Exercise 3 – Spatial Regression?

Dr. Tessio Novack
(novack@uni-Heidelberg.de)

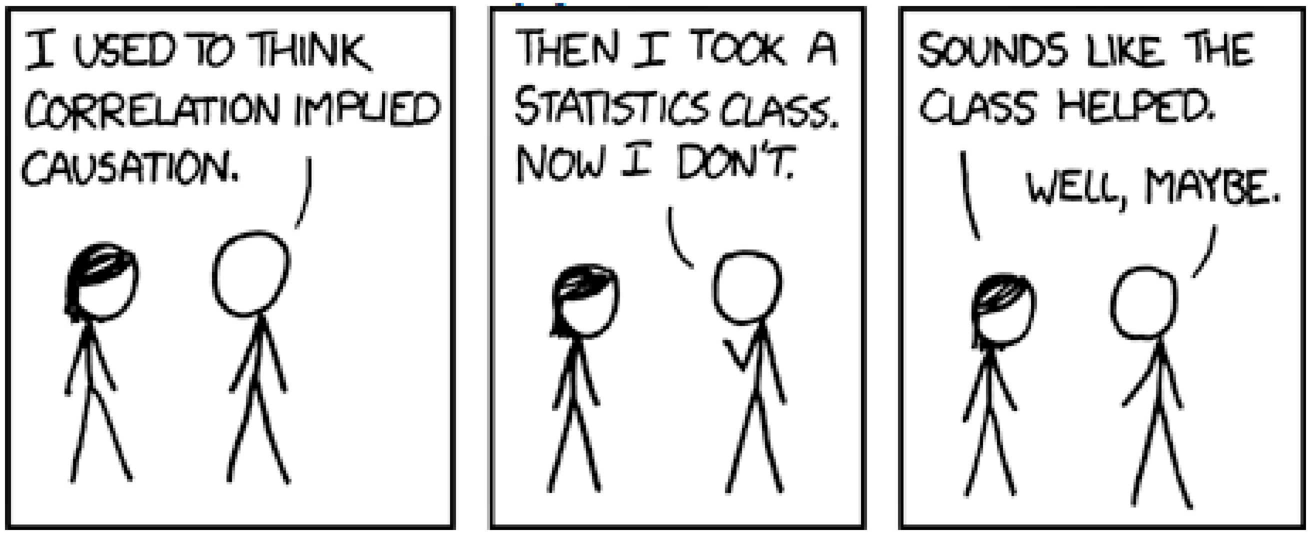
GeoDa Software

- Open source software to introduce spatial analysis
- Functions to explore and model spatial patterns
- Multi-window linked view GUI
- Supports vector and csv data

Correlation

- Defined as the measure of how much two variables X and Y change together
- Range from 1 to -1, i.e. it can be positive or negative correlation

- Ice cream consumption and crime correlate!



Correlation

- Defined as the measure of how much two variables X and Y change together
- Can be between numerical (i.e. discrete or continuous) or nominal variables
  - Examples:

Pearson's correlation coefficient:

$$x = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

where

$$cov(X,Y) = E[(X - E[X])(Y - E[Y])]$$

Mutual Information:

$$I(X,Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log\left(\frac{p(x,y)}{p(x)\,p(y)}\right)$$

