: spread of a rumour by word-of-mouth

...

SAR model selection (2)

What process do we think is producing the spatial correlation?

...

- A *spillover* effect: **SAR Durbin model**
  - this must also account for possible diffusion effects
  - leukemia example: exposure to TCE in neighbouring areas, because residents in one area tend to shop or visit in neighbouring areas (??) and so the neighbours add to exposure
  - ecology example: habitat quality in neighbouring forest patches affects bird population in a patch
  - social example: amenities in neighbouring wards affecting desirability of living in a ward

SAR model selection (3) – comparing models

- Compare models with the **Likelihood Ratio** or **Lagrange Multiplier** [5] tests
  - Likelihood Ratio: both models are fit with maximum likelihood, so the two likelihoods are known
  - **likelihood** $\equiv$ probability of the observed data being produced by the model with the given parameters

SAR models – spatial error model

spatial error $\equiv$ the autoregressive process is found only in the **error term**, i.e., not accounted for by any predictor in the linear model

- formula: $Y = X\beta + \lambda W u + \varepsilon$
- W is a matrix representing the spatial structure (e.g., neighbour weights)
- $u = (Y - X\beta)$ are the spatially-correlated residuals
- λ is the strength of autoregression of the errors
- ε is the independent error (not autoregressive)

The concept here is that there is some spatially-structured error, which cause we can not identify, but which we must account for to have a correct model.

SAR models – spatial lag model

spatial lag $\equiv$ the autoregressive process occurs only in the **response variable**

- formula: $Y = \rho WY + X\beta + \varepsilon$
- also can write $(I - \rho W)Y = X\beta + \varepsilon$
- ρ is the strength of autoregression of the response
- Notice how the autocorrelation is applied to the **response** variable, *not* to the linear model **residuals**, as in the SAR error model

The concept here is that the response in *neighbouring* areas affects the response in a *target* area.

SAR models – Durbin model

"Durbin" or "**mixed**": spatial autocorrelation affects both response ('inherent spatial autocorrelation') and explanatory ('induced spatial dependence') variables

- formula: $Y = \rho WY + X\beta + WX\gamma + \varepsilon$
- ρ is the strength of autoregression of the response
- γ is the strength of autoregression of the errors

SAR models in R

- package `spdep`
- SAR error model: functions `errorsarlm`
  - also `spautlom`; this can also compute Conditional Autoregressive (CAR) models
- SAR lag model: function `lagsarlm` with argument `type="lag"`
- SAR Durbin model: function `lagsarlm` with argument `type="mixed"`

Derivation of the SAR spatial error model

- Accounts for spatial autocorrelation of the residuals by a regression on the residuals from adjacent areas
- Residuals are partially the function of some (unobserved) 'hot' (or 'cold') spot of a spatially-distributed covariable
- Each area's residual is modelled as a **linear function** of all the others (depending on neighbours and weights)

$$e_i = \sum_{j=1}^{m} b_{ij} e_j + \varepsilon_i \tag{7}$$

- b_{ij} values: spatial dependence of e_i (residual in one area) on e_j (residual in neighbour area); set $b_{ii} \doteq 0$ (don't self-regress)

SAR error model formulation

$$
\begin{aligned}
Y &= X\beta + B(Y - X\beta) + \varepsilon &\quad (8)\\
(I - B)(Y - X\beta) &= \varepsilon &\quad (9)
\end{aligned}
$$

To estimate: B (spatial dependence), β (regression)
This residual error ε is to be minimized; from the variance:

$$
\mathrm{Var}[Y] = \sigma^2 (I - B)^{-1}(I - B^T)^{-1} \quad (10)
$$

Reparameterize with **explicit spatial autocorrelation parameter** λ and **spatial dependence matrix** W (list of weights):

$$
\mathrm{Var}[Y] = \sigma^2 (I - \lambda W)^{-1}(I - \lambda W^T)^{-1} \quad (11)
$$

and solve for λ by maximum likelihood.

SAR error model example

```
spautolm(formula = Z ~ PEXPOSURE + PCTAGE65P + PCTOWNHOME,
         data = NY8, listw = NY8listwB)
             Estimate Std. Error z value  Pr(>|z|)
(Intercept) -0.618193   0.176784 -3.4969 0.0004707
PEXPOSURE    0.071014   0.042051  1.6888 0.0912635
PCTAGE65P    3.754200   0.624722  6.0094 1.862e-09
PCTOWNHOME  -0.419890   0.191329 -2.1946 0.0281930

Lambda: 0.040487 LR test value: 5.2438 p-value: 0.022026
Asymptotic standard error: 0.016214
```

- `LR test value` compares the models **with and without spatial autocorrelation**.
- p-value: probability that rejecting the **null hypothesis** (the two models are equally likely) would be a Type I error.
- If p-value is low $\rightarrow$ residuals of non-spatial model **are** autocorrelated.

