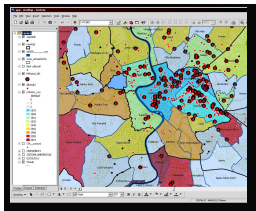
SAPIENZA
UNIVERSITÀ DI ROMA

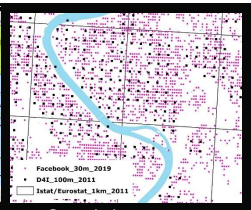
Spatial Data Visualization Analysis & Mapping

2) Introduction spatial statistics

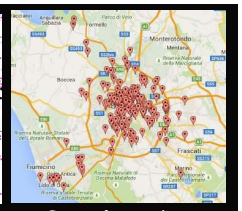
Filippo Celata (filippo.celata@uniroma1.it)



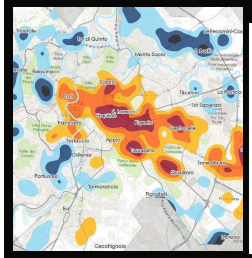
GIS interface and basics



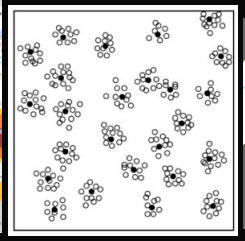
Geodata sources, geometries, problems..



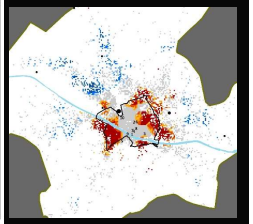
Georeferencing, geocoding



Density maps



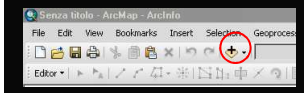
Point patterns



Spatial autocorrelation

Recap / prepare the data: create a point layer of the present population on march 26th 2020 (lockdown) and october 1st 2020 using facebook data per tile, Italy

1) Click add data and add desktop/spatial22/basemap/(add all shapefiles)



2) Click add data and add desktop/ spatial22/data/fb_pop_26mar20.dbf

3) Right click on the table and click 'Display XY data': X Field = **Xutm** and Y Field = **Yutm**

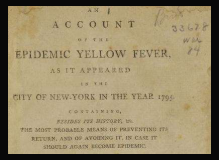
4) Export the 'events' layer to create a shapefile: right click on the 'events' layer and click 'export data', setting 'the same coordinate system as the dataframe', and name the output file 'fbpop_26mar'.

5) Go the layer properties/symbology set 'quantities', 'graduated colors', 'Value' = 'percent_ch' + click on 'classify', set 'Sampling/Maximum sampling size' to 310,000, and set the number of 'Classes' and their 'Break values'.

6) Do the same (1 to 4) for spatial22/data/fb_pop_01oct20.dbf.

The very birth of **spatial statistics:** cartographic analysis of epidemics in the XIX Century (Tom Koch 2005, 2009, 2011)

Valentine Seaman (1797)

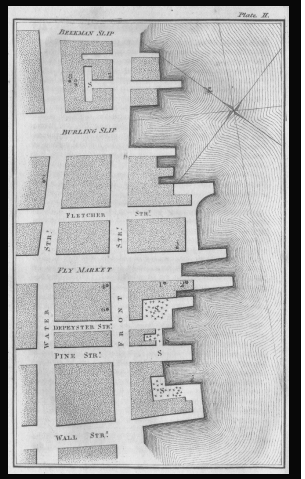

ACCOUNT
OF THE
EPIDEMIC YELLOW FEVER,
AS IT APPEARED
IN THE
CITY OF NEW-YORK IN THE YEAR 1795.

Number: yellow fever cases

S: places of 'miasmas' ("putrid effluvia")

X: crowded places

Spatial proximity = association/correlation = causality

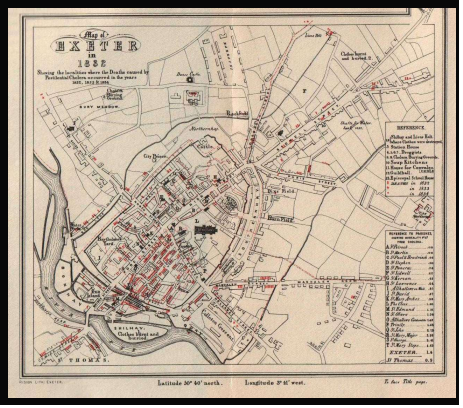


Edwin Chadwick (1800-1890), "*Sanitary Map of Leeds*", 1842. Dark zones: rich; light areas: poors; red points: cholera deaths; blue points: respiratory diseases; 'good' and 'bad' streets [Disease is a class problem! Blame the poor..]

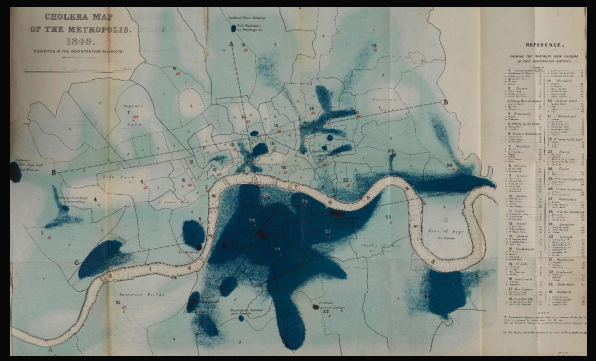
Thomas Shapter (1809-1902), "The History of the Cholera in Exeter in 1832": Map of Exeter "shewing the localities where the deaths caused by Pestilential Cholera occurred in the years 1832, 1833 and 1834", 1849

"A few isolated *spots* in which a remarkable and undue amount of mortality took place" [low-lying, dense, effluvial, odiferous areas (miasmatic airs)]

Sanitation!



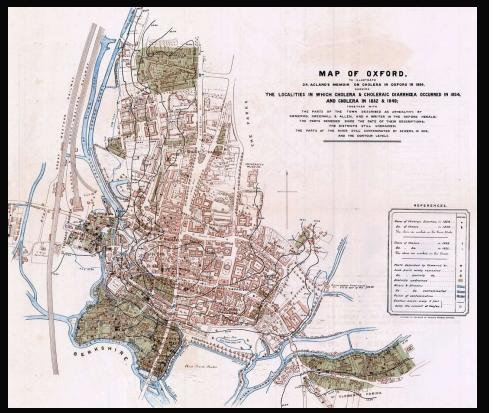
Richard Grainger, "Cholera Map of the Metropolis, 1849", 1850: Deaths' heatmap or density map [Inverse correlation between altitude and disease incidence: miasmatic airs tend to settle around low-lying riverbanks + "Bad ventilation and no drainage" + "over-crowding"]



Henry Acland (1815-1900), "Memoir on the Cholera at Oxford, in the Year 1854", 1855.