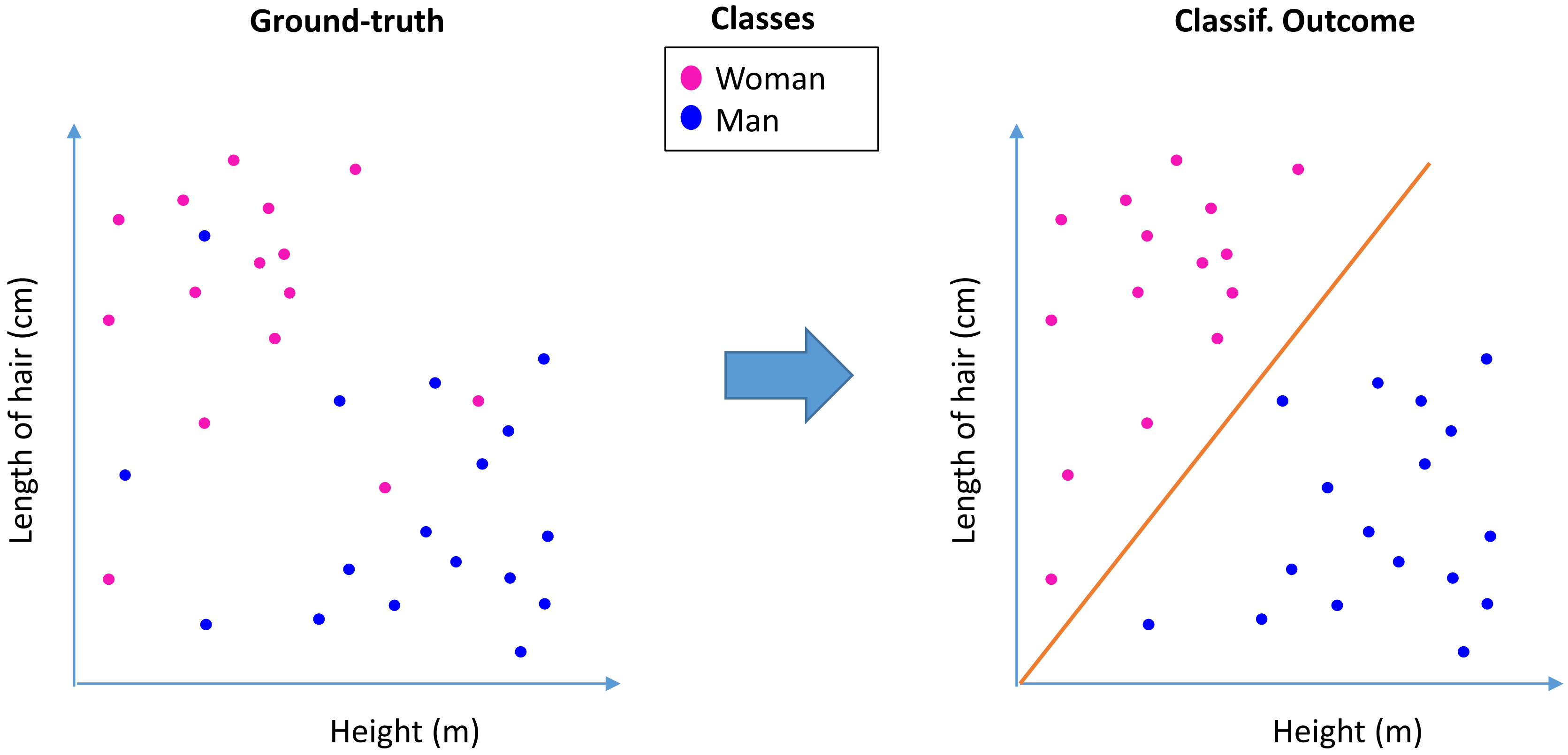
Advantages:

It can handle

- Numerical and nominal variables
- Both linear and highly nonlinear relationships between variables as it does not depend upon fitting a function (linear or otherwise)

Making Predictions

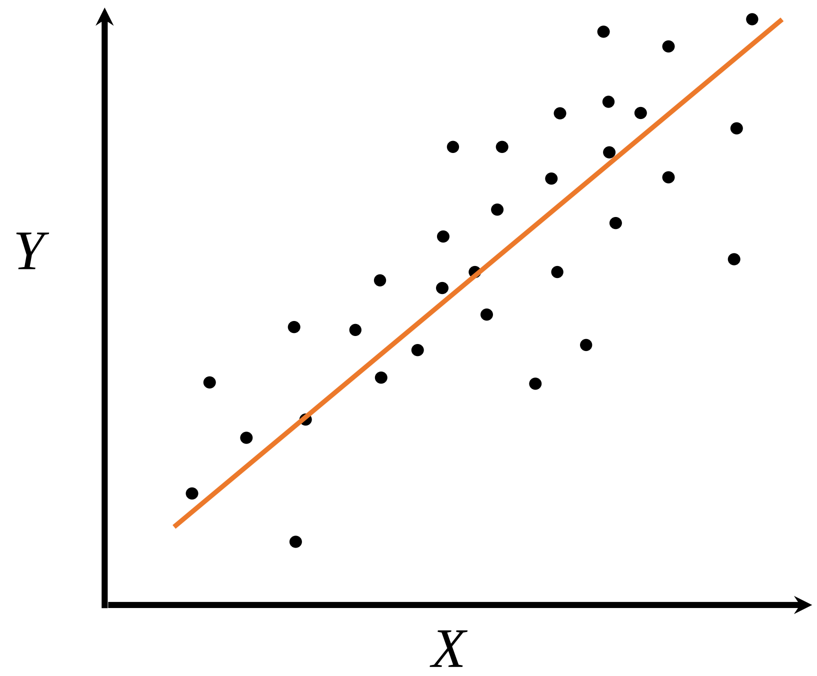
- **Classification:** predicting a nominal variable based on numerical and nominal variables



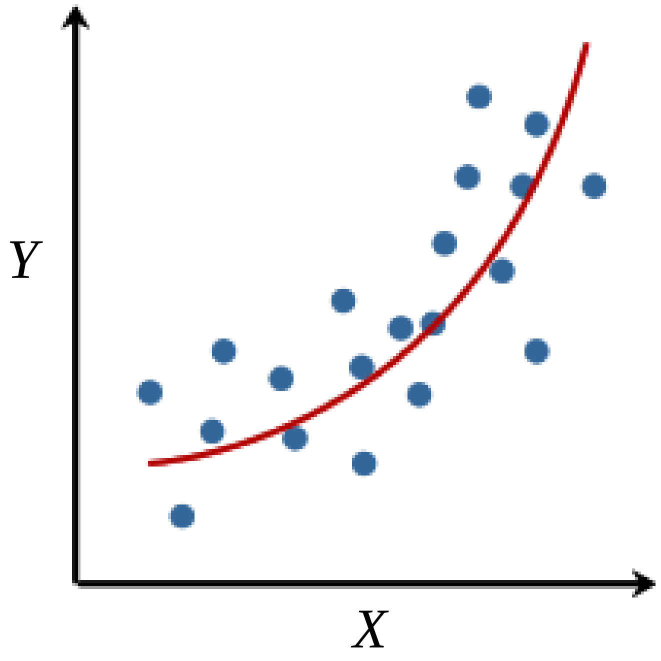
Making Predictions

- **Regression:** estimating the value of unobserved numerical variables based on numerical and nominal variables

$$\bar{Y} = a + b * X$$



$$\bar{Y} = a + bX^2$$



Making Predictions

- **Multivariate Regression:** estimating the value of unobserved numerical variables based on numerical and nominal variables

