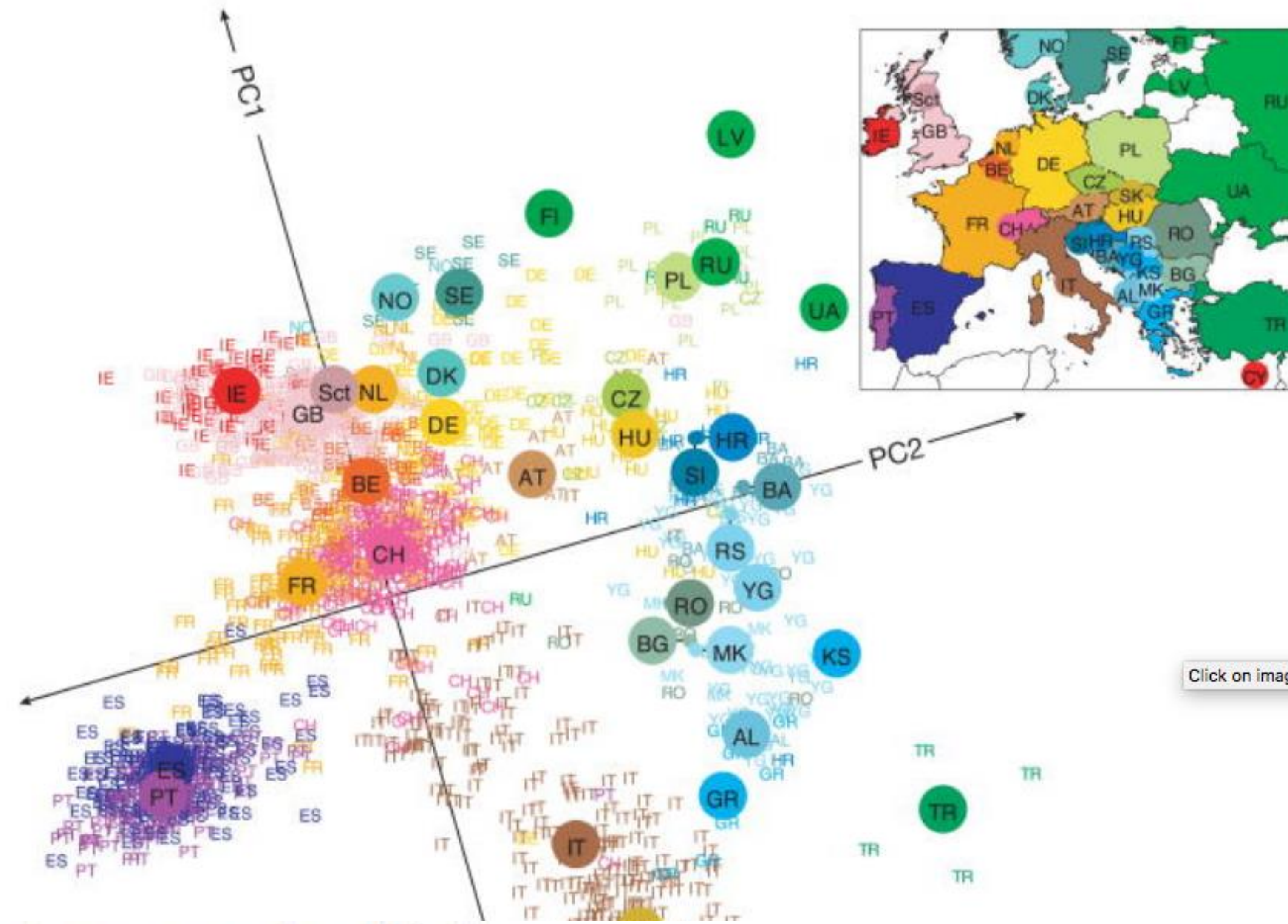
a



Click on image

Genes mirror geography within Europe

John Novembre,[1,2] Toby Johnson,[4,5,6] Katarzyna Bryc,[7] Zoltán Kutalik,[4,6] Adam R. Boyko,[7] Adam Auton,[7] Amit Indap,[7] Karen S. King,[8] Sven Bergmann,[4,6] Matthew R. Nelson,[8] Matthew Stephens,[2,3] and Carlos D. Bustamante[7]

Physalia
Courses

Global and local spatial autocorrelation

Dr Stéphane Joost – Oliver Selmoni (Msc)
Laboratory of Geographic Information Systems (LASIG)
Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

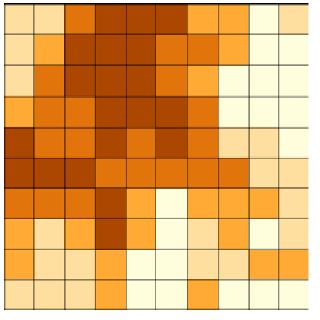
OUTLINE

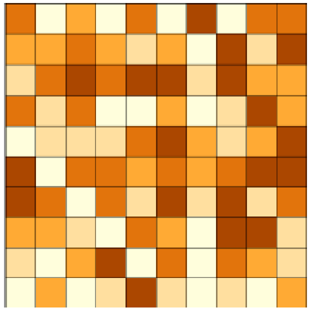
- Introduction: spatial dependence
- Measuring spatial dependence : Moran’s I
- Weighting
- Significance
- Local spatial autocorrelation
- Examples
- Conclusion

MEASURING SPATIAL DEPENDENCE

- We want to measure how similar are the different values of a given variable for a set of spatially distributed individuals…
  - →To quantify the spatial regularity of a given phenomenon
  - →To determine the range of spatial dependence

What is spatial dependence ?



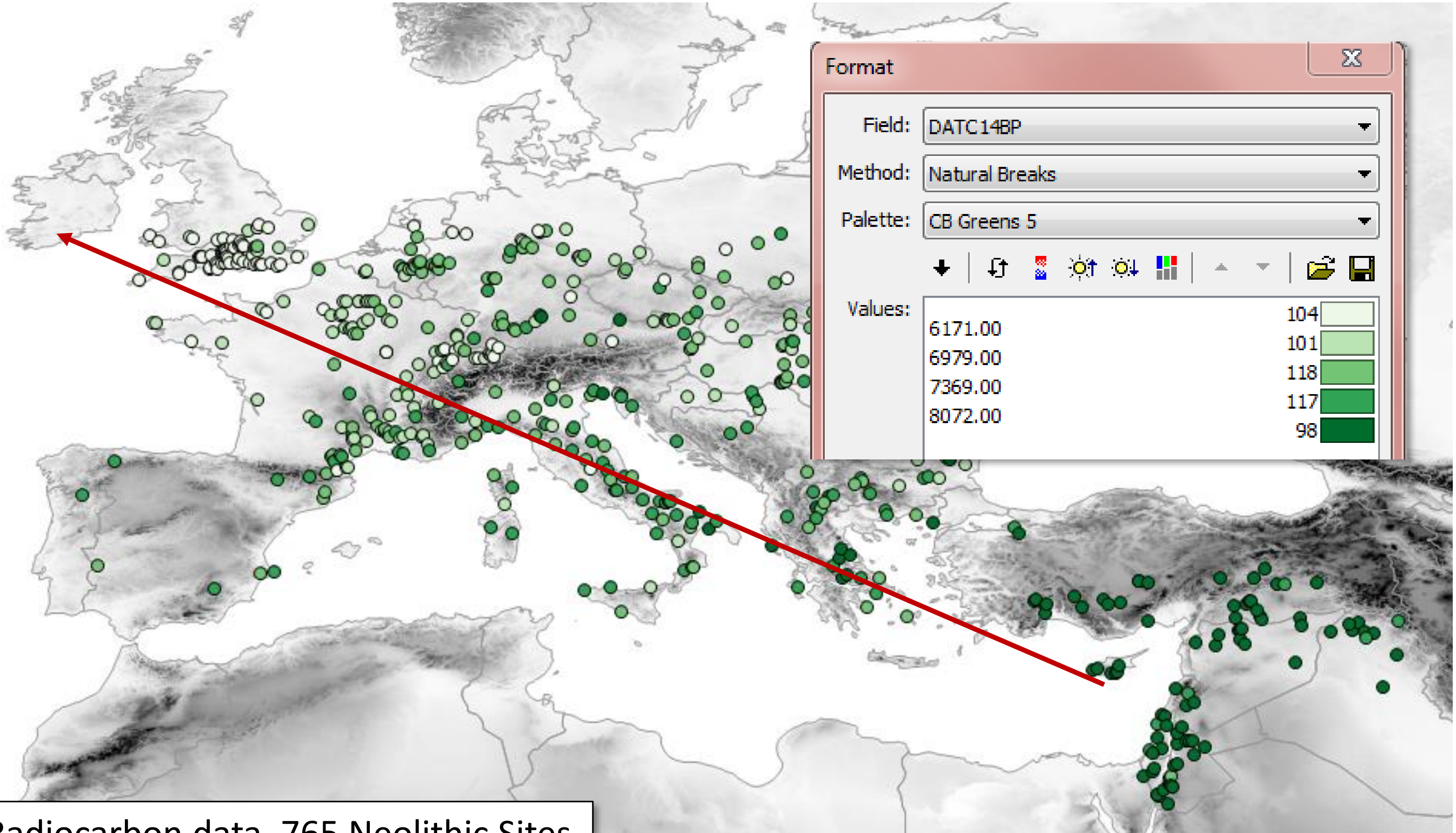


Spatial dependence

- A yellow people is more likely to interact with another yellow people
- Similarly, a r&w people is more likely to interact with another r&w people
- The membership determined the spatial distribution of people
- Spatial dependence induced by this membership is perceptible in space through colors

Tracing the Origin and Spread of Agriculture in Europe

(Pinhasi et al. 2005, PLOS Biology)



Radiocarbon data, 765 Neolithic Sites

SPATIAL AUTOCORRELATION: A PARADOX

- Spatial depdendence can be measured by means of indices of **spatial autocorrelation**
- A **paradox**
- First law of geography: "Everything is related to everything else, but near things are more related than distant things", **W.Tobler** (1970)
- This means that natural phenomena (e.g. temperature) as well as socio-demographic ones (e.g. population density) **are not spatially distributed at random**