SAR error model interpretation

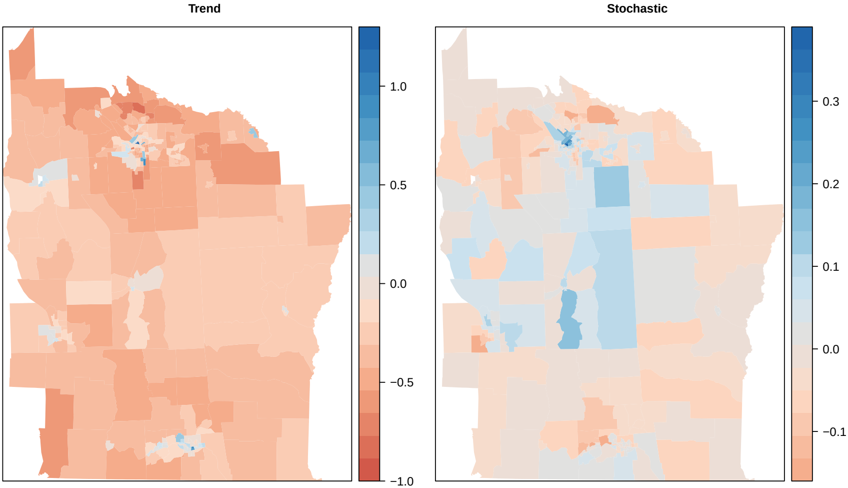
- These coefficients give the influence of **feature-space** predictors, **after** accounting for spatial correlation of residuals
  - i.e., any spatial process is removed (not modelled)
  - computing Moran's I on the SAR residuals should confirm this
- λ gives the **relative strength of the spatial process** vs. the feature-space process
  - can visualize this with trend vs. stochastic residuals fits, see next page
- the **form** of the spatial correlation is modelled by the **form** of **weights**
  - depends on **neighbour list** and **weighting style**
  - weighting style is set by modeller based on hypotheses of how the spatially-correlated error occurs; can compare several for robustness

Comparing regression coefficients: OLS vs. SAR/e

```
# OLS
            Estimate Pr(>|t|)
(Intercept)  -0.5173   0.0012
PEXPOSURE     0.0488   0.1648
PCTAGE65P     3.9509   0.0000
PCTOWNHOME   -0.5600   0.0011

# SAR error model
(Intercept) -0.6182 0.0005
PEXPOSURE    0.0710 0.0913
PCTAGE65P    3.7542 0.0000
PCTOWNHOME  -0.4199 0.0282
```

Substantial change in coefficients; home ownership less important; exposure to TCE more important and now significant at $\alpha < 0.1$.

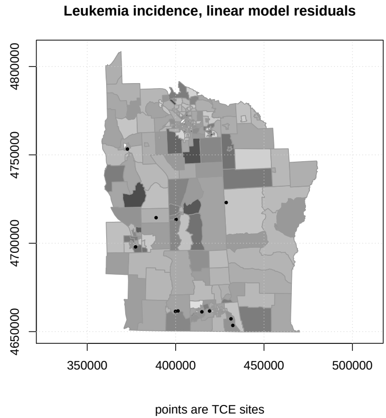
Contributions to model fit

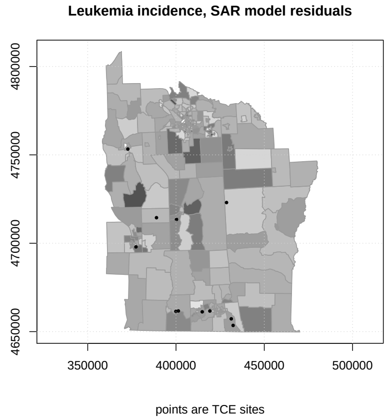


feature space

spatially-correlated residuals

SAR error model residuals





SAR model has *not* removed spatial correlation of residuals, just *changed* it

SAR_error vs. SAR_lag

- The above analysis is with the **SAR error** model:
  - 'induced spatial dependence'
  - process is **exogenous** to the response variable: local hotspots of some unmeasured factor
  - leukemia example: could be local hotspots of carcinogens not included in the `PEXPOSURE` (exposure to TCE) term; could be local hotspots of an unknown or unaccounted for risk factor
- An alternate formulation is the **SAR lag** model:
  - 'inherent spatial autocorrelation'
  - process is **endogenous** to the response variable: diffusion or repulsion effects
  - leukemia example: could be contagious (this is the case for feline leukemia – seems unlikely for humans)