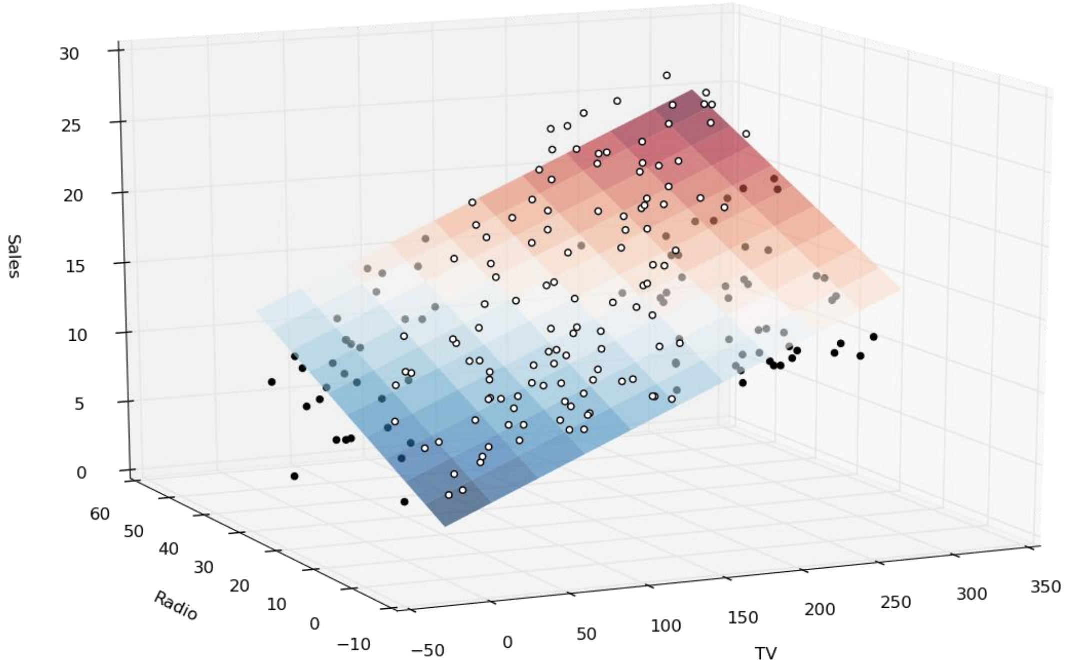
$$\overline{Y} = a + b_1 X_1 + b_2 X_2 + b_3 X_3 + \ldots + b_n X_n$$

a : constant

b_i : coefficient i

X_i : independent variable i

Y : dependent variable



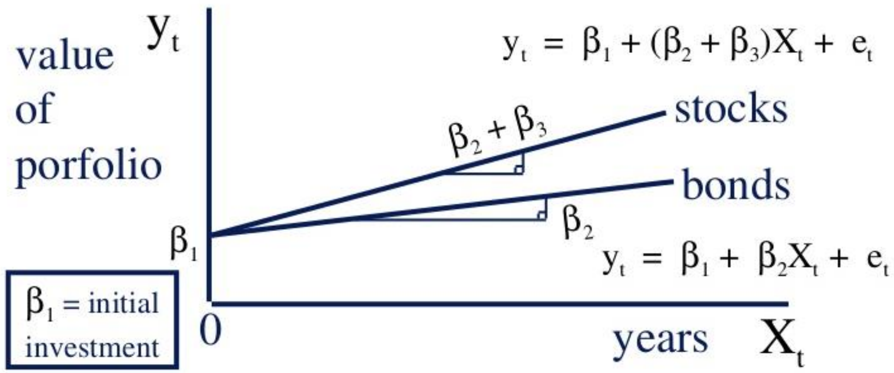
Multivariate Regression

- **Multivariate Regression:** estimating the value of unobserved numerical variables based on numerical and nominal variables