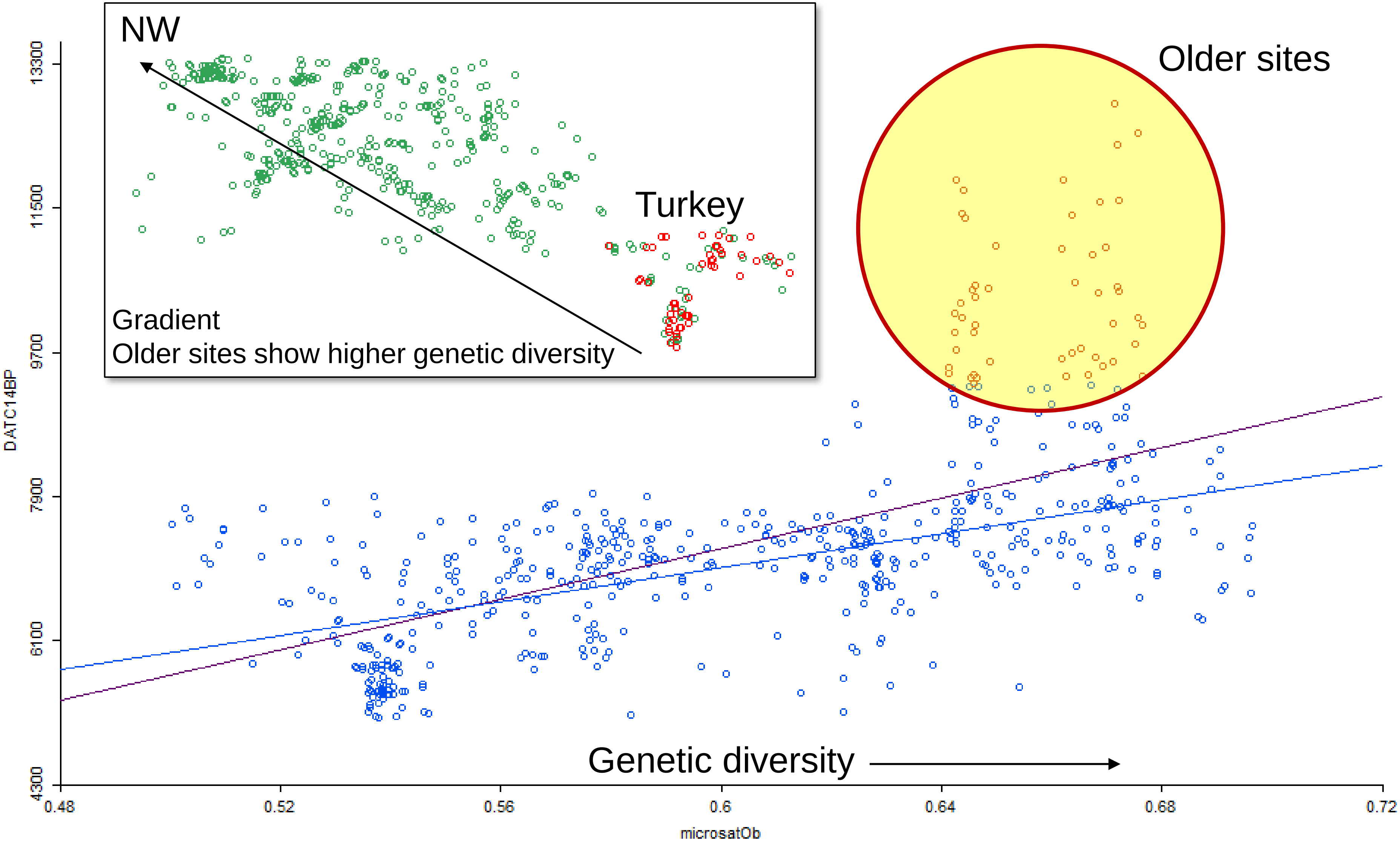
- **But** to measure the spatial structure of these phenomena, we have to use classical statistical tools **requiring a random spatial distribution of samples** and **independence** between them

STANDARD STATISTICS

- Classical statistics are non-spatial
- Based on a neutral geographic space hypothesis
- Geographic space should be the simple **neutral** support for studied phenomena
- Theoretically, the location of a set of observations in space **should not influence their attributes**

Spatial dependance - autocorrelation

- **Geographical space is not neutral** and consequently many statistical tools are not appropriate
- For instance, ordinary linear regressions (OLR) should be implemented **only if observations are selected at random**
- When observations show spatial dependence, estimated values for the whole data set **are distorted/biased**
- Indeed, sub-regions including a concentration of individuals with high values will have an important impact on the model and **lead to a global overestimation** of the variable under study over the whole area
- In other words, a strong correlation between two variables **at a single location** of the territory will influence the measure of the relationship **over the whole territory**



#obs	R^2	const a	std-err a	t-stat a	p-value a	slope b	std-err b	t-stat b	p-value b
538	0.368	-2.24e+003	544	-4.12	4.38e-005	1.58e+004	895	17.7	0
488	0.355	653	390	1.67	0.0948	1.06e+004	649	16.4	0

We have to use standard statistical tools with caution with spatial data

(remember this paradox)

INDICES TO MEASURE SPATIAL AUTOCORRELATION

Developed with standard statistics

- Geary's C
- Ripley's K
- Join Count Analysis
- **Moran's I**

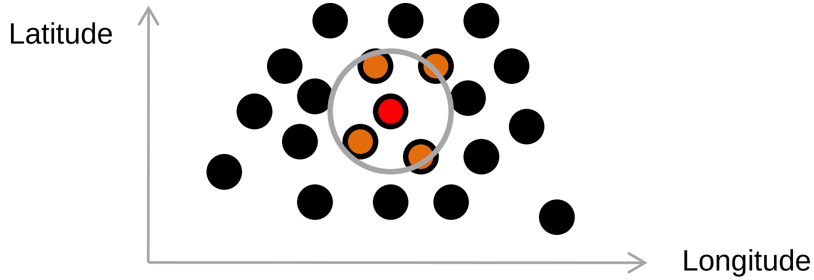
Measuring spatial autocorrelation

- Spatial autocorrelation indices express **the degree of spatial structuring** of a given variable
- Spatial autocorrelation indices permit to **quantify the spatial regularity** of a geographically distributed phenomenon

- Spatial autocorrelation is **positive** when values measured at neighbouring points **resemble each other**
- It is **negative** in case of dissimilarity

Neighborhood relationship and **spatial weighting**

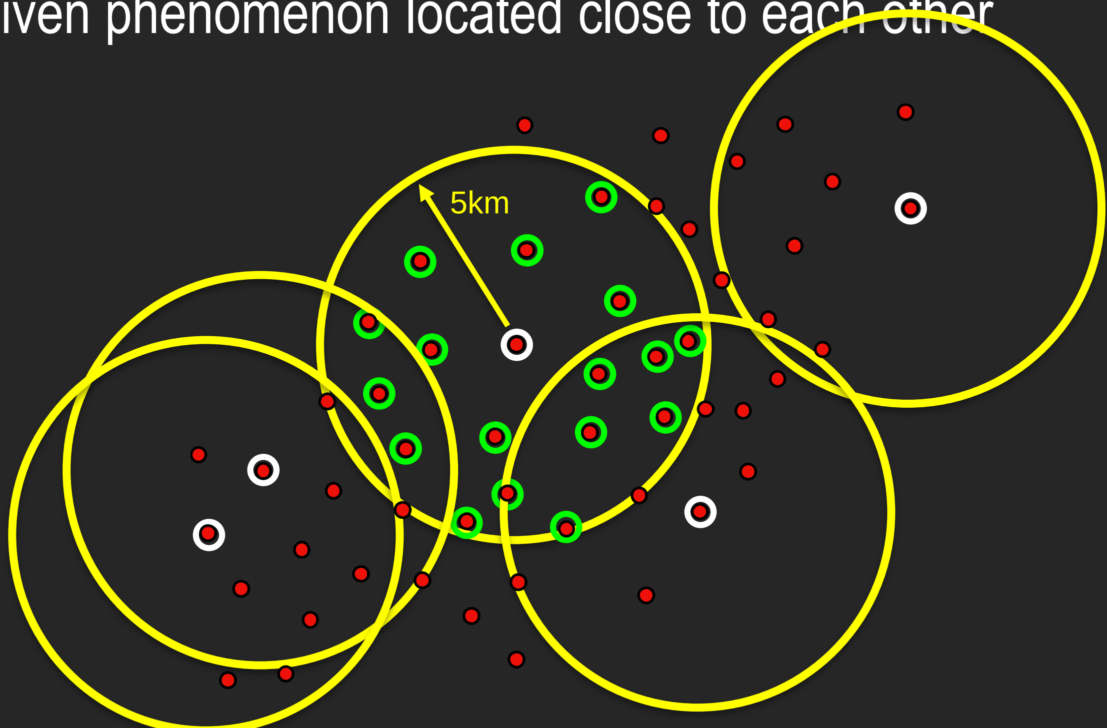
- We want to know how similar is the value *v* of object *o* **in comparison** with the value *v* of other objects located in its neighbourhood ? (to measure spatial dependence)



- We need to define this **neighbourhood**: it may be a distance (100m around each object) or a given number of neighbours (4 nearest neighbours for instance)
- This criterion (100m or 4 nearest neighbours) defines the spatial weight