Cholera deaths in 1854 (black lines and boxes), 1849 (blue lines), 1832 (blue points); miasmatic airs (brown points); sanitized places (brown circles); polluted waters (dotted lines); marshes (green); elevation lines (black)…

Miasmatic airs!

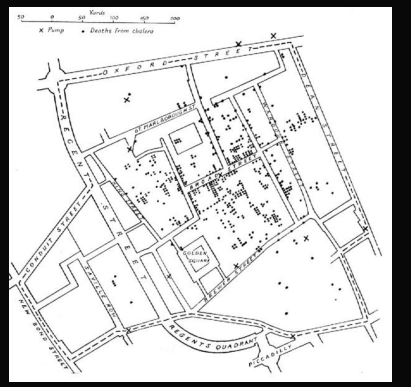


John Snow (1813-1858), "Report on the Cholera Outbreak in the Parish of St. James, Westminster, during the Autumn of 1854", 1855.

The anomalous concentration (hotspot) of cholera cases (black points) in a particular neighbourhood (Broad Street) is not due to miasmatic airs. Cases 'cluster' around a single pump (X).

Cholera is waterborne!

(Close that pump!)



[Celata F. (2020) Storia semiseria della cartografia esattissima delle epidemie, Anno Domini 1690-2020. *Micromega*]

Spatial statistics:

- f(location, distance..)

- to identify (and map) invisible geographical properties of data (eg. distribution patterns)

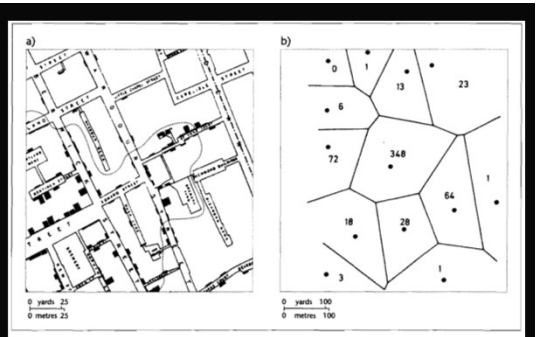


Figure 2. Two views of Snow: (a) Cliff and Haggett (1988) used Snow's cholera map as a tool to describe districting using a Voronoi network of Thiessen polygons. Numbers in individual polygons represent deaths occurring within each. In his 1855 map, (b) Snow used an irregular distance measure to create a single boundary around the Broad Street pump. Reproduced by permission of the authors.

Spatial association: to verify the degree of similarity of spatial events as a function of their distance

Types of spatial association:

1. That are due to **spatial dependence** between geographical features (eg. similar plants require similar soils)

2. That are due to **spatial autocorrelation**: <u>the presence of a certain event increases the probability of finding similar events nearby,</u> due to a reciprocal influence.

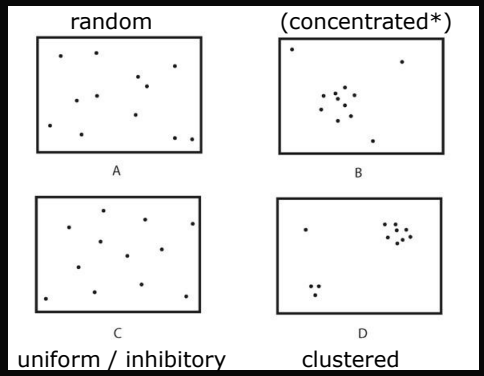
['apparent' vs. 'real contagion': eg. similar plants cluster because they are generated by other similar plants]

Methods:

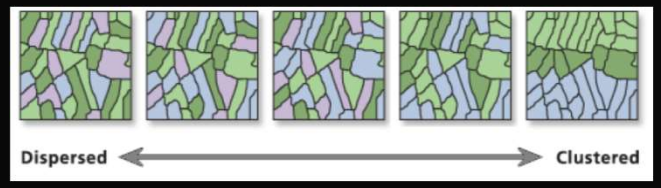
A. To analyze the **spatial distribution** of a pre-selected set of similar events (point patterns or **point processes**)

B. **Autocorrelation analysis**: the degree to which nearby features are more similar than distant ones (to identify relations between proximity and intensity)

A. Point processes and clustering: to verify if the spatial distribution of (similar) events is clustered, dispersed (uniform or inhibitory) **vs.** a "complete spatial randomness hypothesis"



B. Spatial autocorrelation: the degree to which nearby geographical features are similar (vs. CSR)



First law of geography (Tobler, 1970) = "Everything is related to everything else, but near things are more related than distant things".

Spatial association is when the observed distribution of events (point processes) or of their intensity (autocorrelation) is more concentrated than we would expect given a complete spatial random distribution.

Complete spatial randomness (Diggle, 1983) = the event has the same probability to locate anywhere =

- The number of events in any subregion is distributed as a Poisson

-The location of events is not depending upon the location of similar events (indipendence)

- The number of events in two nonoverlapping

regions are independent

3) The average number of events per unit area (intensity) is homogeneous throughout the area (spatial statitionery)

Random distributions implies a certain degree of concentration and/or clustering. This distribution is clustered whenever the degree of concentration is higher than what we would expect in case of complete spatial randomness.

Different techniques imply different CSR hypothesis

Some spatial statistics tools

	Spatial distribution (events)	Spatial dependence (intensities)
Global indexes	Average nearest neighbour (Multi-scale) K-Ripley	Global indexes of autocorrelation: Moran's I Geary's C
Local indexes	Kernel density maps K-Means Clustering Nearest neighbour hierarchical clustering Risk-Adjusted Nearest Neighbor Hierarchical Clustering	Local indicators of spatial association (LISA): Local Anselin of Moran's I (Cluster and outlier analys.) Getis-Ord Gi (Hot-spot analysis)

Where do I find those tools..?

ArcGIS: (kernel) density, average nearest neighbour, Moran's I, Geary's C, Local Anselin of Moran's I ('cluster and outlier analysis'), Getis Ord Gi ('hot spot analysis' and 'optimized hot spot analysis'), Grouping analysis (K-Means), Density based clustering, multivariate clustering, optimized outlier analysis, etc.

QGis: (kernel) density (heatmap), Nearest Neighbor Analysis, K-Means clustering, etc.

Geoda (Anselin et al.): Moran, Local Moran cluster map and + others uni-and bivariate (local) tests of autocorrelation, Local Differential Moran's I (time), colocation join count, k-means, k-medians & co., correlogram (range), Co-location map (categorical data), etc.