Case: different categories, different slopes

$$y_t = \beta_1 + \beta_2 X_t + \beta_3 D_t X_t + e_t$$

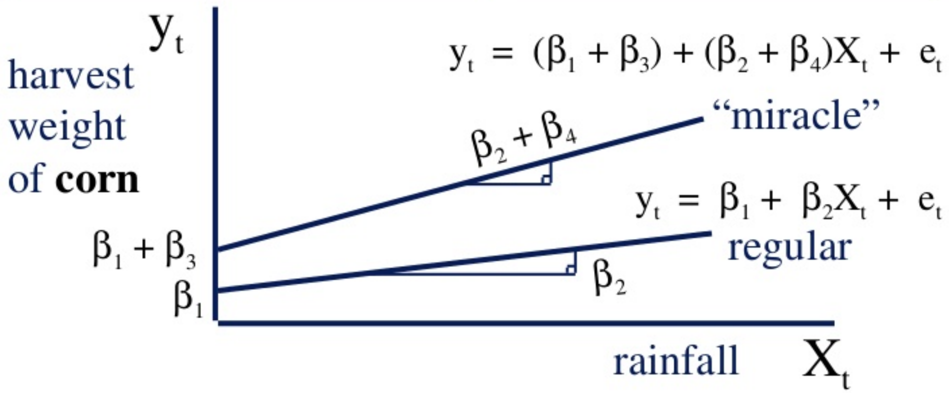
Stock portfolio: $D_t = 1$	Bond portfolio: $D_t = 0$



Case: different categories, different intercepts and slopes

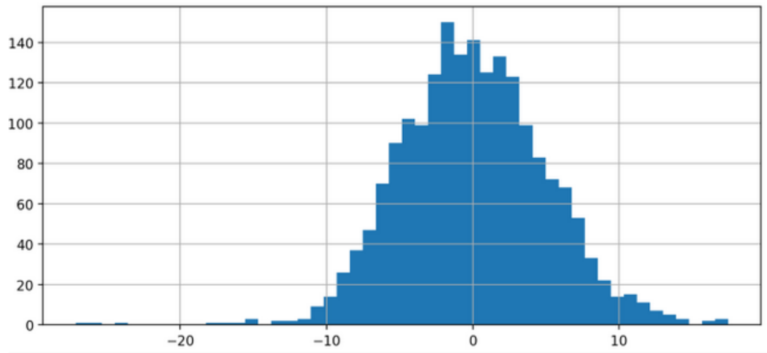
$$y_t = \beta_1 + \beta_2 X_t + \beta_3 D_t + \beta_4 D_t X_t + e_t$$

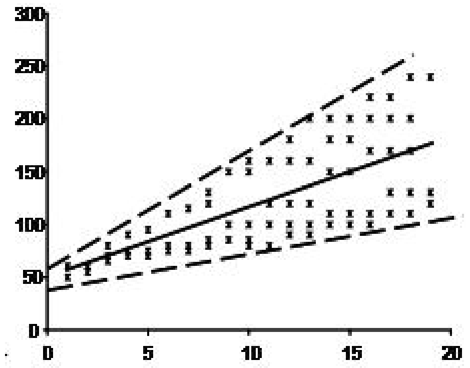
“miracle” seed: $D_t = 1$	regular seed: $D_t = 0$

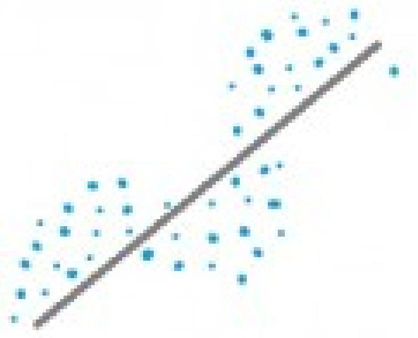


Linear Regression

- **Assumptions:**
  - Linear relationship -> tested with scatterplots
  - Normality of residuals -> Kolmogorov-Smirnov test
  - No or little collinearity -> correlation matrix
  - No auto-correlation (i.e. residuals are independent)
  - Homoscedascity -> Goldfeld-Quandt test
- Rule of thumb: 20 samples per predictor
- Remember to eliminate outliers!