Spatial lag model

```
lagsarlm(formula = Z ~ PEXPOSURE + PCTAGE65P + PCTOWNHOME,
    data = NY8, listw = NY8listwB, type = "lag")
             Estimate Std. Error z value  Pr(>|z|)
(Intercept) -0.514495   0.156154 -3.2948  0.000985
PEXPOSURE    0.047627   0.034509  1.3801  0.167542
PCTAGE65P    3.648198   0.599046  6.0900 1.129e-09
PCTOWNHOME  -0.414601   0.169554 -2.4453  0.014475

Rho: 0.038893, LR test value: 6.9683, p-value: 0.0082967
Asymptotic standard error: 0.015053
```

This is more “likely” (statistically!!) an explanation than the spatial error model: compare LR test values 6.97 (lag) vs. 5.24 (error).
Also, standard error 0.015 (SAR lag) is lower than 0.016 (SAR error).
But this is difficult to justify with our domain knowledge.

Comparing regression coefficients: SAR_error vs. SAR_lag

```
# SAR_error
(Intercept) -0.6182 0.0005
PEXPOSURE    0.0710 0.0913
PCTAGE65P    3.7542 0.0000
PCTOWNHOME  -0.4199 0.0282

# SAR_lag
            Estimate Pr(>|z|)
(Intercept)  -0.5145   0.0010
PEXPOSURE     0.0476   0.1675
PCTAGE65P     3.6482   0.0000
PCTOWNHOME   -0.4146   0.0145
```

Less effect of all predictors after accounting for endogenous spatial autocorrelation in the leukemia incidence.

SAR models: Relation to Generalized Least Squares (GLS)

- Both incorporate spatial correlation structure of the model **residuals** in a mixed model
  - GLS can include many other kinds of correlation structures
- Spatial correlation in GLS depends on an authorized covariance function of separation between point-pairs
  - If polygons are reduced to their centroids, GLS can be used on area data
- SAR uses weighted adjacency matrices to model the linear dependence of residuals on each other
  - so can work with polygons of any shape and size

To remember:

1. **areal** data: **aggregated** over (usually irregular) polygons
2. **apparent** spatial autocorrelation may depend on:
   - a spatial process of that variable;
   - spatially-structured covariable(s);
   - both.
3. Moran's I measures strength of spatial autocorrelation
4. spatial structure depends on assumed process $\rightarrow$ weights matrix; based on distance, common boundary count or length ...
5. paradigm: (1) formulate hypotheses; (2) build model to match hypotheses; (2) test model to see if there is evidence for/against hypotheses
6. **Try to explain** based on **domain knowledge**; beware of the **ecological fallacy**

Outline

Further reading

Theory: Anselin [4, 5], Naimi et al. [15], Openshaw [16]

Use in ecology, difference between SAR model types: Kissling and Carl [14]

Applications: see slide 5

In R: Bivand et al. [8, §9–10]

8-county leukemia study: Ahrens et al. [1]; original data Iwano [12]

Web pages

:

- **GeoDa** Center for Spatial Data Science (Univ. Chicago, Luc Anselin); GeoDa computer program for exploratory ADSA
  `http://spatial.uchicago.edu/geoda/`
  Text: Anselin and Rey [6]
- Workshop: Applied Spatial Statistics in R. by Yuri M. Zhukov, Department of Government, Harvard University; `https://scholar.harvard.edu/zhukov/classes/applied-spatial-statistics-r`

References I

[1] Christina Ahrens, Naomi Altman, George Casella, Malaika Eaton, J. T. Gene Hwang, John Staudenmayer, and Catalina Stefanescu. Leukemia clusters in upstate New York: how adding covariates changes the story. *Environmetrics*, 12(7): 659–672, November 2001. doi: 10.1002/env.490.

[2] L. Anselin. *Spatial econometrics: methods and models*. Kluwer Academic, Boston, 1988.