Crimestat: point processes (e.g. Nearest neighbour and Risk-Adjusted Nearest Neighbor Hierarchical Clustering), etc

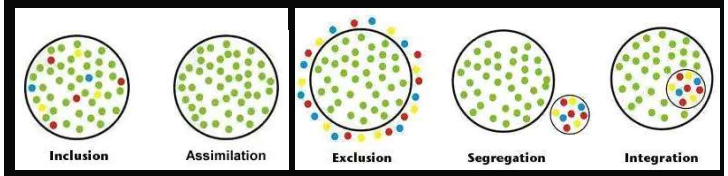
Spatial thinking!

Concentration/dispersion +

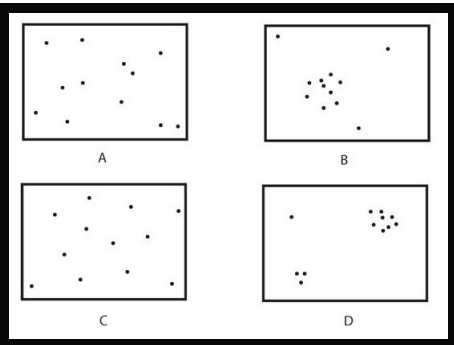
Homogeneity/Heterogeneity =

Co-location, attraction, contagion, agglomeration, interaction, influence, interdependency, centrality, diffusion, inhibition, polarization, uneveness, inequality, segregation, discrimination, homophily, propinquity, etc.

Eg.:



Spatial clustering



Spatial cluster: the distribution of (similar) events is (more) 'clustered' than the CSR and/or the general/global distribution of the process | **Clustering:** tendency of (similar) events to concentrate/co-locate | **Hot-spot:** area with an anomalous concentration of (similar) events

Clustering: <u>"global"</u> indexes (to measure the 'global' degree of clustering for the whole set of events) -> methods based on quadrats (joint count) vs. on distances

<u>AVERAGE NEAREST NEIGHBOUR:</u> the distance between events is less (clustering) or more (pattern inibitorio) of the expected distance in case of complete spatial randomness? (Clark-Evans, '50s)

Nearest neighbour ratio = observed mean distance / expected mean distance (CSR) ->

$$\bar{D}_E = \frac{0.5}{\sqrt{n/A}}$$

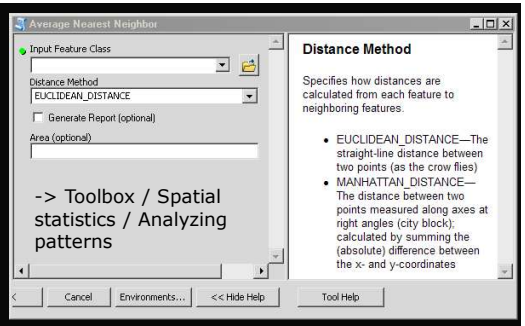
Input:

Points: unweighted (0/1) / Projected coordinate system!

(Polygons and lines: convert into points with x, y = centroids)

Output:

- Observed Mean Distance
- Expected Mean Distance
- Nearest Neighbor Index
- Graphic report
- **Test variables:**

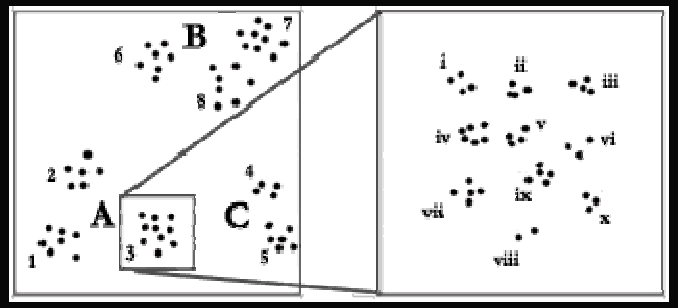


p-value: probabilty of the spatial distribution to be random

z-score: standard deviation of the real values from expected values

z-score (Standard Deviations)	p-value (Probability)	Confidence level
< -1.65 or > +1.65	< 0.10	90%
< -1.96 or > +1.96	< 0.05	95%
< -2.58 or > +2.58	< 0.01	99%

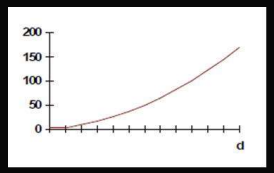
Clustering processes at different scales

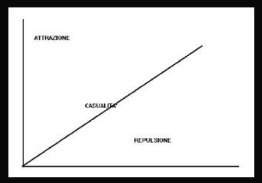


In the figure: 10 clusters of first order, 8 clusters of second order, 3 of third order, and so on..

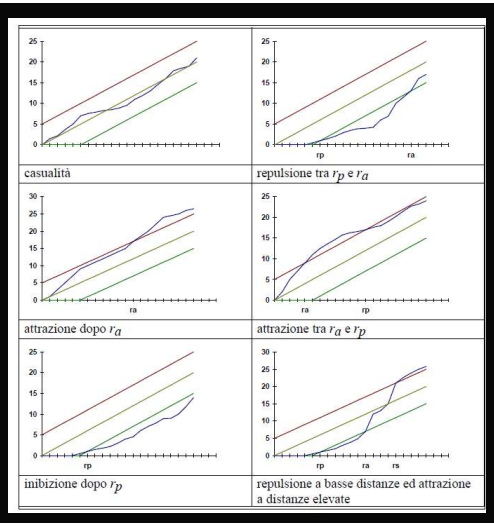
RIPLEY'S K-FUNCTION: A multi-scale global spatial clustering analisys (Ripley 1976, 1981 "Spatial statistics") -> to confirm/reject the null/random hypothesis at various scales/distances + to identify the scale/distance where the clustering/inhibition is more intense/weak

K = expected/observed number of events (in case of CSR: K(d) = πd2

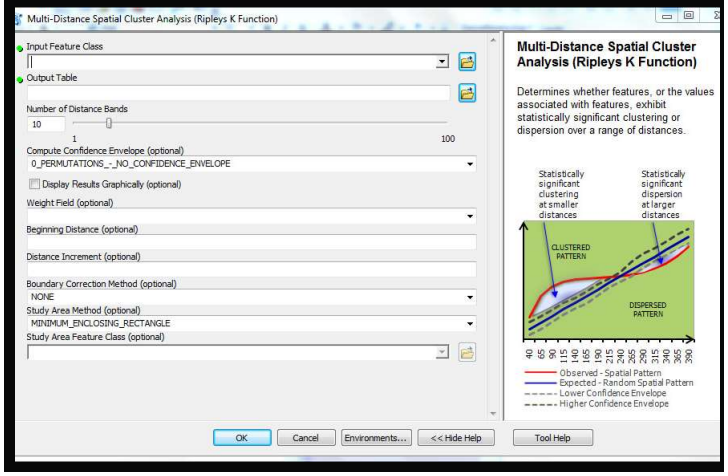




+ lower and upper **confidence envelops:** beyond which results can be considered significant, based on the reiteration of a Montecarlo simulation (simulations work better if the number of points is not small, e.g. > 100)



Spatial statistics / Analyzing patterns / Multi-Distance Spatial Cluster Analysis



Area sensitive index: results are influenced by the area extension -> minimum enclosing rectangle vs. user provided polygonial layer or 'Study Area'

Study area **extension**: bigger the extension, higher the probability that distributions are due to unobserved/ exogenous processes, eg. urbanization

Boundary problem: given the probability of non observed events beyond the study area's boundaries, clustering near the boundaries is under-estimated.