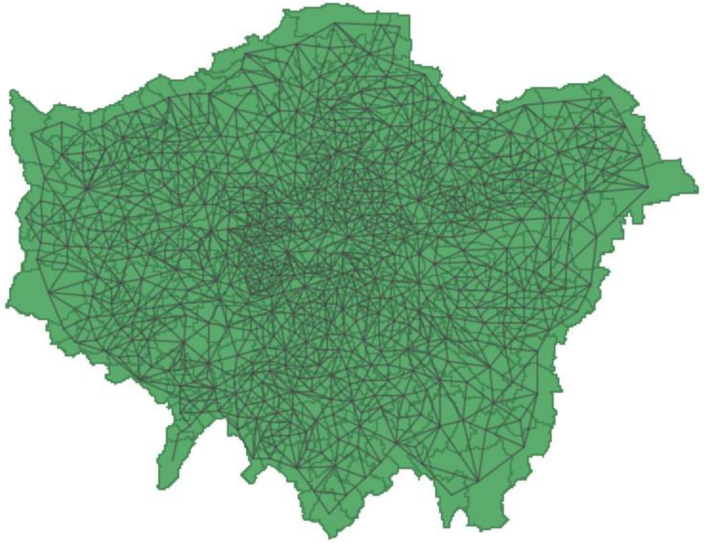


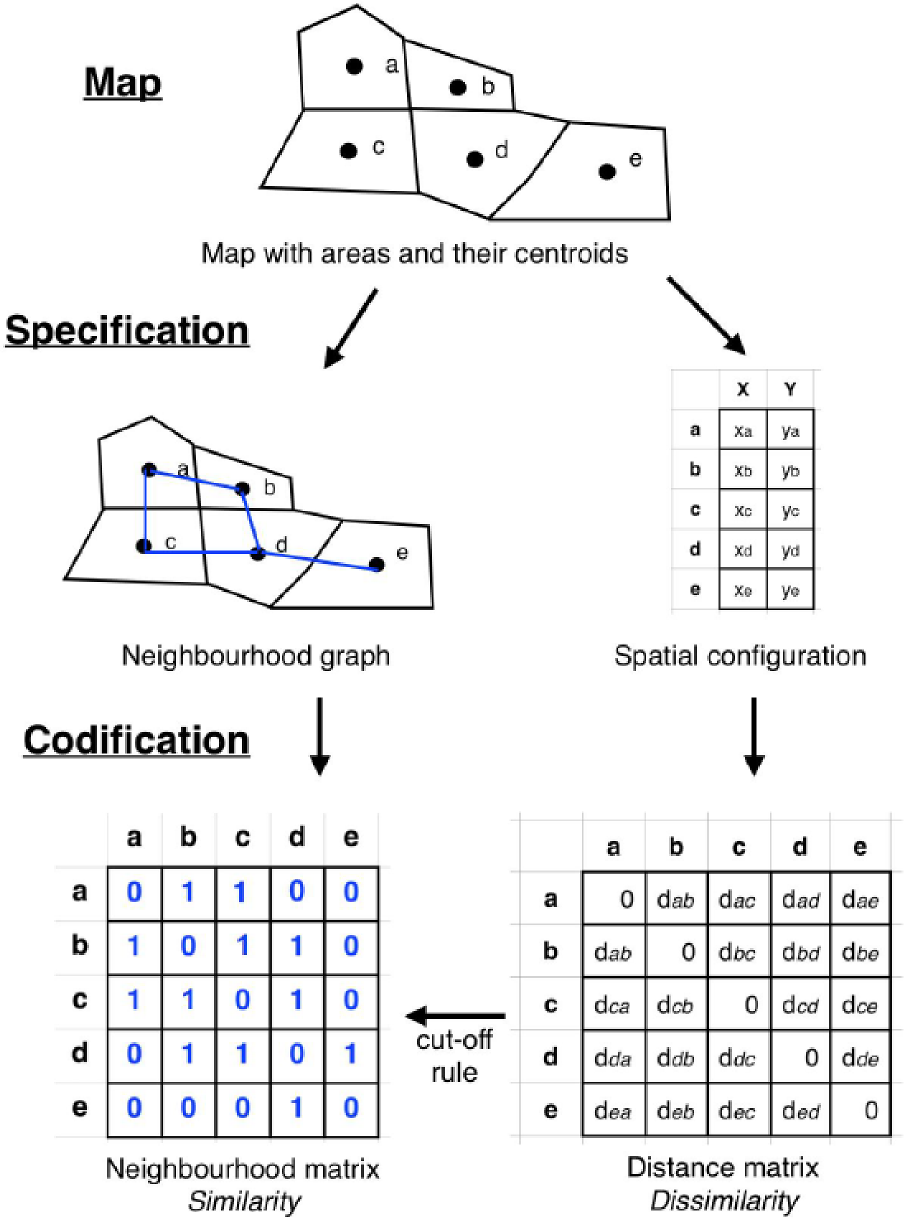




Spatial Weights Matrix

- Defines the spatial relations among locations (i.e. spatial entities)
- Used to account for spatial patterns, where traditional methods may overestimate associations, as linear regression assumes spatial independence
- Main types
  - Contiguity-based
  - Distance-based





	X	Y
a	xa	ya
b	xb	yb
c	xc	yc
d	xd	yd
e	xe	ye

	a	b	c	d	e
a	0	1	1	0	0
b	1	0	1	1	0
c	1	1	0	1	0
d	0	1	1	0	1
e	0	0	0	1	0

	a	b	c	d	e
a	0	dab	dac	dad	dae
b	dab	0	dbc	dbd	dbe
c	dca	dcb	0	dcd	dce
d	dda	ddb	ddc	0	dde
e	dea	deb	dec	ded	0

Spatial Autocorrelation

- Spatial autocorrelation is the correlation among values of a single variable strictly attributable to their relatively close locational positions on space, introducing a deviation from the independent observation assumption of classical statistics.
- Is often measured to avoid violating assumptions of certain regression methods, e.g. OLS
- Can be tested for and quantified

“Everything is related to everything else, but near things are more related than distant things” – Tobler, W. (1970)

Spatial Autocorrelation

- It indicates that that there is something of interest in the distribution of map values that calls for further investigation in order to understand the reasons behind the observed spatial variation
- Might be positive or negative

Positive autocorrelation

Negative autocorrelation

No spatial autocorrelation

Measuring Spatial Autocorrelation

- Moran's *I*: measures the global spatial autocorrelation
- Significance is measured computationally (i.e. pseudo p-value)