Neighborhood relationships

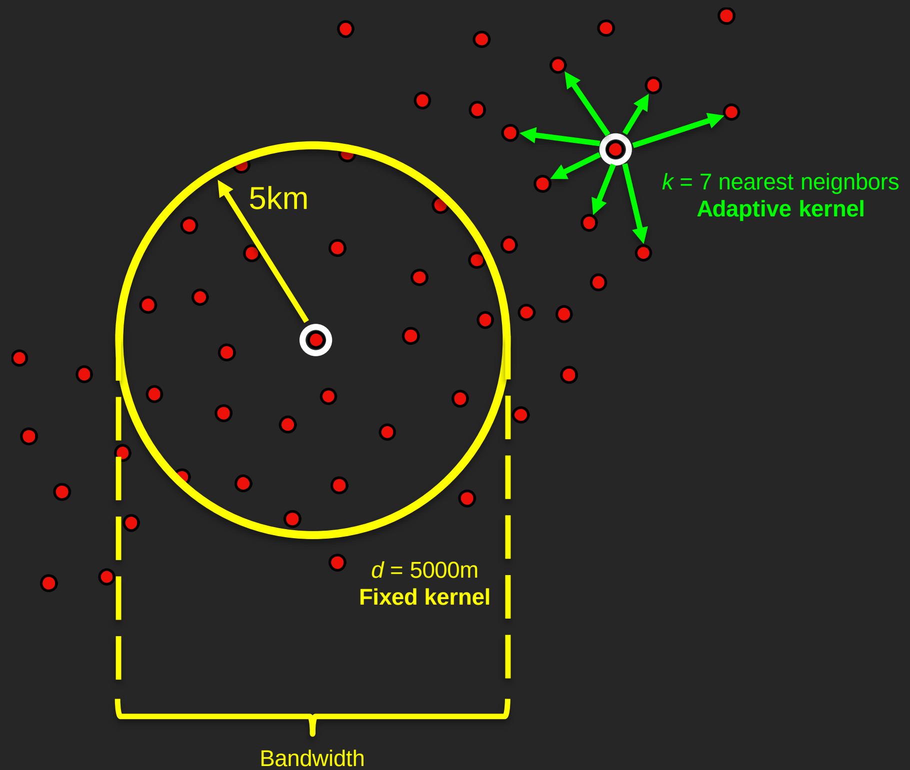
- Spatial autocorrelation is characterized by a correlation between measures of a given phenomenon located close to each other

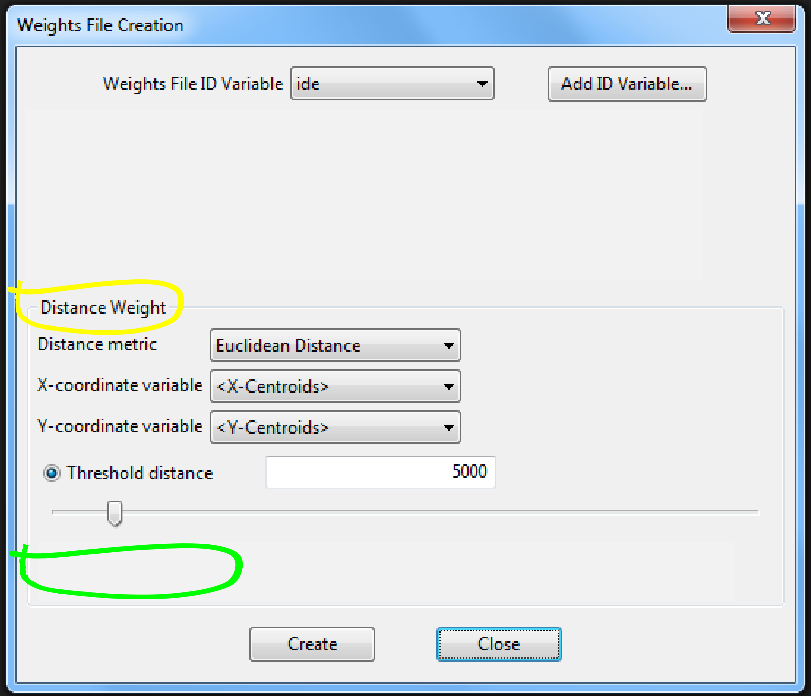


Etc.

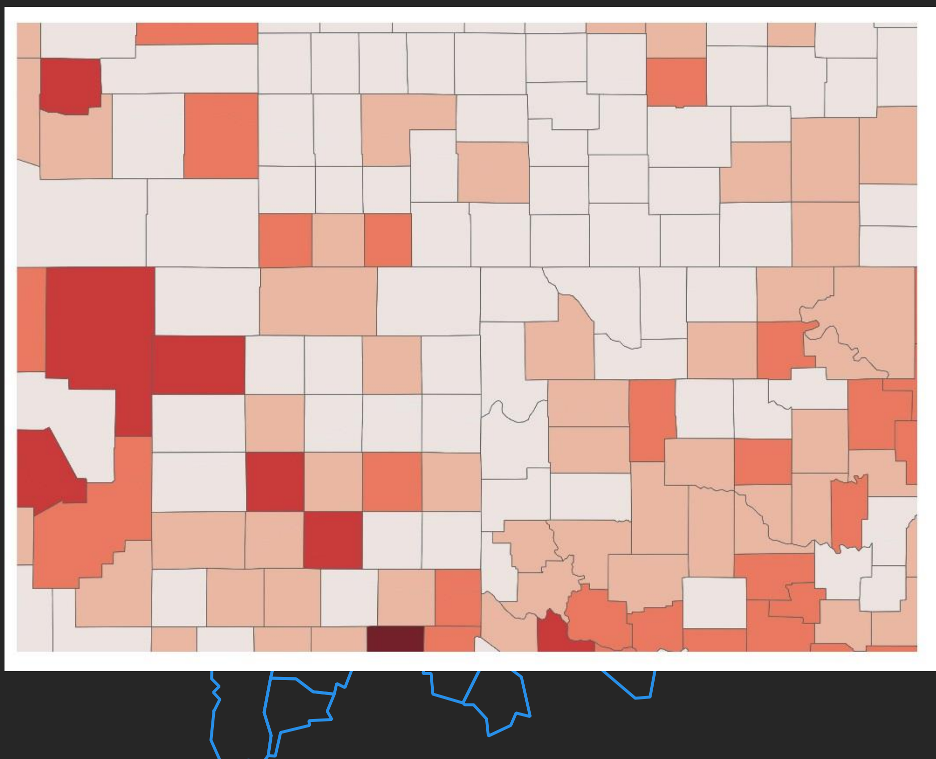
To quantify the spatial dependence and produce a measure of global spatial autocorrelation, it is necessary to take into account the neigborhood of each of the considered geographic objects

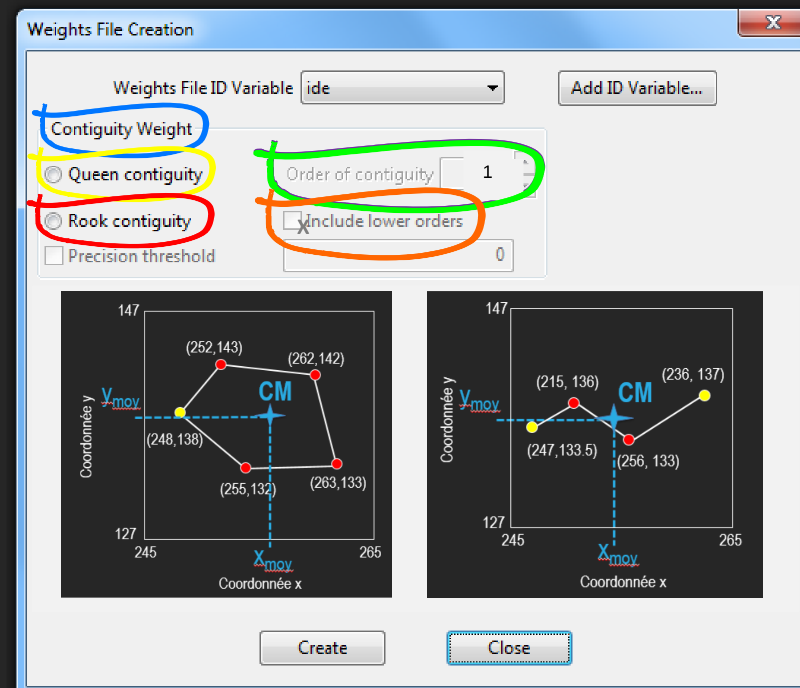
Several ways to define the spatial neighborhood





How to define the neighborhood of polygons





Spatial weighting

Spatial weighting scheme

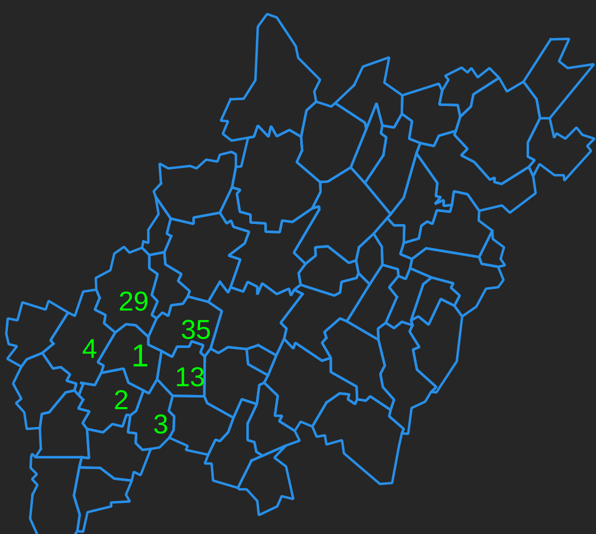
Fixed kernel → Bandwidth *d*

Adaptive kernel → *k* nearest neighbors

Adaptive kernel → Contig order *n*

Spatial weighting file

```
0 54 comvd_prec ide
1 2        1895.70644
1 13       2031.2413
1 4        2062.65071
1 35       2365.3336
1 3        2474.90419
1 29       2533.8499
1 19       3041.1263
...
```



comvd_prec_qu1.gal

```
0 54 comvd_prec ide
1 6
35 29 13 4 3 2
2 5
9 8 4 3 1
3 5
13 9 21 1 2
4 6
8 7 5 1 2 29
5 2
7 4
6 2
10 9
7 3
8 4 5
8 5
10 9 2 4 7
9 5
10 3 2 6 8
10 3
6 8 9
11 8
35 27 26 21 15 13 14 18
```

n = 1

comvd_prec_7k.gwt

```
0 54 comvd_prec ide
1 2        1895.70644
1 13       2031.24132
1 4        2062.65071
1 35       2365.33363
1 3        2474.90419
1 29       2533.84993
1 19       3041.12639
2 3        1813.98042
2 1        1895.70644
2 9        1992.83081
2 8        2039.59784
2 4        2479.38678
2 7        2693.1369
2 13       3076.66614
```

k = 7

MORAN'S I

- Moran's autocorrelation coefficient I is an extension of Pearson product-moment correlation coefficient