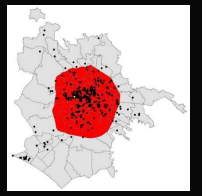
Boundary correction methods:

a. none (eg. mask)

b. simulate_outer_boundary_values

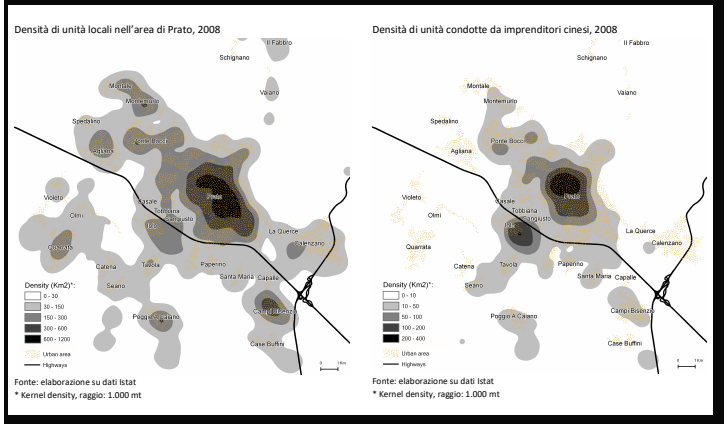
c. reduce_analysis_area

d. Ripley's edge correction formula (weights)



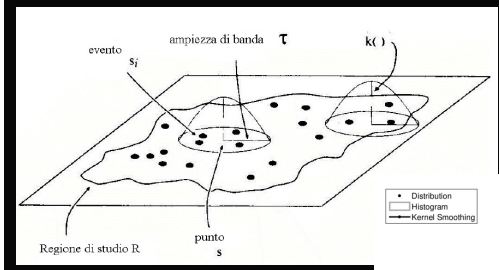
Clustering: 'local' indexes

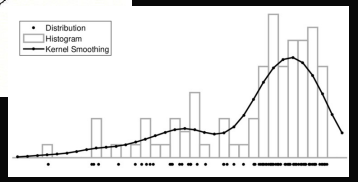
(KERNEL) DENSITY MAPS / HEATMAPS



Kernel density

Silverman, *Density estimation for statistics and data analysis*, 1986



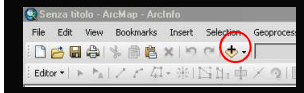


Measure the density for each x/y in the map, based on the number of features (points, lines, or weighted points/lines) which are located within a certain radious, by performing a spatial moving average and 'smoothing' the results

Produce a density map of the difference between the present population during and before the lockdown using facebook disaster maps per tile, Italy

if you didn't already:

1) Click add data and add desktop/spatial22/basemap/(add all shapefiles)



2) Click add data and add desktop/ spatial22/data/fb_pop_26mar20.dbf

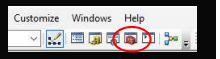
3) Right click on the table and click 'Display XY data': X Field = Xutm and Y Field = Yutm

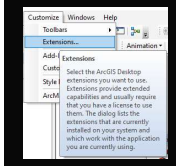
4) Export the 'events' layer to create a shapefile: right click on the 'events' layer and click 'export data', setting 'the same coordinate system as the dataframe', and name the output file 'fbpop_26mar'

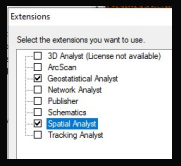
5) Do the same for desktop/spatial22/data/fb_pop_01oct20.dbf, and name the file 'fbpop_01oct'

Produce a density map of the difference between the present population during and before the lockdown (1)

6) Open ArcToolBox and Spatial analyst/Density/ Kernel density



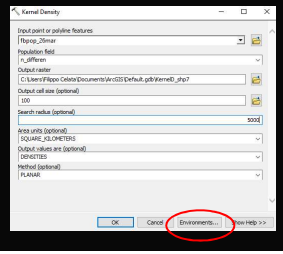




If the tool is not accessable, you need to activate the extensions in the customize/extensions menu

7) Set the kernel density paramaters: Input point feature: 'fbpop_26mar'; Population field: n_difference; Output cell size: 100 (mts); Search radius: 5.000 (mts)

8) Go to the 'enivornments' menu

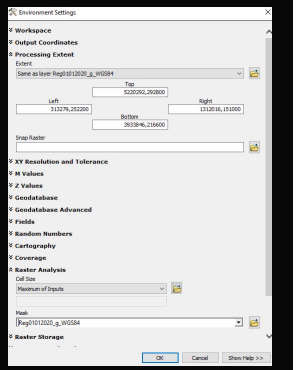


Density map of the difference between the present population during and before lockdown (2)

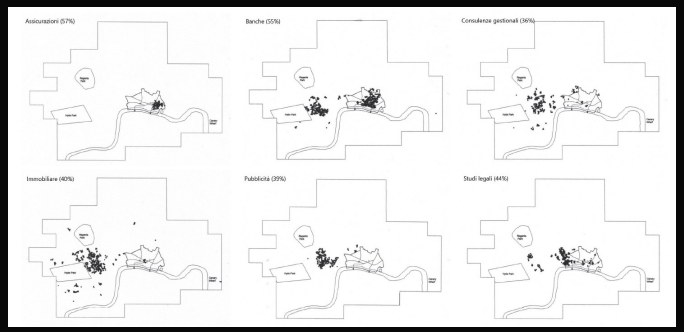
8) In order for the output raster's extent and shape to correspond to that of Italy, you need to 'clip' it: in the 'Environments' menu set both the 'Processing extent' and 'Raster analysis'/'Mask' to that of administrative regions.

9) Fix the symbology: right-click on the output raster, properties, symbology, 'classify' (use geometric intervals or quantiles)

10) In 'layout view', insert the legend, right-clik on the legend and customize it, and export the map via 'file', 'export' (300 dpi)



Urban business clusters in London: insurance, banks, management consulting, real estate, advertisements, lawyers



A clustered firm is defined as one whose average distance to its 10 nearest neighbours (in its sector) is less than 100 metres. Maps show where the clustering in a sector is taking place by, basically, removing all non-clustered firms. The degree of clustering is indicated by the percentage of total firms in each sector that are part of the cluster.