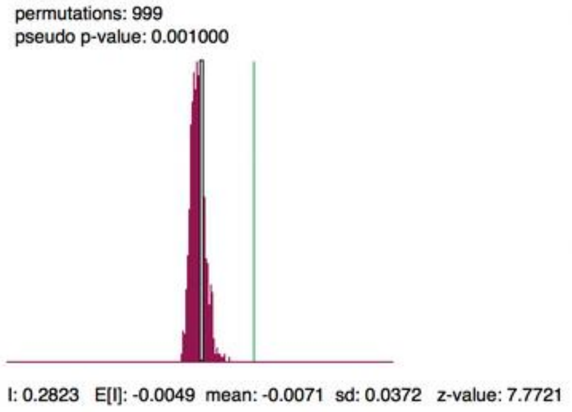
$$I = \frac{N}{W} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

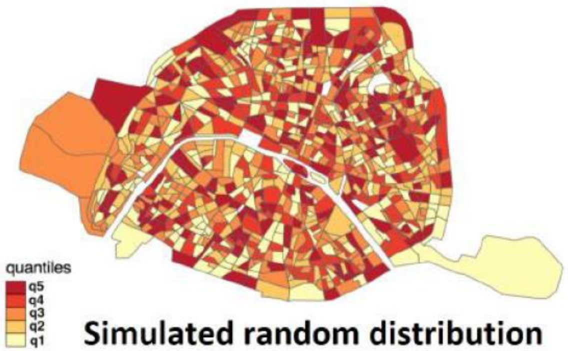
where N is the number of spatial units indexed by i and j;

x is the variable of interest; $\bar{x}$ is the mean of x;

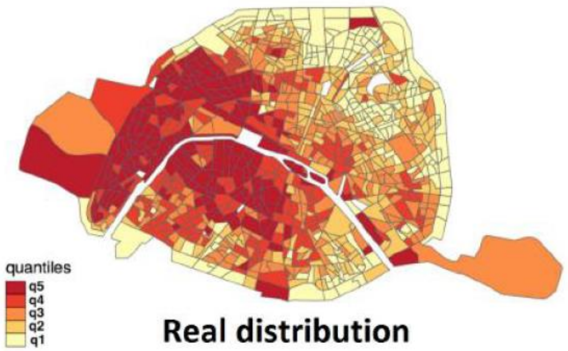
w_{ij} is a matrix of spatial weights with zeroes on the diagonal (i.e., $w_{ii} = 0$);

and W is the sum of all w_{ij}.