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

Where

- $\boldsymbol{N}$ is the number of observation units
- $\boldsymbol{W_{ij}}$ is a spatial weight applied to define the comparison between locations $\boldsymbol{i}$ and $\boldsymbol{j}$
- $\boldsymbol{X_i}$ is the value of the variable at a location i
- $\boldsymbol{X_j}$ is the value of the variable at a location j
- $\boldsymbol{\bar{X}}$ is the mean of the variable

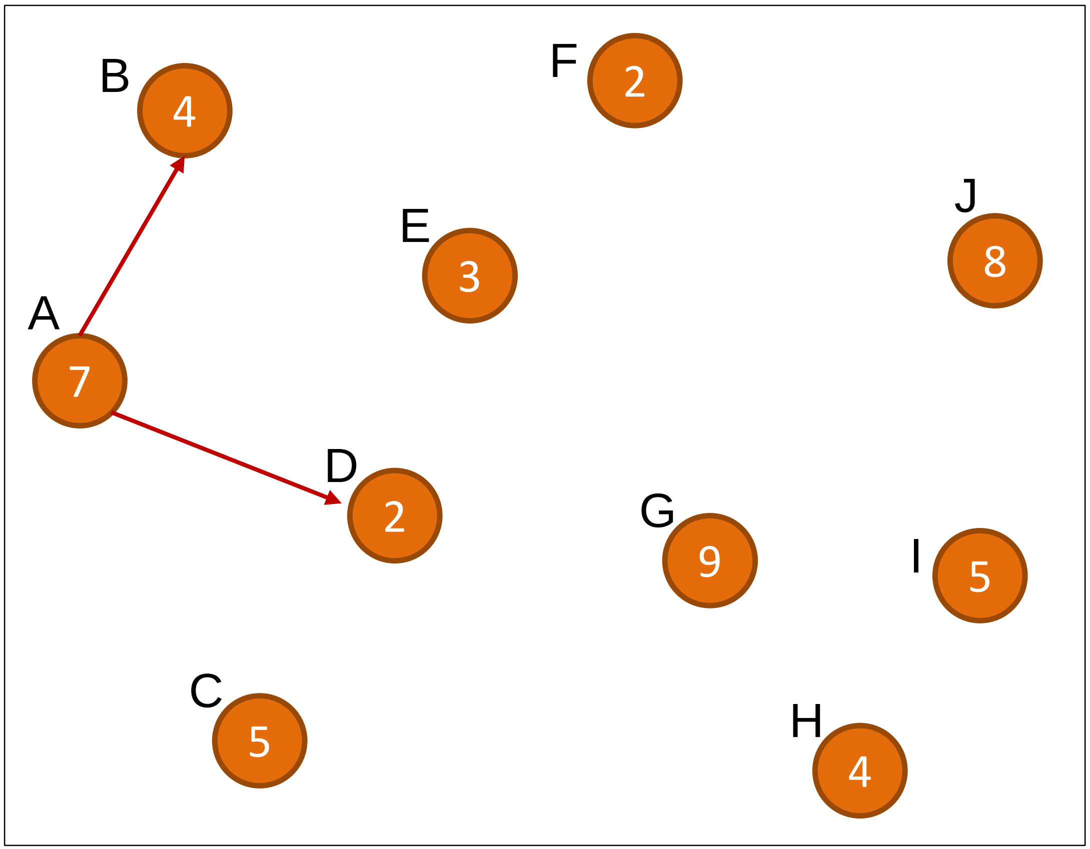
Moran’s I as a regression coefficient

- Anselin (1996): Moran’s I can be interpreted as a regression coefficient
- $\rightarrow$ regression of *WZ* on *Z*
- The *weighted value of Z* (mean of *Z* according to the weighting criteria) **on** *Z* (the observed value at the point of interest)
- This interpretation provides a way to visualize the linear association between *Z* and *WZ* in the form of a bivariate scatterplot
- Anselin (1996) referred to this plot as the Moran scatterplot
- He pointed out that **the least squares slope** in a regression through the origin **is equal** to **Moran’s I**

PROCESSING MORAN'S I

Regression of Y on X

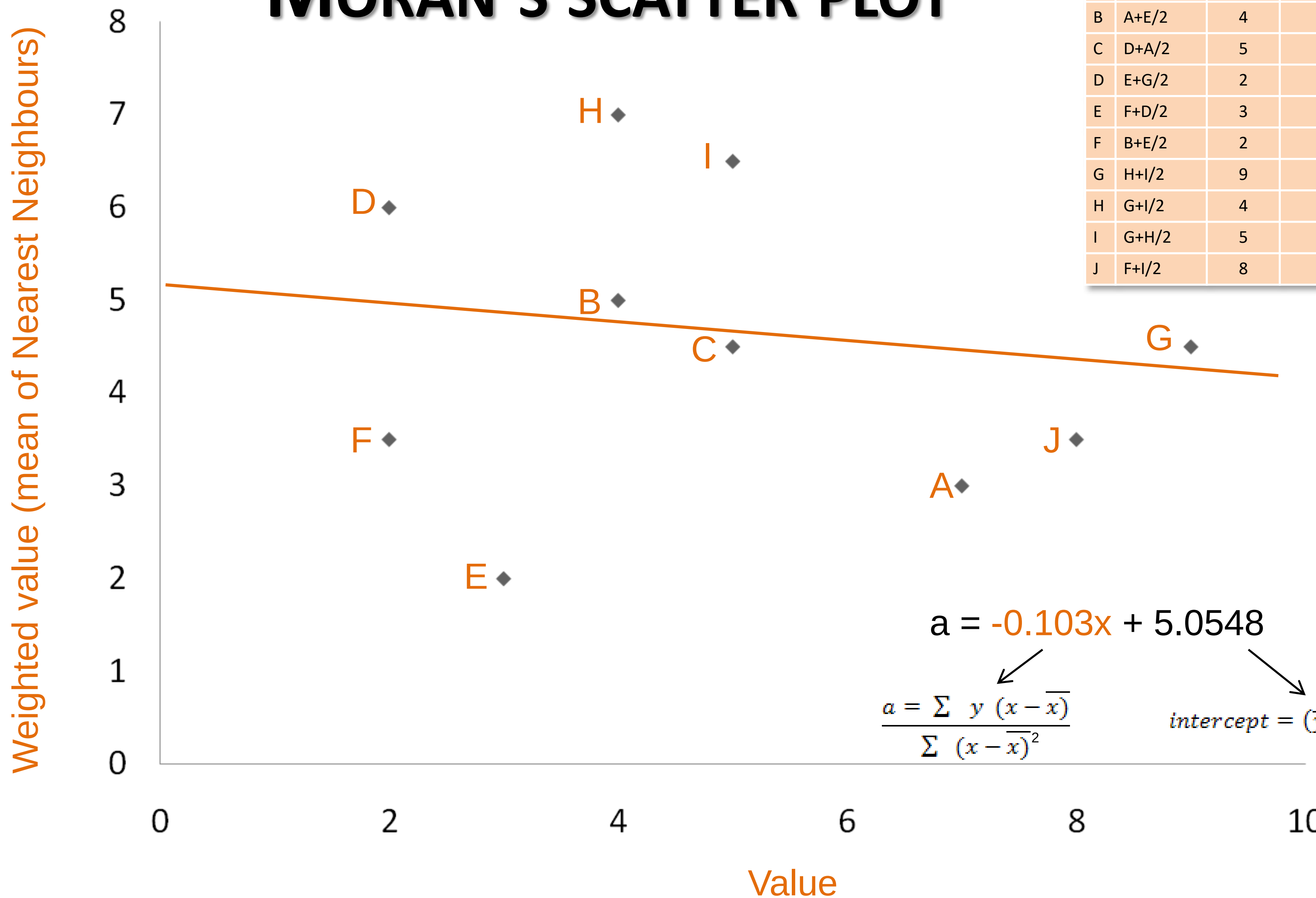
Weighting criteria: 2 nearest neighbours
The value of the variable is displayed in the centroids



	Weighting	Value (X)	Mean of nearest neighbours values (Y)
A	B+D/2	7	3
B	A+E/2	4	5
C	D+A/2	5	4.5
D	E+G/2	2	6
E	F+D/2	3	2
F	B+E/2	2	3.5
G	H+I/2	9	4.5
H	G+I/2	4	7
I	G+H/2	5	6.5
J	F+I/2	8	3.5

Moran's scatter plot

	Weighting	Value	Mean of NN
A	B+D/2	7	3
B	A+E/2	4	5
C	D+A/2	5	4.5
D	E+G/2	2	6
E	F+D/2	3	2
F	B+E/2	2	3.5
G	H+I/2	9	4.5
H	G+I/2	4	7
I	G+H/2	5	6.5
J	F+I/2	8	3.5