[3] L. Anselin. Local indicators of spatial association - LISA. *Geographical analysis*, 27 (2):93–115, 1995. doi: 10.1111/j.1538-4632.1995.tb00338.x.

[4] L. Anselin. Under the hood issues in the specification and interpretation of spatial regression models. *Agricultural Economics*, 27(3):247–267, 2002. doi: 10.1111/j.1574-0862.2002.tb00120.x.

[5] Luc Anselin. Lagrange multiplier test diagnostics for spatial dependence and spatial heterogeneity. *Geographical Analysis*, 20(1):1–17, 1988. doi: 10.1111/j.1538-4632.1988.tb00159.x.

[6] Luc Anselin and Sergio J. Rey. *Modern spatial econometrics in practice: A guide to GeoDa, GeoDaSpace and PySAL*. GeoDa Press LLC, Dec 2014. ISBN 978-0-9863421-0-3.

References II

[7] R. S. Bivand, E. J. Pebesma, and V. Gómez-Rubio. *Applied Spatial Data Analysis with R*. UseR! Springer, 2008. `http://www.asdar-book.org/`.

[8] Roger S. Bivand, Edzer J. Pebesma, and V. Gómez-Rubio. *Applied Spatial Data Analysis with R*. Springer, 2nd edition, 2013. ISBN 978-1-4614-7617-7; 978-1-4614-7618-4 (e-book). URL `http://www.asdar-book.org/`.

[9] Paul Elliott and Daniel Wartenberg. Spatial epidemiology: Current approaches and future challenges. *Environmental Health Perspectives*, 112(9):998–1006, 2004. doi: 10.1289/ehp.6735.

[10] Arthur Getis and J. K. Ord. The analysis of spatial association by use of distance statistics. *Geographical analysis*, 24(3):189–206, 1992. doi: 10.1111/j.1538-4632.1992.tb00261.x.

[11] M.F. Goodchild, Luc Anselin, and E. Diechmann. A framework for the areal interpolation of socioeconomic data. *Environment & Planning A*, 25:383–397, 1993.

[12] Eric J. Iwano. *A comparison of cluster detection procedures*. PhD thesis, Cornell University, Ithaca, N.Y., 1989.

References III

[13] Dennis E. Jelinski and Jianguo Wu. The modifiable areal unit problem and implications for landscape ecology. *Landscape Ecology*, 11(3):129–140, 1996. doi: 10.1007/BF02447512.

[14] W. Daniel Kissling and Gudrun Carl. Spatial autocorrelation and the selection of simultaneous autoregressive models. *Global Ecology and Biogeography*, 17(1): 59–71, Jan 2008. ISSN 1466-8238. doi: 10.1111/j.1466-8238.2007.00334.x.

[15] Babak Naimi, Nicholas A. S. Hamm, Thomas A. Groen, Andrew K. Skidmore, Albertus G. Toxopeus, and Sara Alibakhshi. ELSA: Entropy-based local indicator of spatial association. *Spatial Statistics*, 29:66–88, March 2019. doi: 10.1016/j.spasta.2018.10.001.

[16] S. Openshaw. *The modifiable areal unit problem*. Concepts and techniques in modern geography,. Geo Books, Norwich, 1983.

[17] J. K. Ord and Arthur Getis. Local spatial autocorrelation statistics: distributional issues and an application. *Geographical Analysis*, 27(4):286–306, 1995. doi: 10.1111/j.1538-4632.1995.tb00912.x.

References IV

[18] W. S. Robinson. Ecological correlations and the behavior of individuals (reprint of American Sociological Review, Vol 15. No 3 (Jun., 1950),351–35). *International Journal of Epidemiology*, 38(2):337–341, April 2009. doi: 10.1093/ije/dyn357. WOS:000264890300003.

[19] S. V. Subramanian, Kelvyn Jones, Afamia Kaddour, and Nancy Krieger. Revisiting Robinson: The perils of individualistic and ecologic fallacy. *International Journal of Epidemiology*, 38(2):342–360, April 2009. doi: 10.1093/ije/dyn359. WOS:000264890300004.

[20] C. J. Vilalta y Perdomo. The local context and the spatial diffusion of multiparty competition in Urban Mexico, 1994-2000. *Political Geography*, 23(4):403–423, May 2004. doi: 10.1016/j.polgeo.2003.12.009.

[21] L. A. Waller and C. A. Gotway. *Applied spatial statistics for public health data*. Wiley-Interscience, Hoboken, N.J., 2004.