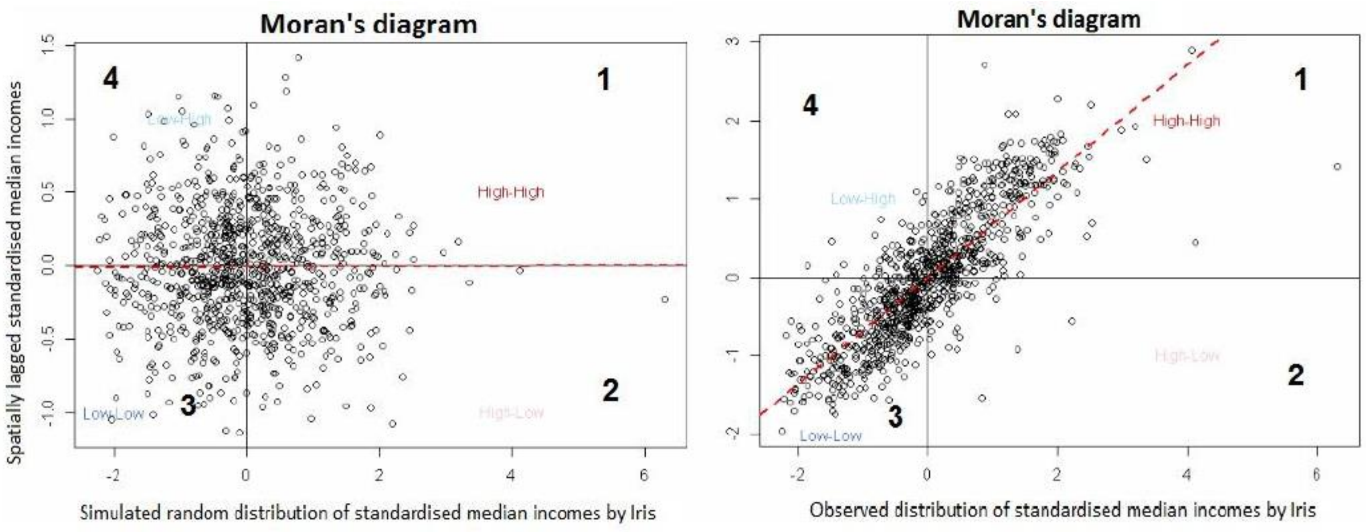
Simulated random distribution



Real distribution

Measuring Spatial Autocorrelation

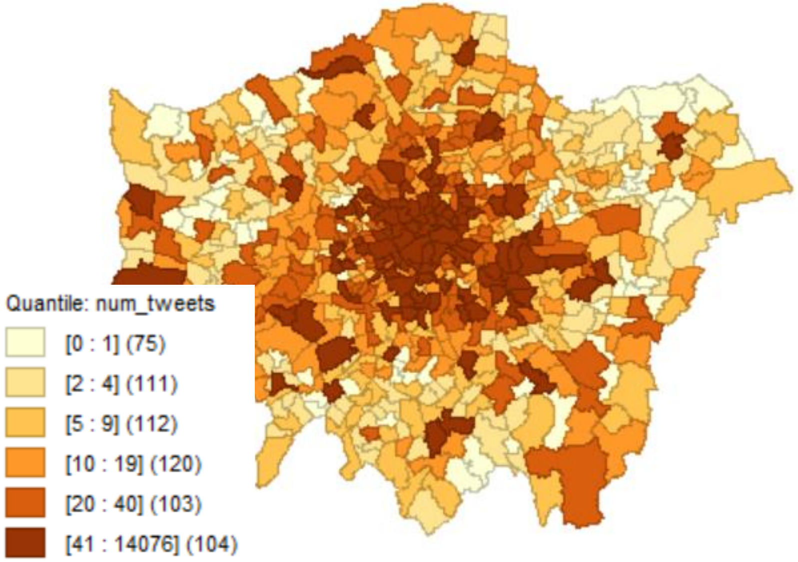
- **Moran scatter plot:** is a plot with the spatially lagged variable on the y-axis and the original variable on the x-axis
- The Moran scatter plot provides a classification of spatial association into four categories, corresponding to the location of the points in the four quadrants of the plot



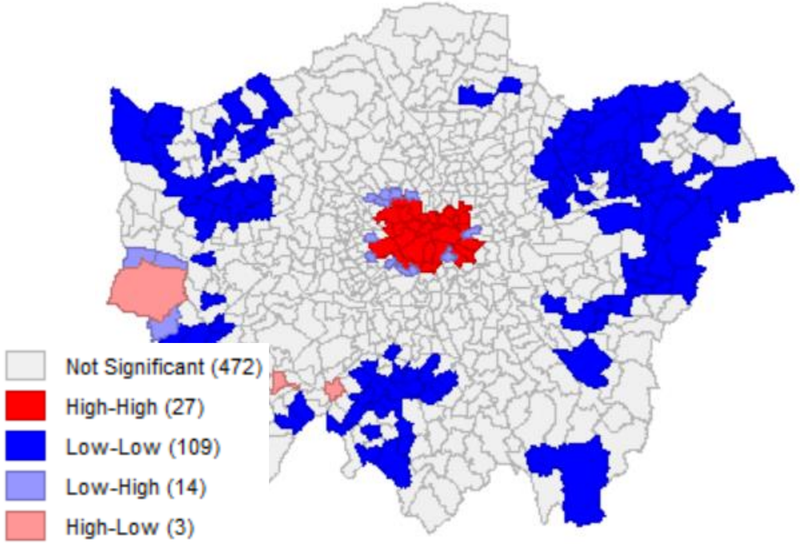
Visualizing Spatial Autocorrelation

- Through so-called Local Indicator of Spatial Association (LISA)
- LISA satisfies two requirements:
  - For each observation: indicates the significance of spatial clustering around that observation
  - The sum of LISAs for all observations is proportional to a global indicator of spatial association