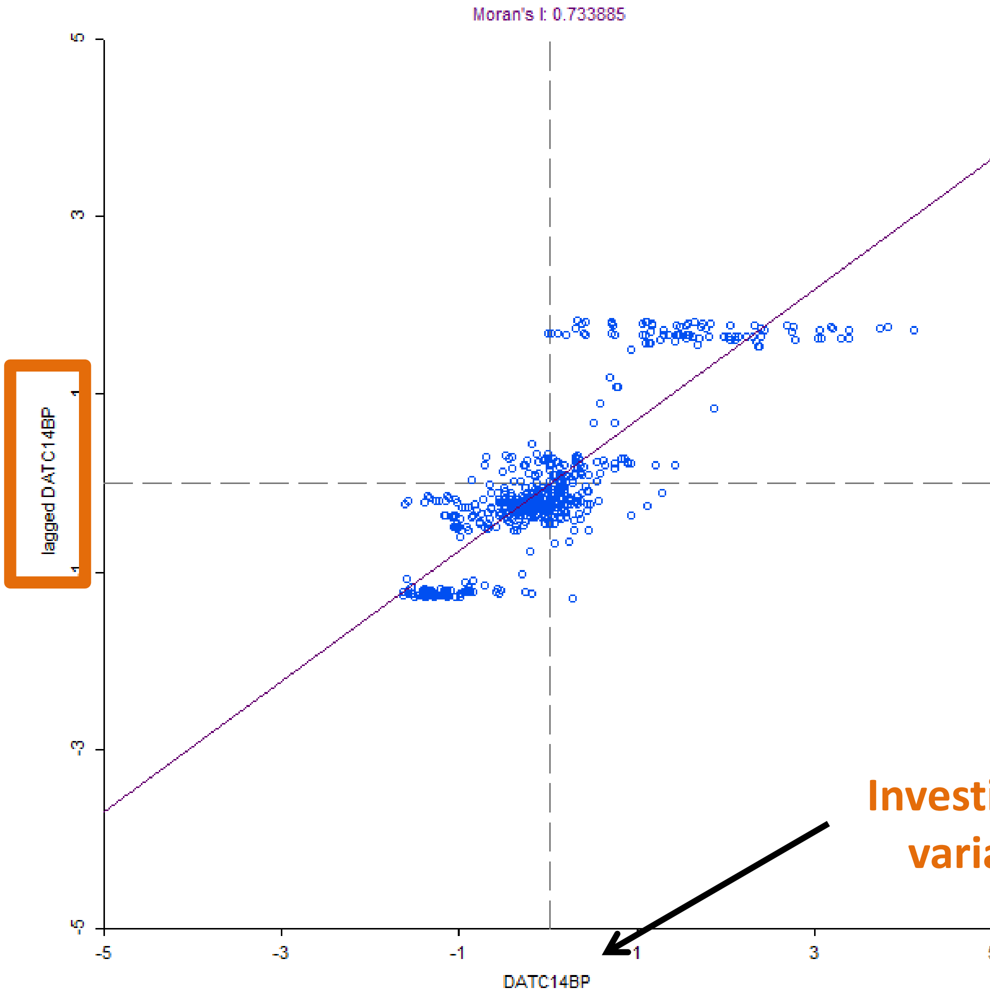
$$a = -0.103x + 5.0548$$

$$a = \frac{\sum y\,(x-\overline{x})}{\sum (x-\overline{x})^2}$$

$$intercept = (\overline{y} - a)\,\overline{x}$$

I = Slope of the regression

The slope of the regression (**0.734**) is Moran's I for the Variable **DATC14BP**



#obs	R^2	const a	std-err a	t-stat a	p-value a	slope b	std-err b	t-stat b	p-value b
538	0.718	-0.0168	0.0198	-0.849	0.396	0.734	0.0199	36.9	0

Weighted variable: mean of neighbouring spatial units according to the selected weighting criteria (50k)

Investigated variable

Moran's I range of values

- Moran's I statistics ranges **from -1 to 1**
- A value close to 1 shows a strong positive spatial autocorrelation
- A value close to -1 shows a strong negative spatial autocorrelation (opposition between individuals)
- **0** = no spatial autocorrelation **= independence between individuals** = neutral geographic space

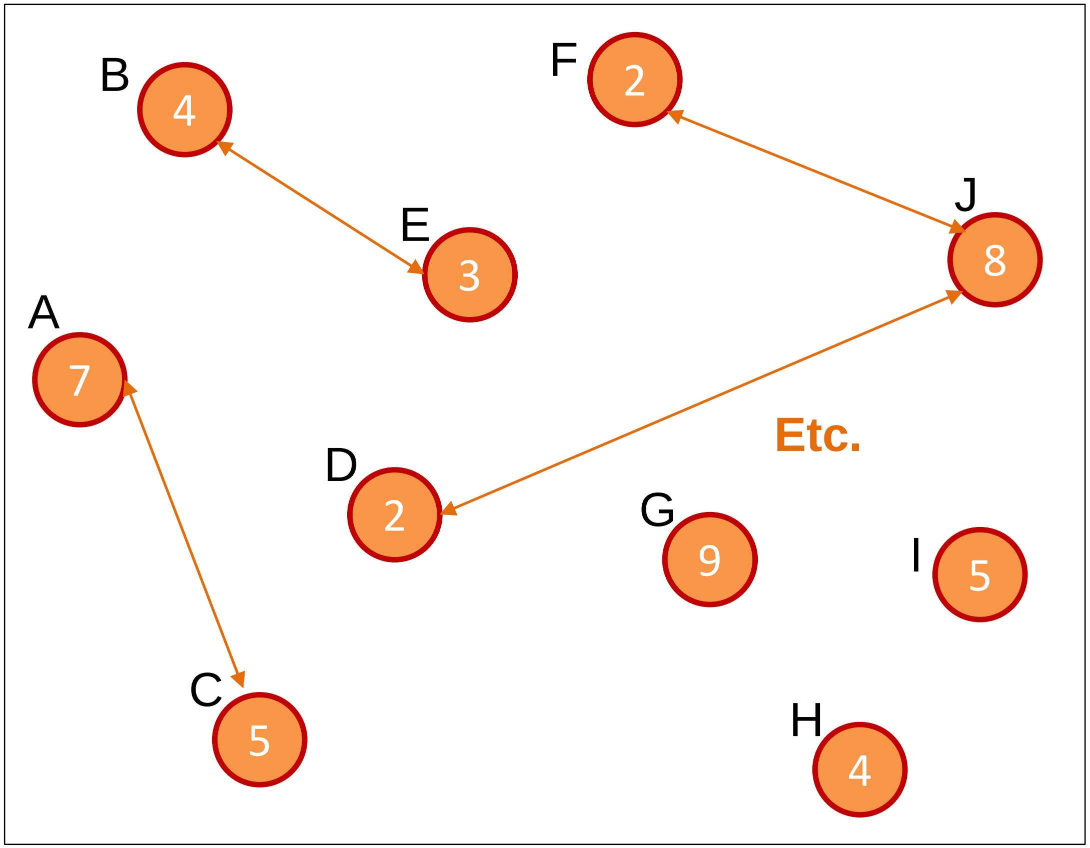
MORAN'S I SIGNIFICANCE

- We need a statistical test to determine whether data attached to individuals are distributed at random over the territory
- We use random **permutations** between individuals to perform this test
- How does the observed situation behave **in comparison with all other possible configurations** ?

RANDOM PERMUTATIONS

- For each run, the **attributes** of all individuals in the data set **are randomly moved** between **all** possible **locations**
- Many runs are performed by means of Monte-Carlo method (e.g. 500, 1000, or more permutations)
- More runs = more significance

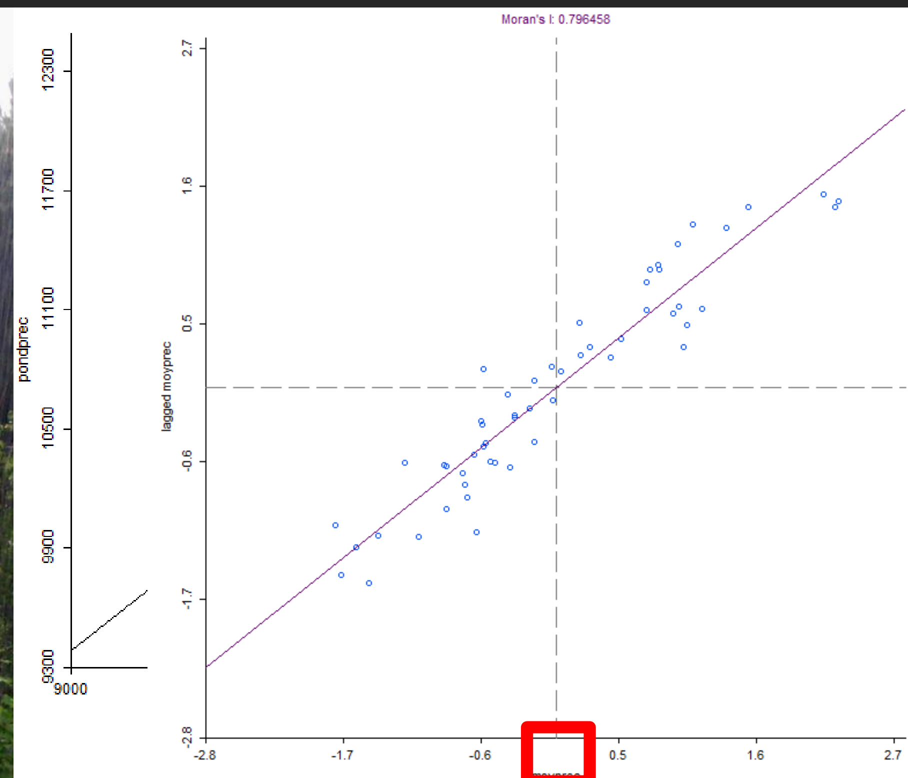
Random permutations



- Moran’s I is **calculated for each run**
- The total number of possible permutations is **n!** (here 10! = 3’628’800)