The identification of clusters: methodological challenges

Table 2: Varieties of Cluster and the Cluster Measurement Problem

Cluster Concept	Conceptual/ Definitional Depth	Empirical Methodology	Ease of Measurement	Empirical Support
Co-location	Shallow	Top-Down	Easy to Measure (Quantitative)	Indirect Evidence
Co-location and Technological Proximity	↓	↓	↓	↓
Input-Output Table and Complementarities	↓	↓	↓	↓
Co-location and Superior Performance	↓	↓	↓	↓
Marshallian Externalities	↓	↓	↓	↓
Network Firms	↓	↓	↓	↓
Explicit Collaboration	↓	↓	↓	↓
Informal Knowledge Spillovers	Deep	Bottom Up	Hard to Measure (Qualitative)	Direct Evidence

Adapted from Swann (2002)

Martin R., Sunley P. (2003), **Deconstructing clusters: chaotic concept or policy panacea?** Journal of Economic Geography 3, pp. 5-35.

What happened next? Map the difference of the present population between March 26 and October 1*

11) Do a kernel density map of 'fbpop_26mar' with the same paramaters as in the previous slide -> in the main upper bar, click 'geoprocessing'/ 'results' and than click 'current session' and (the last) 'kernel density', and change the population field to '**n_crisis**'

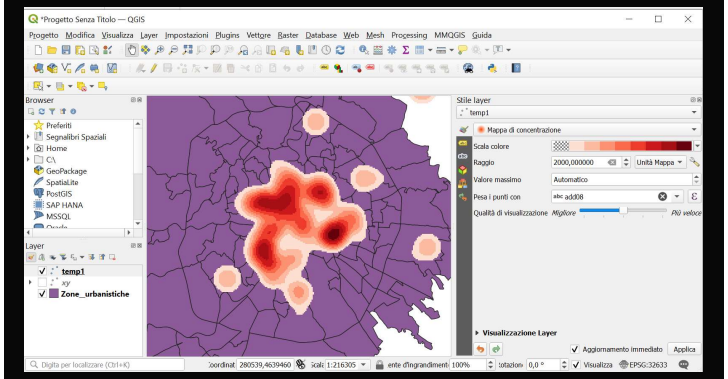
12) Repeat the same process for October 1st 2020, using 'fbpop_01oct.shp'

13) Subtract the 1 October density raster to the 26 March density raster using ArcToolBox/Spatial analyst/Map algebra/Raster calculator.

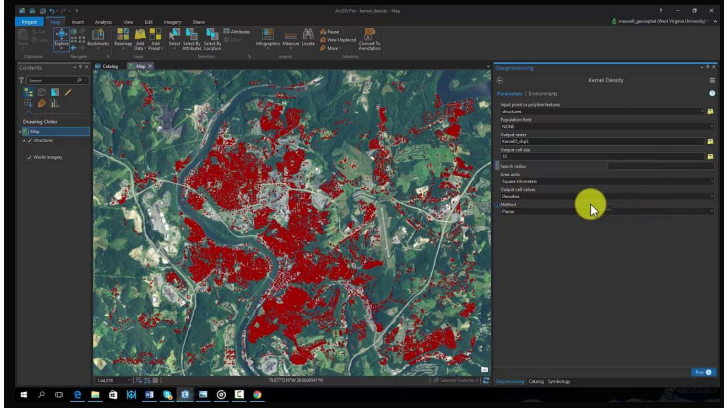
14) Fix the symbology, layout view, legend, and export the map as an image file (300 dpi).

-> surface-based indicators

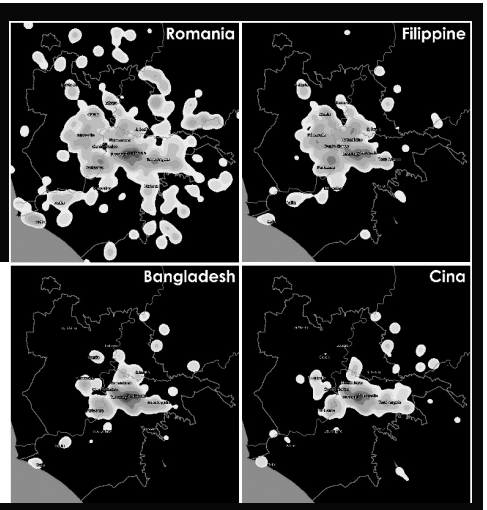
QGis: heatmaps



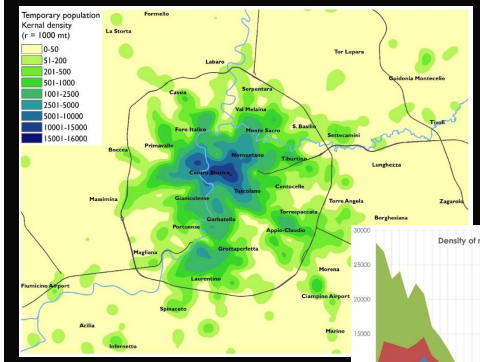
Kernel density in ArcPRO

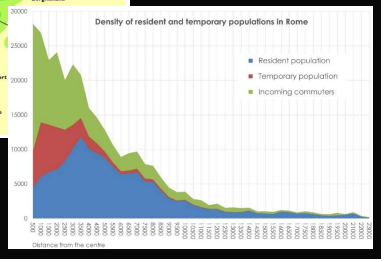


Eg. The spatial distribution of foreigners and segregation/ integration/ inclusion/ assimilation



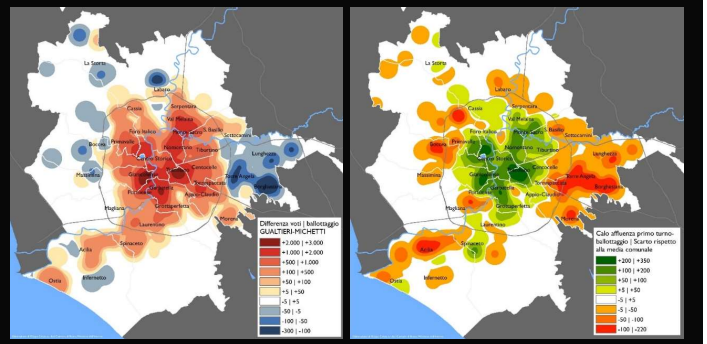
Eg. density of temporary or 'floating' populations in Rome