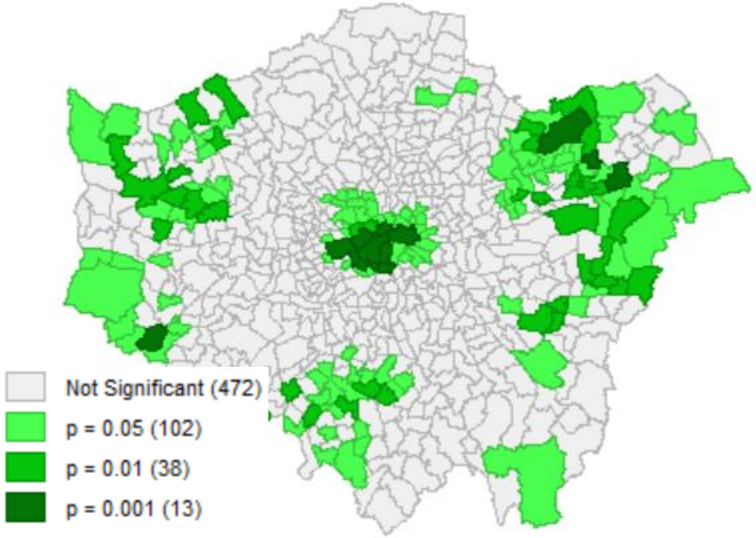
Quantile maps



Categories maps



Significance maps



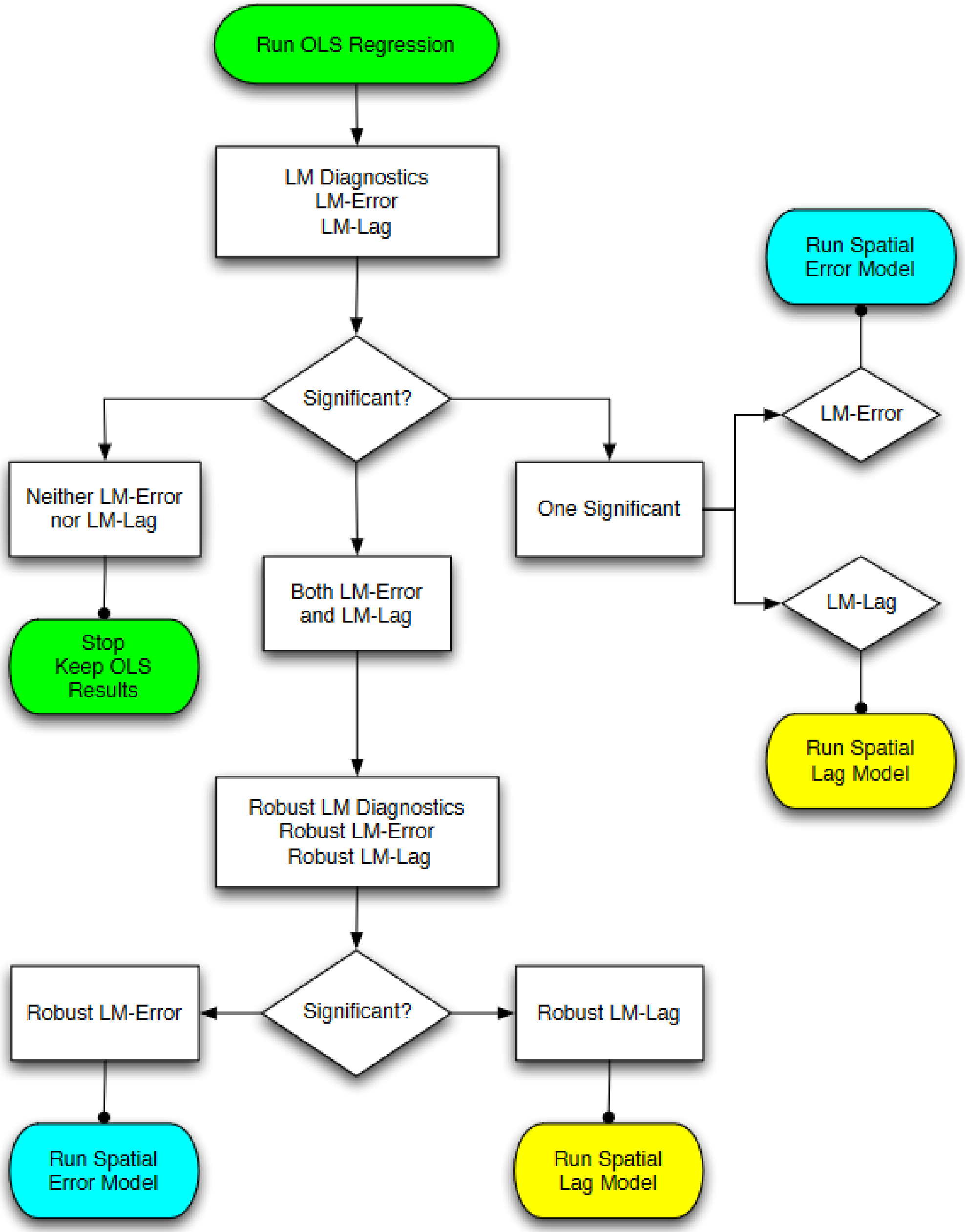
Measuring Spatial Autocorrelation

- **Local Moran's I**
- For identifying local clusters and local spatial outliers
- Significance is based on conditional permutations
- Significance should be assessed in conjunction with Moran's scatter plot

$$I_i = \frac{(x_i - \bar{x})}{\frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2} \cdot \sum_{j=1}^{n} w_{ij}(x_j - \bar{x})$$

Spatial Regression

- Spatial regression decision process
- Tells which model to apply (if necessary)
- Based on Lagrange Multiplier tests



Spatial Error Model

- Consider a spatially lagged error term
- Capture measurement errors generated by unobserved attributes
- Or due to inadequate delineation of the regions

y Is a NX1 vector of observations on the dependent variable

X : is the NXK vector of observations of the independent variables

β : is a KX1 vector of regression coefficients

λ : is the autoregressive coefficient

$W\epsilon$: vector of the spatial lag for the errors

ϵ : is a NX1 vector of spatially autocorrelated error terms

u : is another error term

$$y = X\beta + \varepsilon$$
$$\varepsilon = \lambda \mathbf{W}\varepsilon + u$$

Spatial Lag Model

- Spatial dependencies are captured by the spatial lag *Wy* of the dependent variable *Y*
- Motivation:
  - (1) to obtain the proper inference on the coefficients of the other covariates in the model,
  - (2) capture the strength of spatial dependencies
- Applied when the variable in one place actually increases the likelihood of outcome variable in nearby locales

$$y = \varrho \mathbf{W} y + X\beta + \varepsilon$$

Comparing Models

- The proper measures for goodness-of-fit are based on the likelihood function and include the value of the maximized likelihood, the **Akaike** Information Criterion (AIC) and the Schwartz Criterion (SC).

- The model with the highest log likelihood, or with the lowest AIC or SC is the best.