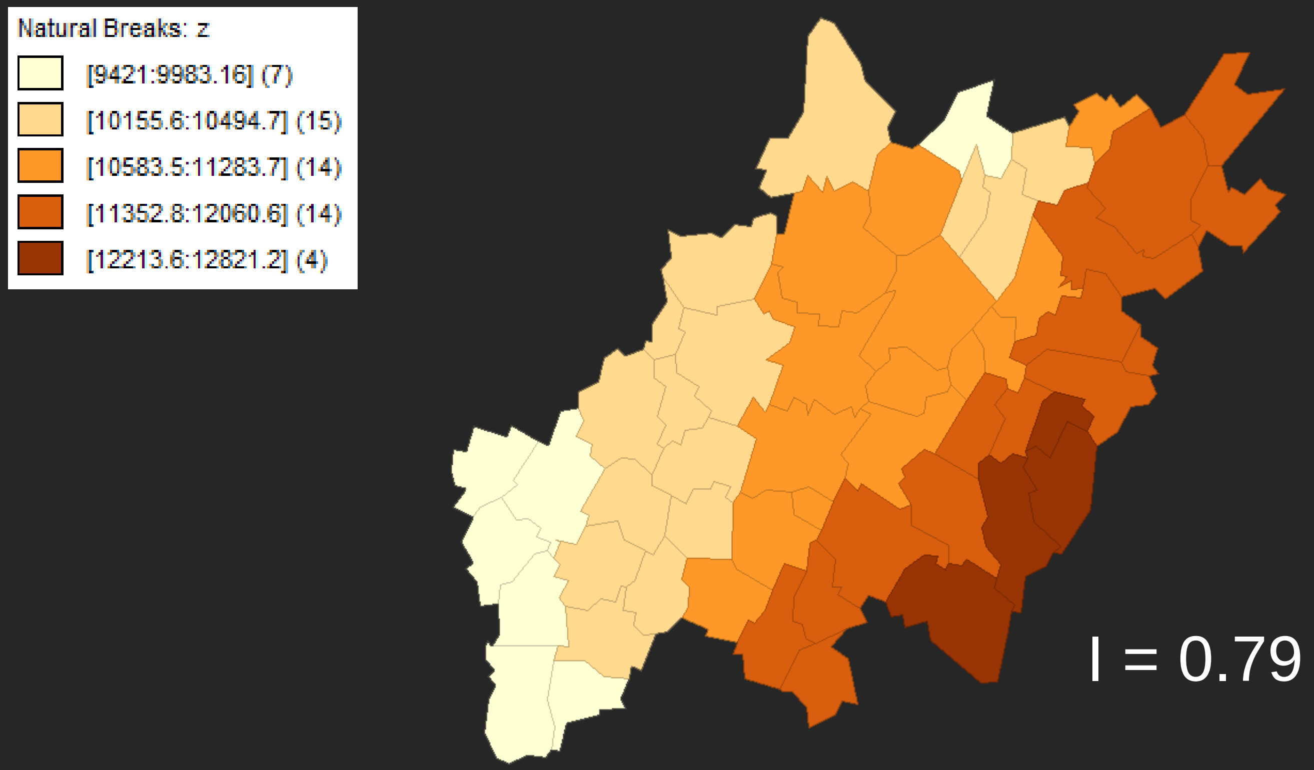
Moran's I as a coefficient of regression

ide >	commune	z	z_barre
1	Bettens	10291.14	10244.1741
2	Bournens	10166.64	10085.3390
3	Boussens	10494.71	10399.5093
4	Daillens	9708.88	9904.0585
5	Lussery-Villars	9642.24	9583.5026
6	Mex (VD)	9983.16	9897.9559
7	Penthalaz	9458.12	9636.0535
8	Penthaz	9557.04	9825.9107
9	Sullens	10374.92	9924.5102
10	Vufflens-la-Ville	9421.00	9971.7072
11	Assens	10763.34	10953.1723
12	Bercher	10413.63	10655.9440
13	Bioley-Orjulaz	10432.68	10526.8663
14	Bottens	11705.89	11404.4981
15	Bretigny-sur-Morrens	11519.39	11430.8364
16	Cugy (VD)	11901.36	11436.0724
17	Dommartin	11519.02	11617.8423
18	Echallens	10583.54	10857.6636
19	Eclagnens	10281.59	10325.2212
20	Essertines-sur-Yverdon	10406.37	10680.8931
21	Etagnières	10732.18	10760.8724
22	Fey	10626.72	10719.9970
23	Froideville	12821.19	12162.2771



...

