Systèmes d’Information Géographique

https://go.epfl.ch/sig

La dépendance spatiale

Stéphane Joost, Gabriel Kathari (GEOME-LGB)

Première loi de la géographie

A COMPUTER MOVIE SIMULATING URBAN GROWTH IN THE DETROIT REGION

W. R. TOBLER

University of Michigan

In one classification of models [16] the simulation to be described would be considered a demographic model whose primary objectives are instructional.[1] The model developed here may be used for forecasting, but was not constructed for this specific purpose, and it is a demographic model since it describes only population growth, with particular emphasis on the geographical distribution of this growth.

As a premise, I make the assumption that everything is related to everything else. Superficially considered this would suggest a model of infinite complexity; a corollary inference often made is that social systems are difficult because they contain many variables; numerous people confuse the number of variables with the degree of complexity. Because of closure, however, models with infinite numbers of variables are in fact sometimes more tractable than models with a finite but large number of variables [27]. My point here is that the utmost effort must be exercised to avoid writing a complicated model. It is very difficult to write a simple model but this, after all, is one of the objectives. If one plots a graph with increasing complexity on the abscissa, and increasing effectiveness on the other axis, it is well known that science is only asymptotic to one hundred percent effectiveness. No scientist claims otherwise. But the rate at which this effectiveness is achieved is extremely important, *ceterus paribus*. In other words, the objective is high success with a simple model. Statistical procedures which order the eigenvalues are popular for just this reason. Because a process appears complicated is also no reason to assume that it is the result of complicated rules, examples are: the game of chess, the motion of the planets before Copernicus; evolution before Darwin and the double helix, geology before Hutton, mechanics before Newton, geography before Christaller, and so on [5]. The plausibility of models also varies, but this is known to be an incomplete guide to the scientific usefulness of a model. The model I describe, for example, recognizes that people die, are born, and migrate. It does not explain why people die, are born, and migrate. Some would insist that I should incorporate more behavioral notions, but then it would be necessary to discuss the psychology of urban growth; to do this properly requires a treatise on the biochemistry of perception, which in turn requires discussion of the physics of ion interchange, and so on. My attitude, rather, is that since I have not explained birth, death, or migration, the model might apply to any phenomenon which has these characteristics, e.g., people, plants, animals, machines (which are built, moved, and destroyed), or ideas. The level of generality seems inversely related to the specificity of the model. A model of urban growth should apply to all 92,200 cities [9, p. 81] (not just to one city), now and in the future, and to other things that grow. These are rather ambitious aims. Conversely, the model attempts to relate population totals only on the basis of prior populations, and neglects employment opportunities, topography, transportation, and other distinctions between site qualities. Consequently the only difference between places in the model is their population density, and other demographic differences are ignored. Similarly, the population model attempts to relate

[1] For a review of urban models see Lee [21].

population growth only to population in the immediately preceding time period. Since, by assumption, everything is related to everything else, such a neglect of history may prove disastrous. To include all history, however, is known to require integral equations of the Volterra type [34] and these complicate the presentation.[2] We may also determine empirically whether a neglect of history has serious consequences, at least in the short run. In summary, the many simplifications of the model are acknowledged as advantages, particularly for pedagogic purposes.

Conceptually, I have been influenced by Borchert's model of the twin cities region [2]. This was later applied to Detroit by Deskins, and I have used his data [8]. As formulated by Borchert and Deskins the model is in graphical form and suggests that the lines of growth coincide with extrapolations, modified by local conditions, of the orthogonal trajectories to the level curves of population density. The difficult step is to estimate the amount of growth along these trajectories. Presumably this is proportional to the population pressure, or the gradient of the population density [23].

Following Pollack [26] specific equations may now be postulated, letting $\frac{dP}{dt}$ denote population growth at any location:

$\frac{dP}{dt} = k$, constant regional growth, or

$\frac{dP}{dt} = kP$, proportional growth, or

$\frac{dP}{dt} = k(1-\alpha)P$, logistic growth, or

$\frac{dP}{dt} = k\left[\left(\frac{\partial P}{\partial x}\right)^2 + \left(\frac{\partial P}{\partial y}\right)^2\right]^{1/2}$, growth is proportional to the population gradient, or

[2] Also see Brown [4].

$\frac{dP}{dt} = k\left(\frac{\partial^2 P}{\partial x^2} + \frac{\partial^2 P}{\partial y^2}\right)$, growth is proportional to the rate of change of the population gradient, or

$\frac{d^2P}{dt^2} = k\left(\frac{\partial^2 P}{\partial x^2} + \frac{\partial^2 P}{\partial y^2}\right)$, the acceleration of growth is proportional to the population curvature, and so on.