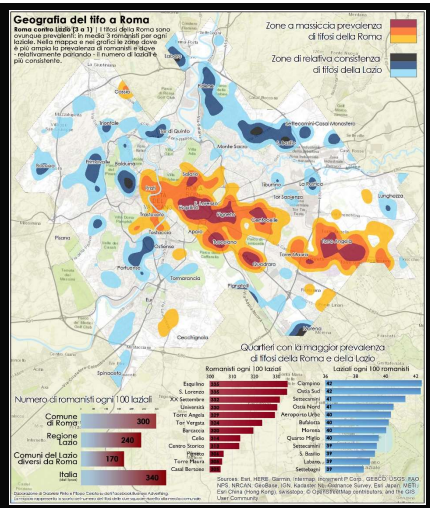




Estimation based on Facebook data

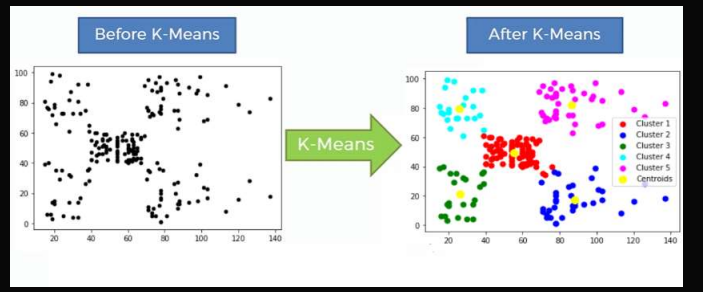
Eg. Electoral results, 2021 municipal elections in Rome





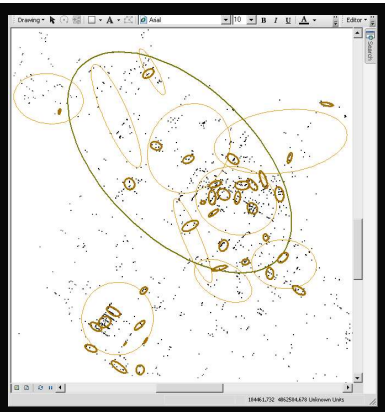
Celata F, Pinto G (2022) *La geografia del tifo a Roma*: www.ricercaroma.it/geografia-tifo-a-roma/

K-Means (QGis): to assign events to a pre-determined number of spatial clusters (based on reiterations for all potential sets of centroids until results are stable)



NEAREST NEIGHBOR HIERARCHICAL CLUSTER

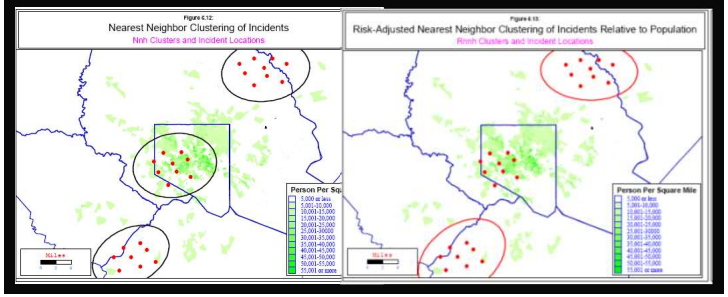
Constant-distance clustering routine for non-weighted events, **hierarchical:**
first order clusters are considered points which may cluster at the second order and so on, until criteria are satisfied (for each order).



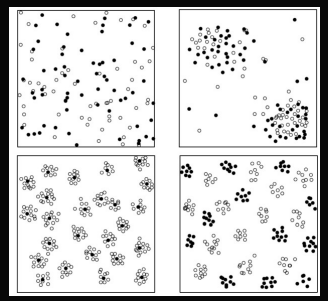
Output (dbf, shp): n. cluster, mean center, deviational ellipse and convex hull (spezzata) of points beloning to each cluster, area and cluster density.

Risk-Adjusted Nearest Neighbor Hierarchical Spatial Clustering

Clustering index in which the probability of identifying clusters for certain categories of events is assessed in relation to the spatial distribution of all events, by using an interpolation between the (kernel) density surfaces of the primary file (e.g. crimes) and the secondary files (eg. population)



Bivariate point patterns : co-agglomeration, co-location, competition/cooperation, etc.: Bivariate/Cross K function, Pairwise interaction point process.. (Geoda, R, Crimestat)



Multi-variate point patterns… (…). -> Bivariate point patterns analysis for each couple of patterns

Spatial autocorrelation

Why to account for **spatial autocorrelation**:

- To estimate the degree to which nearby features potentially **influence** each other (=interaction, interdependence, attraction, contagion, clustering, segregation, etc…)

- To understand the process (or the variety of processes..) which explain the **geographical distribution** of intensities

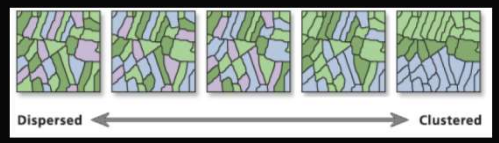
-To verify the degree to which the observations are (not) **statistically indipendent** (eg. autocorrelation reduces the dataset's information content or obscures what is specific about each area, because intensities in one area are partially influenced by what is happening nearby)

- To account for the **MAUP**, by considering values per each feature + those of nearby features (-> spatial econometrics)

- (to test the spatial autocorrelation of models' **residuals**)

- (to assist in the identification of the (spatial) **sample** size)

Spatial auto-correlation: <u>global indexes</u>



MORAN'S I:

Global co-variance index adapted from the analysis of the 'memory' effect in time series (Moran '40s, Whittle 1954).

Measures the "gobal" degree of similarity between the (upper and lower) intensities (-/+) of nearby features

Moran's I

$$I = \frac{N \sum_i \sum_j W_{ij} (X_i - \bar{X})(X_j - \bar{X})}{(\sum_i \sum_j W_{ij}) \sum_i (X_i - \bar{X})^2}$$

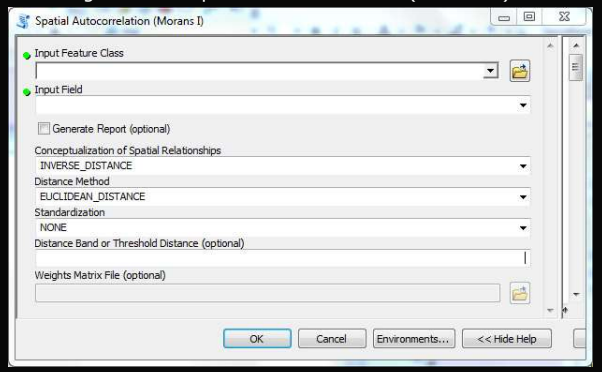
Xi – X = intensity in point Xi – average intensity

(Xi-X)(Xj-X): Cross-product, high if values are similar

Wij: spatial weights (/influences) matrix *

Clustered/high autocorrelation if I is high (I>0), dispersed/low autocorrelation if I is low (I<0), vs. the CSR hypothesis Iexp=-[1/(n-1)]

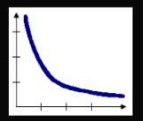
Eg. ArcGIS: Spatial autocorrelation (Moran's I)



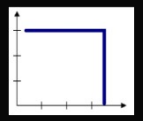
Vs. Geary's autocorrelation index: the probability for high or low values (+) to be clustered or dispersed (similar to Average Nearest Neighbour). Moran is more robust

Conceptualization of spatial relationships:

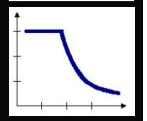
Inverse distance (squared): spatial relationships between features are inversely proportional to their (squared) distance. Computational problems with small distances (crimestat: "adjust for small distances") and no threshold (n to n)



Fixed distance band: within the threshold (band) any feature weights 1. Appropriate in the case of non-uniform polygons, and for large point datasets.