Visualization of the spatial structure



Moran’s I significance and random permutations

Observed situation : I= I_0

Draw 1: I= I_1

Draw 2: I= I_2

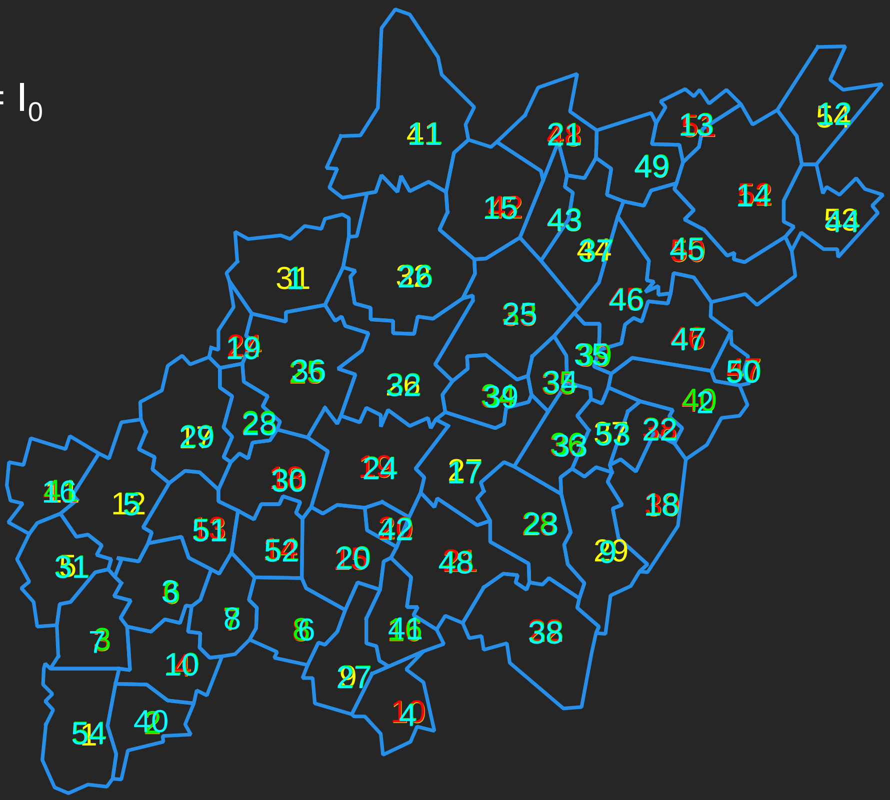
Draw 3: I= I_3

Draw 4: I= I_4

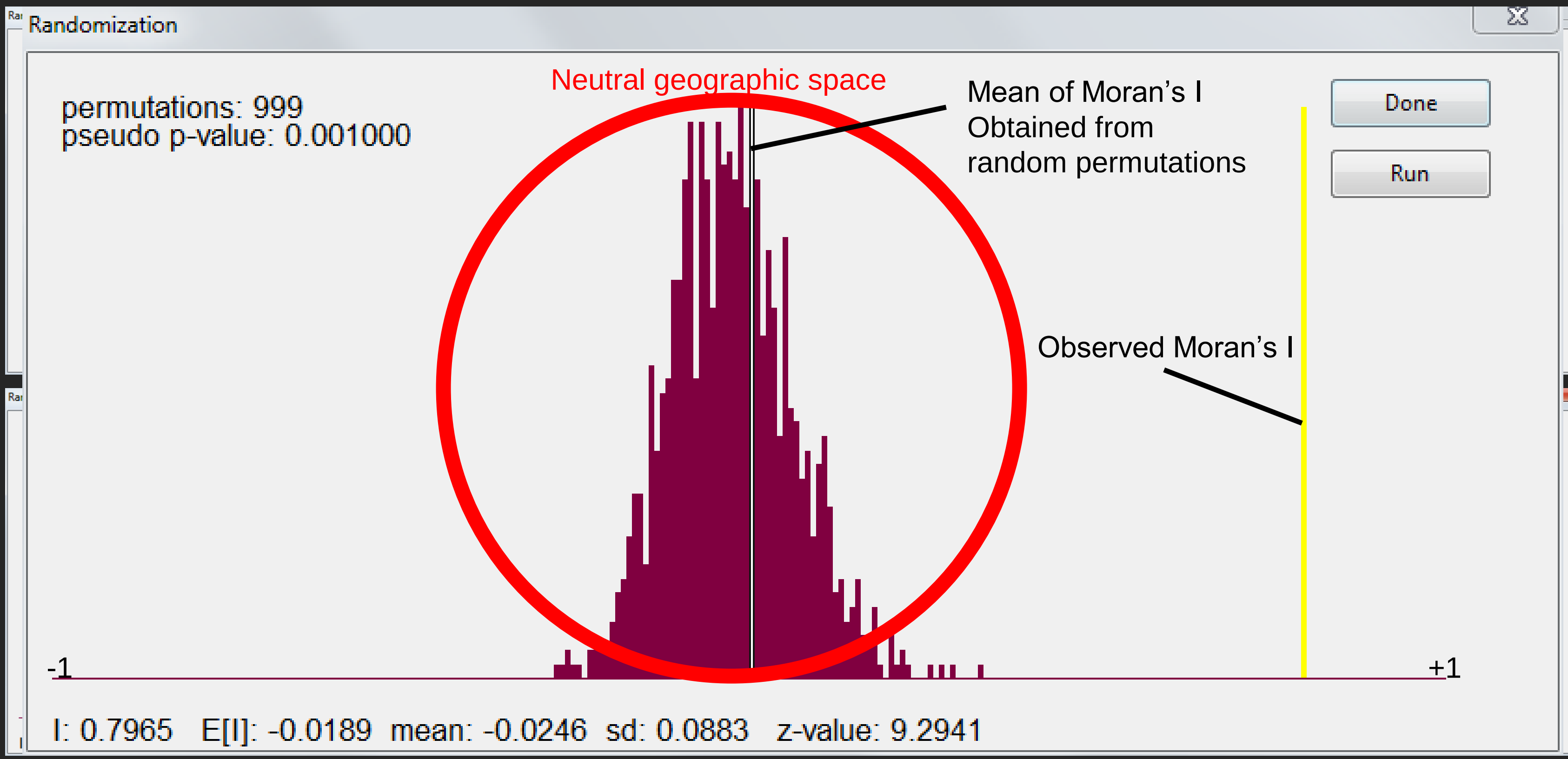
…

…

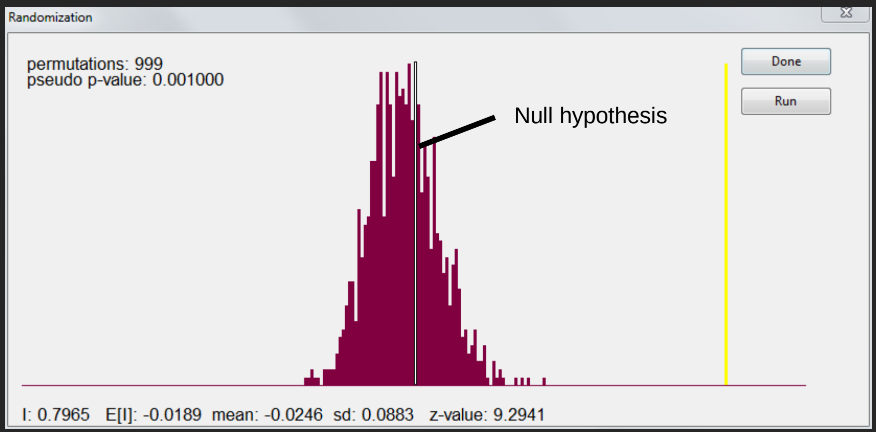
54! possible configurations



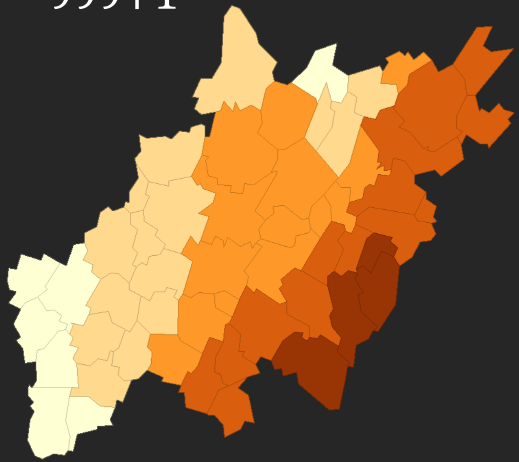
Histogram of random situations and p-values



How to calculate the significance and the p-value



$$\text{p-value} = \frac{0+1}{999+1} = 0.001$$



$$\text{p-value} = \frac{Nb\ I_{al} \geq I_{obs}+1}{Nb\ permutations+1}$$

$$ou\ \frac{Nb\ I_{al}\ \leq\ I_{obs}+1}{Nb\ permutations+1}$$

A Moran's I of 0.79 translates a spatial structure significantly different from a random spatial distribution

LISA : LOCAL INDICATORS OF SPATIAL ASSOCIATION

- It is also possible to calculate **local Moran's indices**
- Moran's I is decomposed into several local coefficients, whose sum over the whole studied area is proportional to the global Moran's I
- The significance is assessed locally (same procedure like Global I except that the value of the point of interest remains fixed, and neighbours are among N-1 values)

Calculation of Local I

45	44	44
43	42	39
38	32	34

Rook weighting scheme

$$I_i = \left[\frac{Z_i}{S^2}\right] \sum_{j=1}^{n} w_{ij} z_j \text{ , } j \neq i$$

y = value
Z = dev. from the mean (40.1)
w = weight

y_i	z_i	w_{ij}	$w_{ij}z_j$
45	4.889	0	0
43	2.889	0.25	0.722
38	-2.111	0	0
44	3.889	0.25	0.972
42	1.889	0	0
32	-8.111	0.25	-2.028
44	3.889	0	0
39	-1.111	0.25	-0.278
34	-6.111	0	0
		1	**-0.611**

Mean = 40.1
S^2 = Variance = 21.8 $\frac{\sum (x-\bar{x})^2}{(n-1)}$

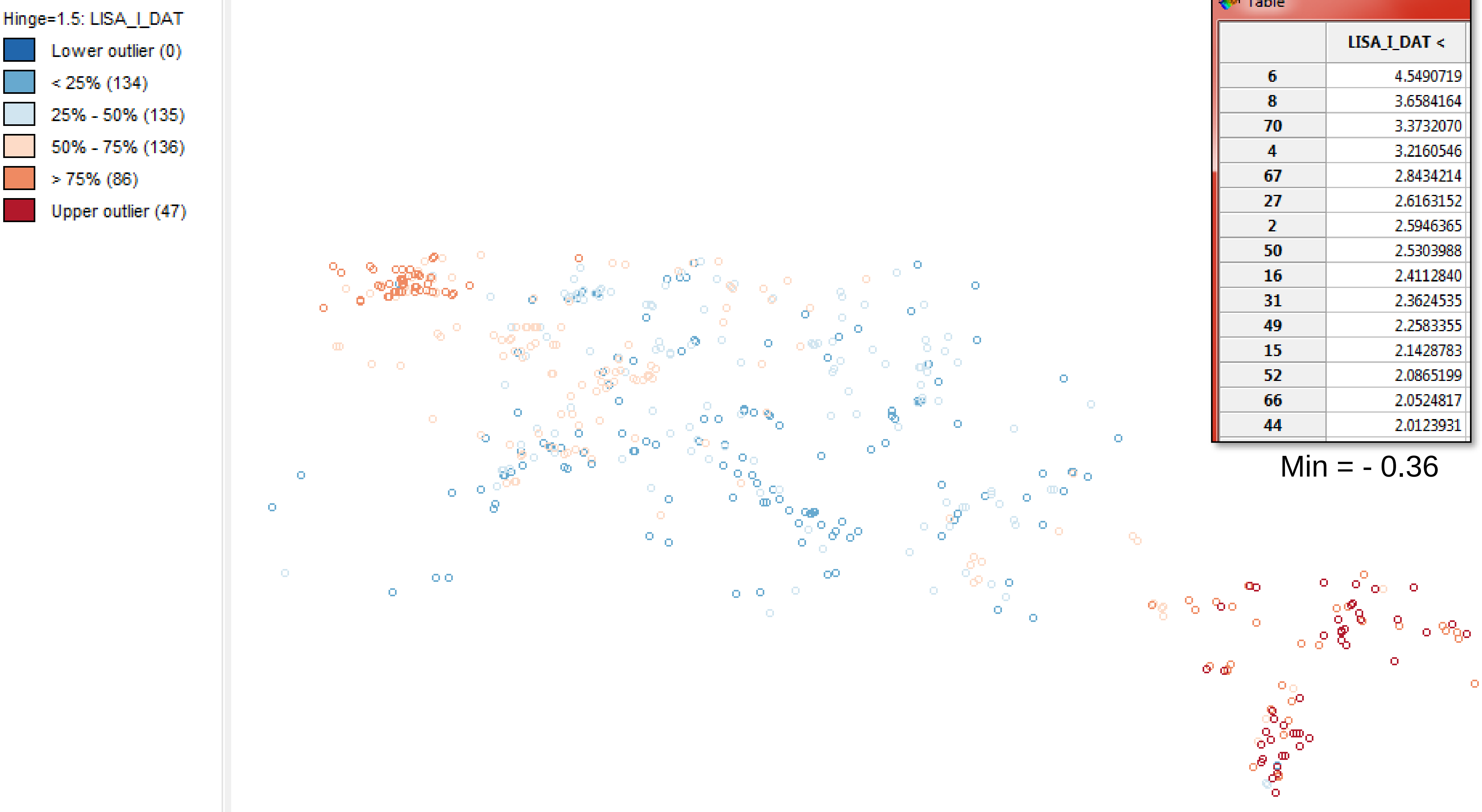
- Here weights are standardized (they sum to 1). We have 4 values (because of the Rook criterion).
- **For this location**, z_i = 42 - 40.111 = 1.889
- The sum of the weights multiplied by the deviations from the mean = -0.611
- I_i = 1.889/21.861 x -0.611 = -0.053
- Local I values sum up to global Moran's I

LISA

Thus it is possible:

- to map indices and corresponding p-values to show how local spatial autocorrelation varies over the territory
- to highlight local regimes of spatial autocorrelation

Local Index of Spatial Association (DATC14BP)