Each of these equations could now be examined in some detail, or converted to finite difference form for empirical estimation purposes, but I prefer to generalize in a different direction.

The simulation of urban growth raises questions of geographical syntax. As an example, recall that many predictive models are of the form

$$C = BA$$

where A is an n x 1 vector of known observations, B is an m x n transformation matrix of coefficients or transition probabilities, and C is the m x 1 vector to be predicted. This scheme seems inadequate as a geographical calculus. The geographical situation is better represented, in a simplified special case, as

$$D = NGE$$

where G and D are now m x n matrices, isomorphic to maps of the geographical landscape [32], and N and E are coefficient matrices representing North-South and East-West effects. The matrix D could of course be converted into a long column vector (mn x 1) by partitioning along the columns and the placing of these one above the other. But this destroys the isomorphism to the geographical situation. Since "the purpose of computing is insight, not numbers," I aim for a simple structure [13]. Using geographical state matrices seems more natural than using state vectors.

To some extent attempts to simulate urban growth are also related to the problem of comparing geographical maps, a question which occurs frequently in geography [30]. Let me clarify this analogy. Suppose I have a map showing

«Tout interagit avec tout dans l'espace géographique, mais deux objets proches ont plus de chances de le faire que deux objets éloignés»

W.Tobler, 1970

Dépendance spatiale

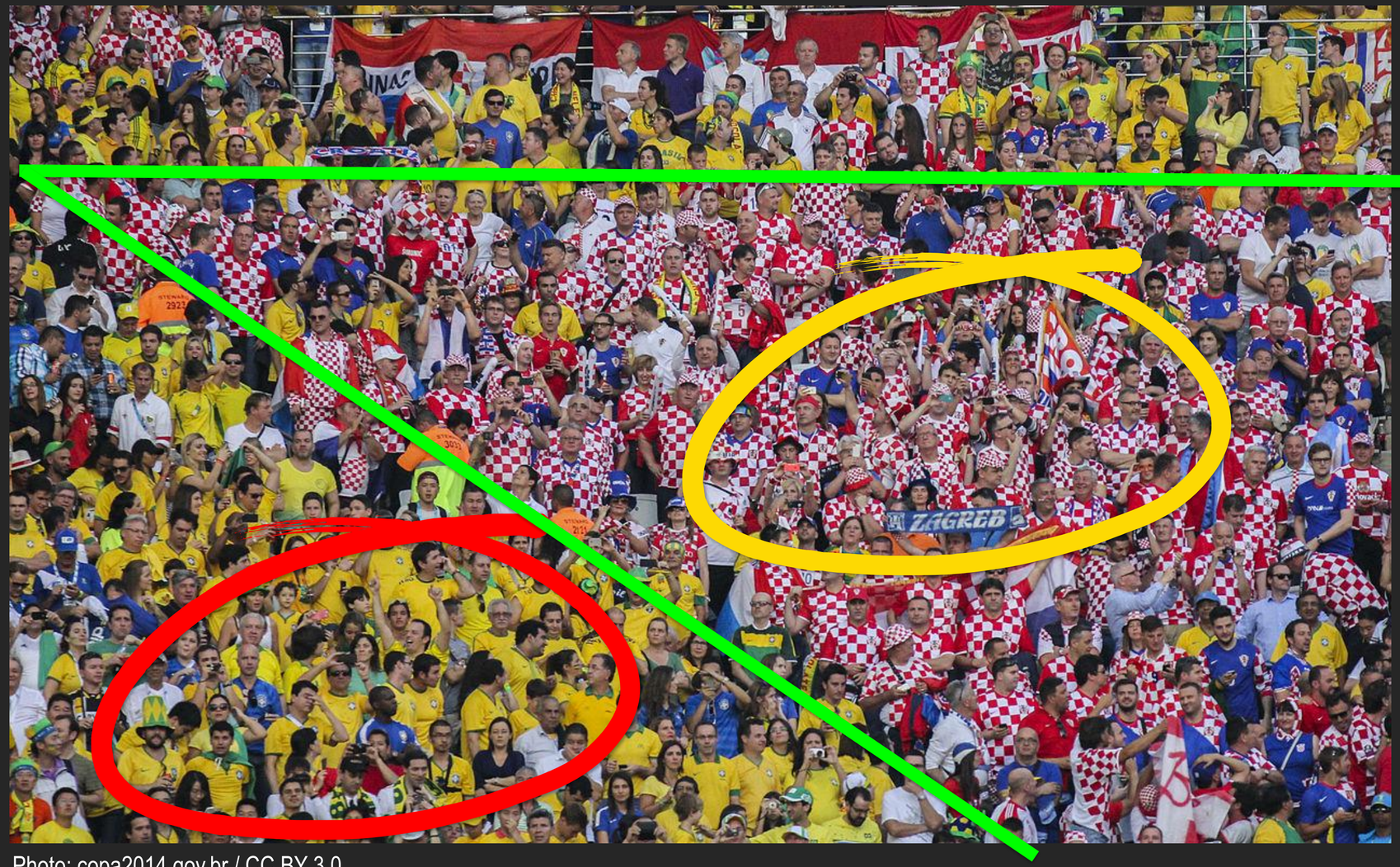


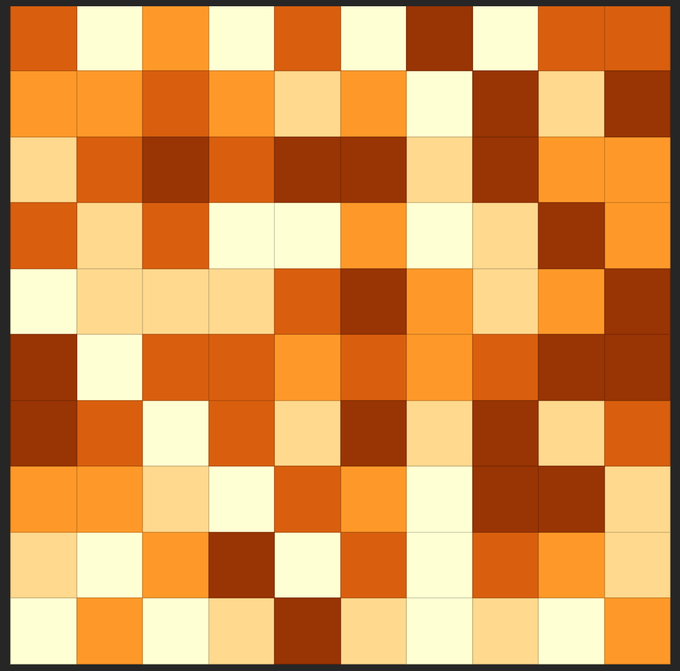
Photo: copa2014.gov.br / CC BY 3.0

Dépendance spatiale

Photo: EPFL

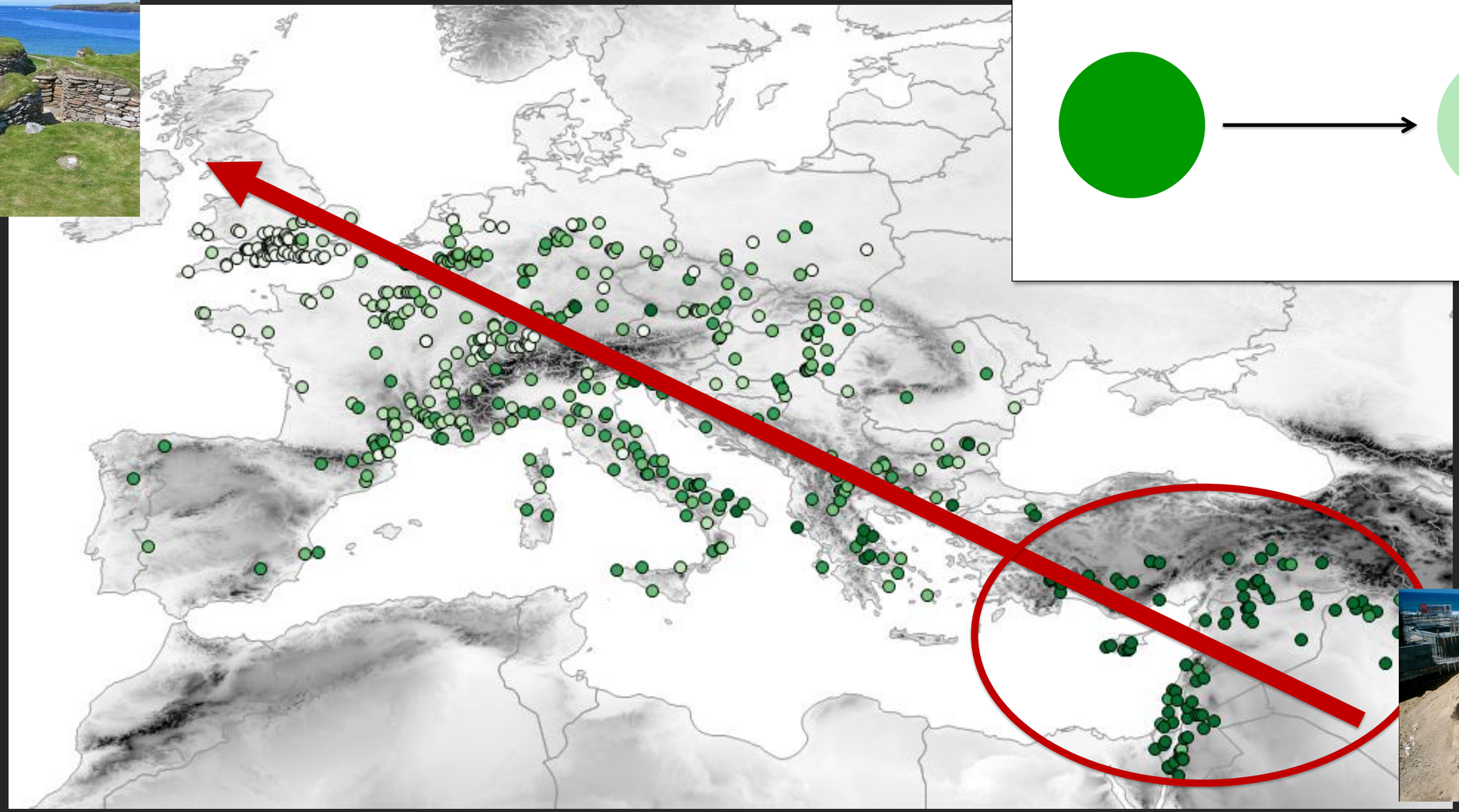
Dépendance spatiale (d’un attribut, d’un phénomène)