Zone of indifference: neighbors (or features within the distance threshold) weight 1. Other features' weight is inversely proportional to their distance. Appropriate as above, when the influence of distant features is relevant. Computational problems.



Polygon contiguity (adjacency!): considers only bordering features (1 if bordering, 0 all the others). Appropriate only for regular polygons (original Moran's I. Generalized by Cliff and Ord 1973. Widely used in spatial econometrics)

More conceptualizations of spatial relationships..:

K_NEAREST_NEIGHBORS: considers only a "K" number of the most proximate features

CONTIGUITY_EDGES_ONLY: considers only features which share a boundary ("rooke")

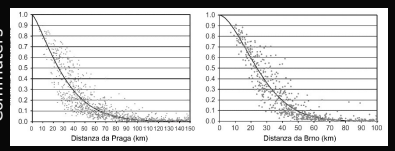
CONTIGUITY_EDGES_CORNERS: considers only features which share a boundary and/or vertex ("queen")

DELAUNAY_TRIANGULATION: create overlapping triangles connecting polygons' centroids, and considers only features which share a triangle's vertex..

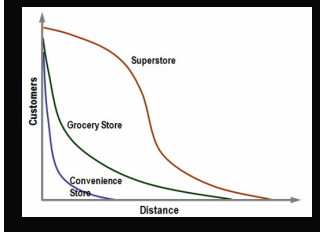
Distance Band or Threshold Distance (mostly for large datasets): threshold beyond which influence is null (with "inverse distance" = i) 0: all features are considered; ii) Empty: applies a default threshold distance (min distance at which any feature has a neighbour); iii) defined by the user

Proximity/distance and spatial interaction: actual and potential interactions (of all kinds..) decrease esponentially as distances increase ('tyranny of distance') = **distance decay function** (inverse distance squared when b=2) -> potential accessibility -> gravitational models

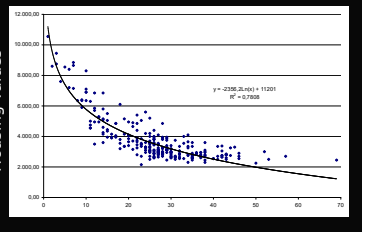
Commuters



$$I = 1 / D^b$$



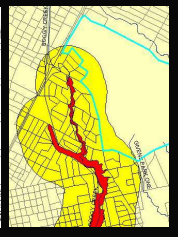
Housing values



(Conceptualizations of distance): Euclidean vs. Manhattan | Geographic vs. functional | Topographic vs. topological

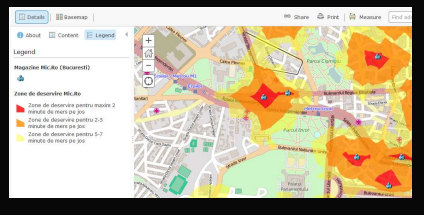
Cost-weighted distance: f('cost' of movement raster) (eg. slope)

(Proximity analysis): buffers



Travel-time or **travel-distance buffers**

Eg. ArcGIS Online (free <1000),
TravelMap (free 20 per minute)
Iso4app (free < 1000)



Spatial Weights Matrix (SWM)

To identify the presence/absence (0/1) or intensity (1/n) of the relationships between each couple of geographical features, as an expression of their distance/spatial relationship.

	1	2	3	4	5	6	7	8	9	10
1	0	1	0	1	0	0	0	0	0	0
2	1	0	0	0	1	0	0	0	0	0
3	0	0	0	1	0	0	1	0	0	0
4	1	0	1	0	1	0	0	1	0	0
5	0	1	0	1	0	1	0	0	1	0
6	0	0	0	0	1	0	0	0	0	1
7	0	0	1	0	0	0	0	1	0	0
8	0	0	0	1	0	0	1	0	1	0

Row standardization: values in the swm are standardized in order for their sum to be = 1 (to avoid the index to be influenced by the different number of 'nearby' features: appropriate in the case of sample data and compulsory in the case of polygon contiguity, because (irregular) polygons have a different and arbitrary number of bordering features.

Test variables:

Z-score = standard deviation / **p-value**

(**Normality:** the Z-score displays a 'normal' distribution?)

LOCAL INDEXES OF SPATIAL AUTOCORRELATION

To measure the degree of autocorrelation for each geographical feature: how much and where?

Anselin local of Moran's I (Anselin L. 1995, Local indicators of spatial association – LISA. *Geographical Analysis* 27, 93-115)

To attribute to each feature a degree of high/low autocorrelation based on its (high/low) intensity being similar/dissimilar to nearby features

$$I_i = \frac{(Z_i - \bar{Z})}{S_Z^2} * \sum_{j=1}^{N} [\, W_{ij} * (Z_j - \bar{Z}) \,]$$

Z: intensity, S: variance, W: spatial weight matrix

Input: polygons (crimestat) and points(ArcGIS)

Output:

Cluster type (COType) identifies (and renders):

- Features which are part of high (HH) or low (LL) values clusters, because nearby features have similar values, and are statistical significant (positive and high z-score).

- "outlier" features, with high or low values, surrounded by features with low (HL) or high (LH) values, and are statistical significant (low and negative z-score)

Local indexes of spatial autocorrelation (2):

Getis-Ord Gi, high/low clustering (Hot Spot Analysis)

Identifies features which are part of "hot spots": areas with unusual clustering of high or low values (Cliff & Ord, Spatial autocorrelation, 1973), based on the value of the GiZScore (categorized according to the standard deviation: the higher the GiZscore, the more nearby features have high values, and viceversa.