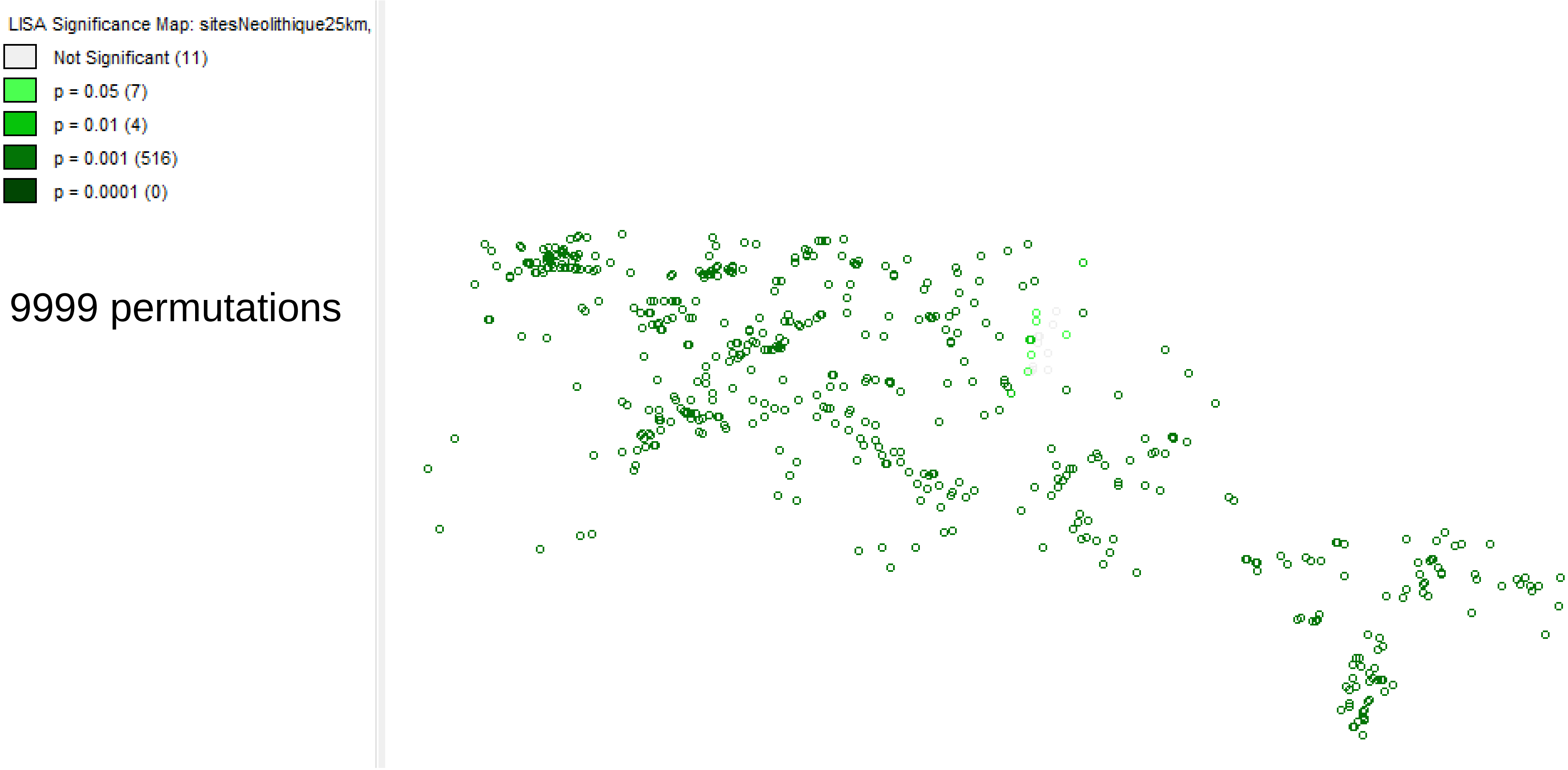


Table

	LISA_I_DAT <
6	4.5490719
8	3.6584164
70	3.3732070
4	3.2160546
67	2.8434214
27	2.6163152
2	2.5946365
50	2.5303988
16	2.4112840
31	2.3624535
49	2.2583355
15	2.1428783
52	2.0865199
66	2.0524817
44	2.0123931

Min = - 0.36

Significance map

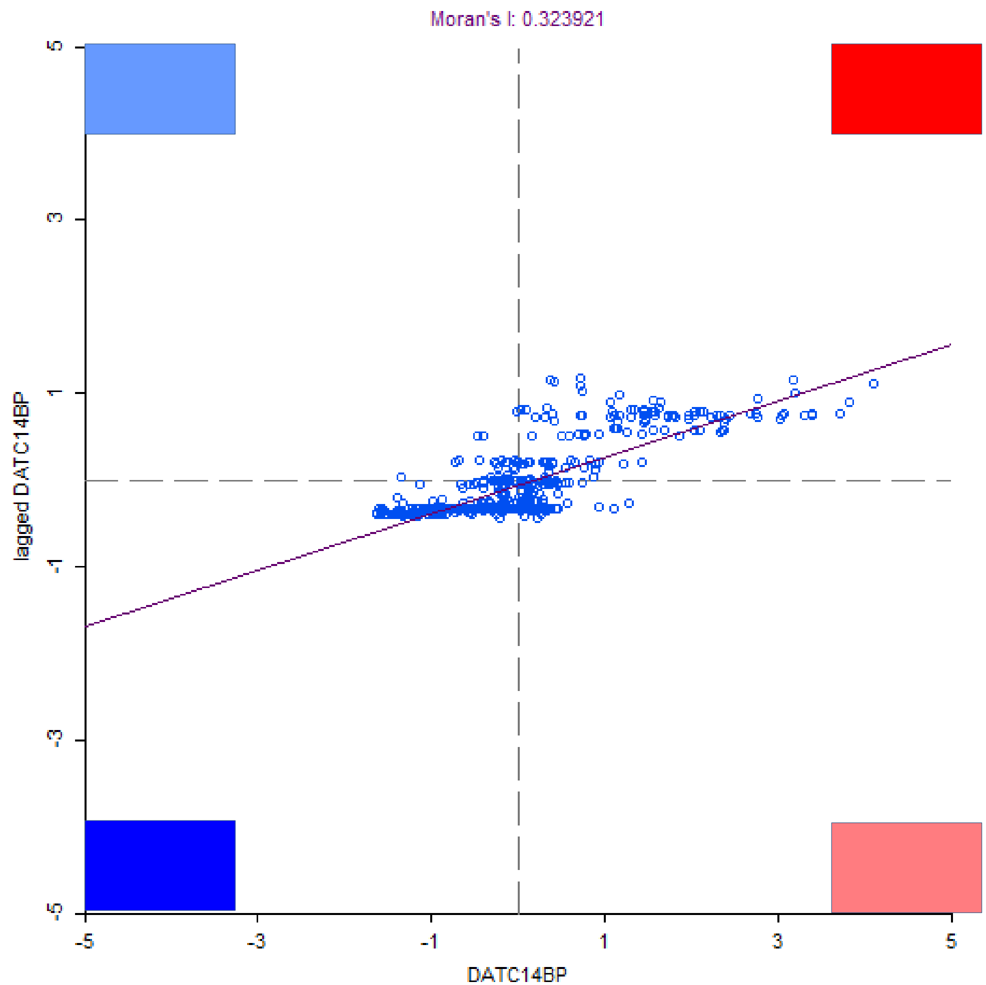


9999 permutations

LISA CLUSTERS

LOW-HIGH

HIGH-HIGH

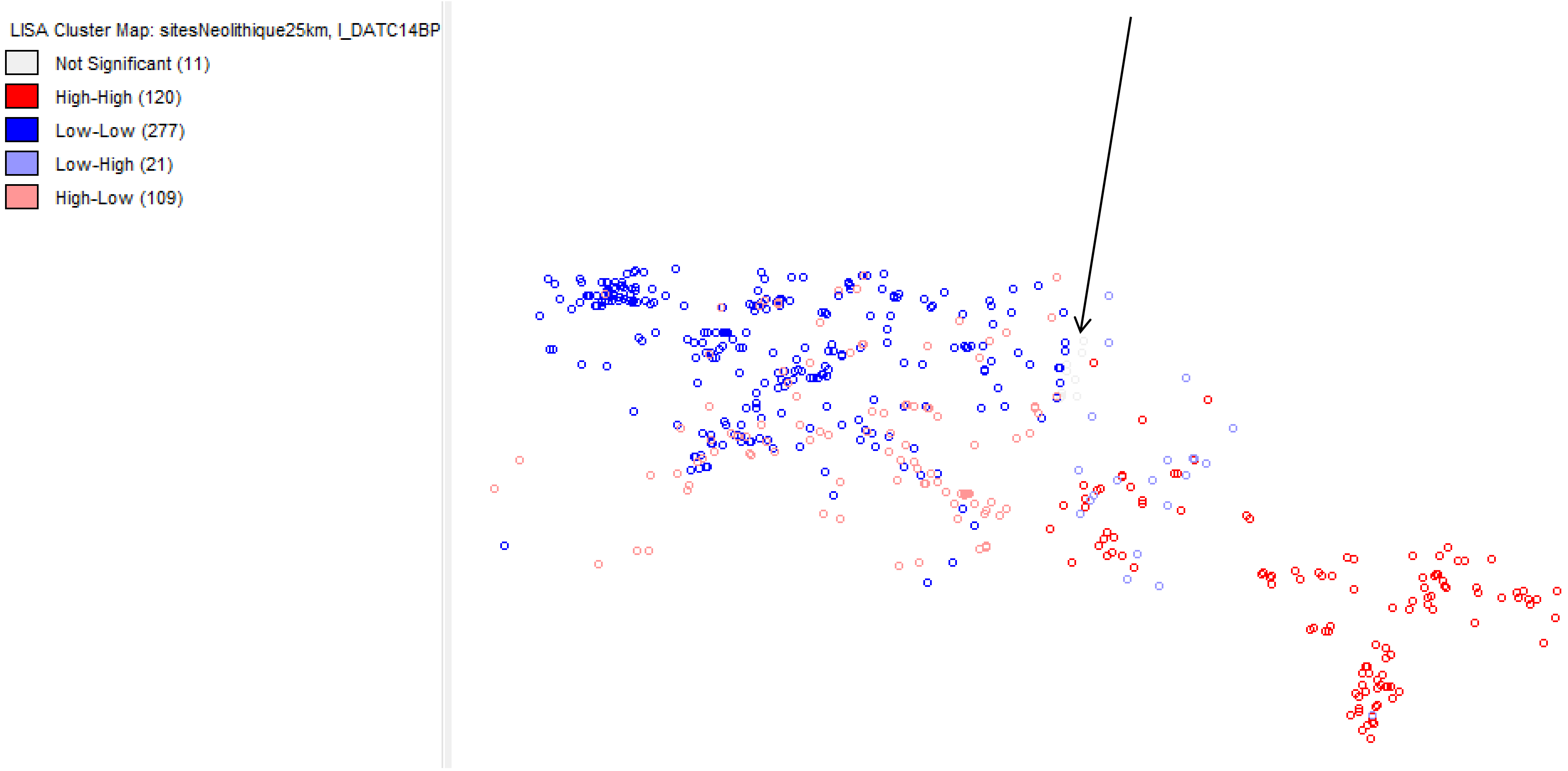


LOW-LOW

HIGH-LOW

#obs	R^2	const a	std-err a	t-stat a	p-value a	slope b	std-err b	t-stat b	p-value b
538	0.612	-0.0658	0.0111	-5.91	5.99e-009	0.324	0.0111	29.1	0

Local spatial autocorrelation on DATC14BP



IN SUMMARY

- To assess spatial dependence:
  - Calculate global or local indicators of spatial autocorrelation
  - Indicators based on the resemblance between points of interest and their neighborhood
  - Significance: establish a comparison with a **neutral geographic space** (null hypothesis = the spatial distribution of attributes is random)
- We calculate spatial dependence with theories relying on contradictory arguments (space is not neutral vs space must be neutral)
- → **Geographically Weighted Regression** (GWR)
- Includes spatial weighting in the processing of the regression

References

- Anselin, L. (1995). Local indicators of spatial association — LISA. Geographical Analysis, 27:93–115.
- Anselin, L. (1996). The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In Fischer, M., Scholten, H., and Unwin, D., editors, Spatial Analytical Perspectives on GIS in Environmental and Socio-Economic Sciences, pages 111–125. Taylor and Francis, London.
- Legendre, P. (1993) Spatial Autocorrelation: Trouble or New Paradigm? Ecology, Vol. 74, No. 6, pages 1659-1673