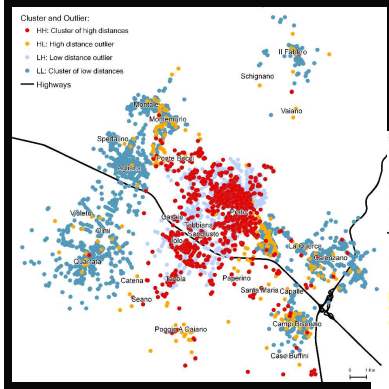
(You may do a density map of using the Z-Score as weight)

Cautions:

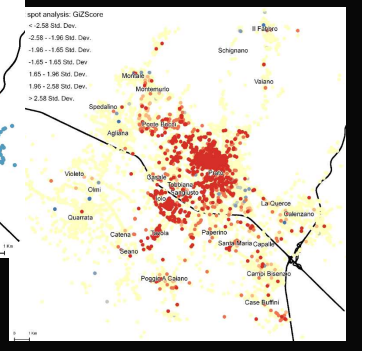
- reliable only with large dataset (>30 features)

- test problems (the significativity test is based on global indexes of spatial autocorrelation)

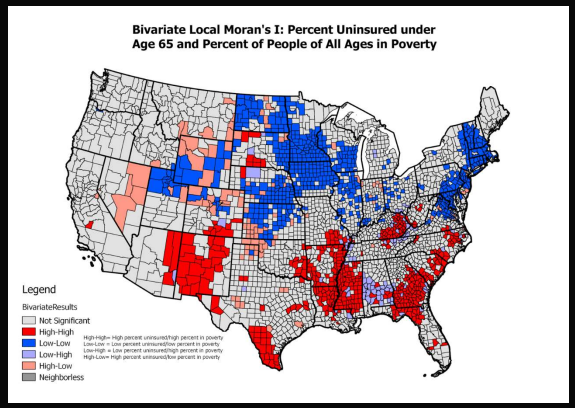
Local indicators of spatial association (LISA)

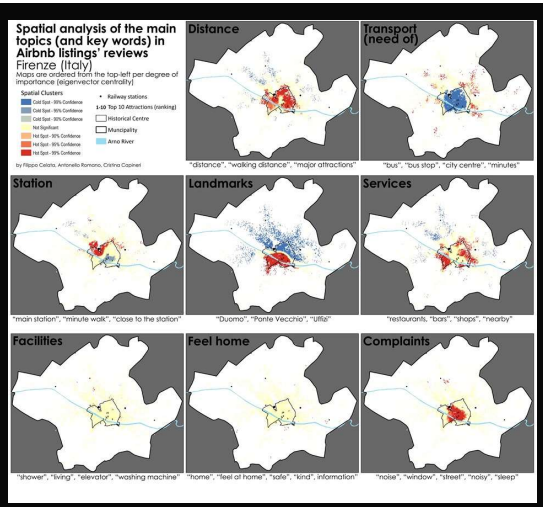


Anselin local of Moran's I



Getis-Ord Gi (Hot Spot Analysis)





Optimized Hot Spot Analysis (ArcGIS)

Optimized Hot Spot Analysis: simplified tool to identify hot spots and cold spots, based on some automatic preliminary operations (e.g. incident data aggregation*), the identification of the 'optimal' parameters, eg. the most appropriate scale of distance (based on the average distance between features); and the execution of tests of significance**

* Methods of aggregation ('Collect events')

** p-value and 'False discovery rate correction' (reduces the critical p-v. based on the number of observations and false positives)

Incident Data Aggregation Method (optional)

The aggregation method to use to create weighted features for analysis from incident point data.

- COUNT_INCIDENTS_WITHIN_FISHNET_POLYGONS—A fishnet polygon mesh will overlay the incident point data and the number of incidents within each polygon cell will be counted. If no bounding polygon is provided in the Bounding Polygons Defining Where Incidents Are Possible parameter, only cells with at least one incident will be used in the analysis; otherwise, all cells within the bounding polygons will be analyzed.
- COUNT_INCIDENTS_WITHIN_AGGREGATION_POLYGONS—You provide aggregation polygons to overlay the incident point data in the Polygons For Aggregating Incidents Into Counts parameter. The incidents within each polygon are counted.
- SNAP_NEARBY_INCIDENTS_TO_CREATE_WEIGHTED_POINTS—Nearby incidents will be aggregated together to create a single weighted point. The weight for each point is the number of aggregated incidents at that location.

Spatial analysis to identify the distribution and clustering of primary schools' quality in Rome

Dataset: desktop/spatial22/2nd/Schools_Roma_XY_dprv.dbf

Index: «DPRV_NORM»

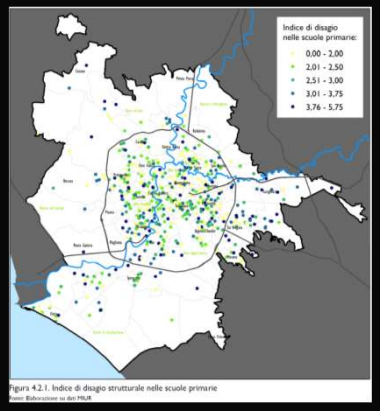


Figura 4.2.1. Indice di disagio strutturale nelle scuole primarie
Fonte: Elaborazione su dati MIUR

Spatial analysis to identify the distribution and clustering of primary schools' quality in Rome

1. Add spatial22/2nd/Zone_urbanistiche.shp to a new blank map
2. Add spatial22/2nd/Schools_Roma_XY_dprv.dbf to the workspace, right click on the layer to 'display XY data' and add the layer to the map
3. Using selection/select by attributes, in the 'events' layer, select all primary schools, and export the selection as a shapefile with the same coordinate system as the dataframe.
4. Perform a global spatial autocorrelation analysis of the variable 'DPRV_INDX' using the 'spatial autocorrelation' tool (Moran's I) in arctoolbox/spatial statistics/analyzing patterns, setting: 'fixed distance band', row standardization, generate report, distance band = 4,000 mts.
5. Perform a local spatial autocorrelation analysis of 'DPRV_INDX' using arctoolbox/spatial statistics/mapping clusters/'optimized hotspot analysis'
6. Perform a local spatial autocorrelation with the same inputs and parameters using 'cluster and outlier analysis'
7. (Describe and interpret the results in a word file)

Spatial analysis of to the distribution and clustering of high-skilled workers in Rome

Dataset: rm_census_dt.shp (census tracks' data)

Relevant variables (on the right end of the table):

- 'tipatt_mda': number of high-skilled workers per census track
- 'tipatt_pcma': percentage of high-skilled workers per census track

1. Produce a density map of high-skilled workers in Rome.
2. Assess the global spatial autocorrelation of zones with a very high or low incidence of high-skilled workers.
3. Identify the zones where high-skilled and low-skilled workers cluster.
4. (Export and describe the results in a word file)

Perform the same analysis for the incidents of graduates (variable 'P47') **over the total population** (variable 'P1')