Dépendance spatiale



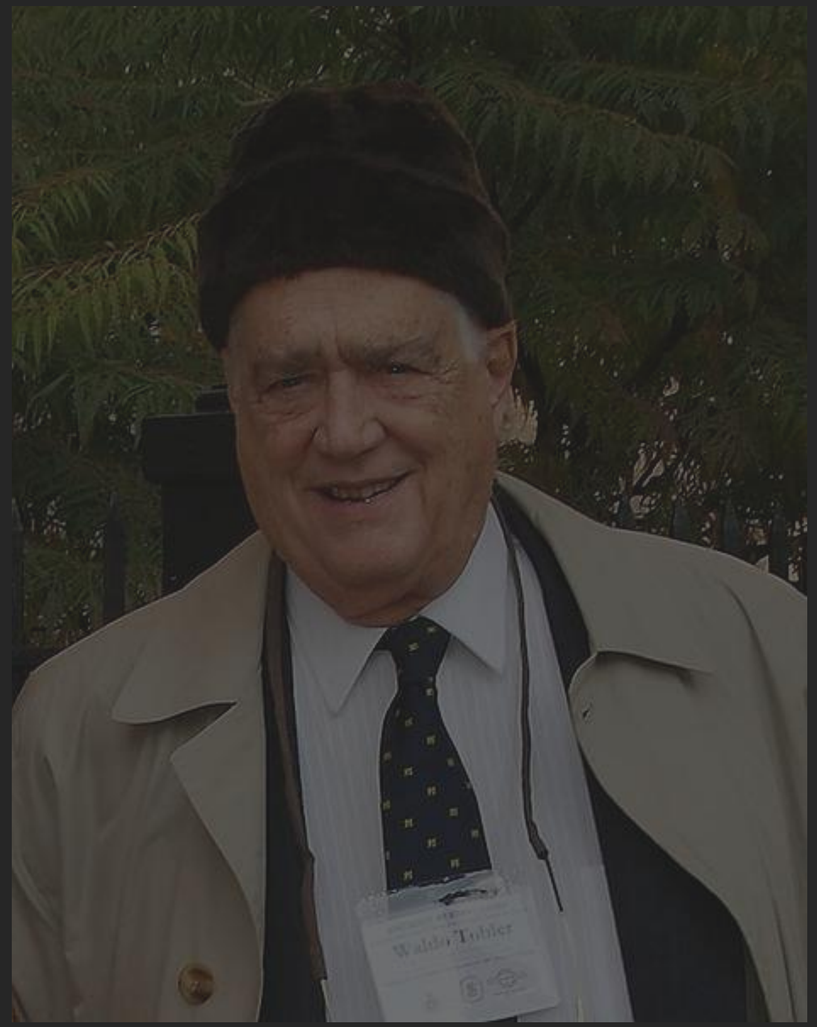
Distribution aléatoire des mêmes valeurs

Datation au carbone 14 de 765 sites néolithiques



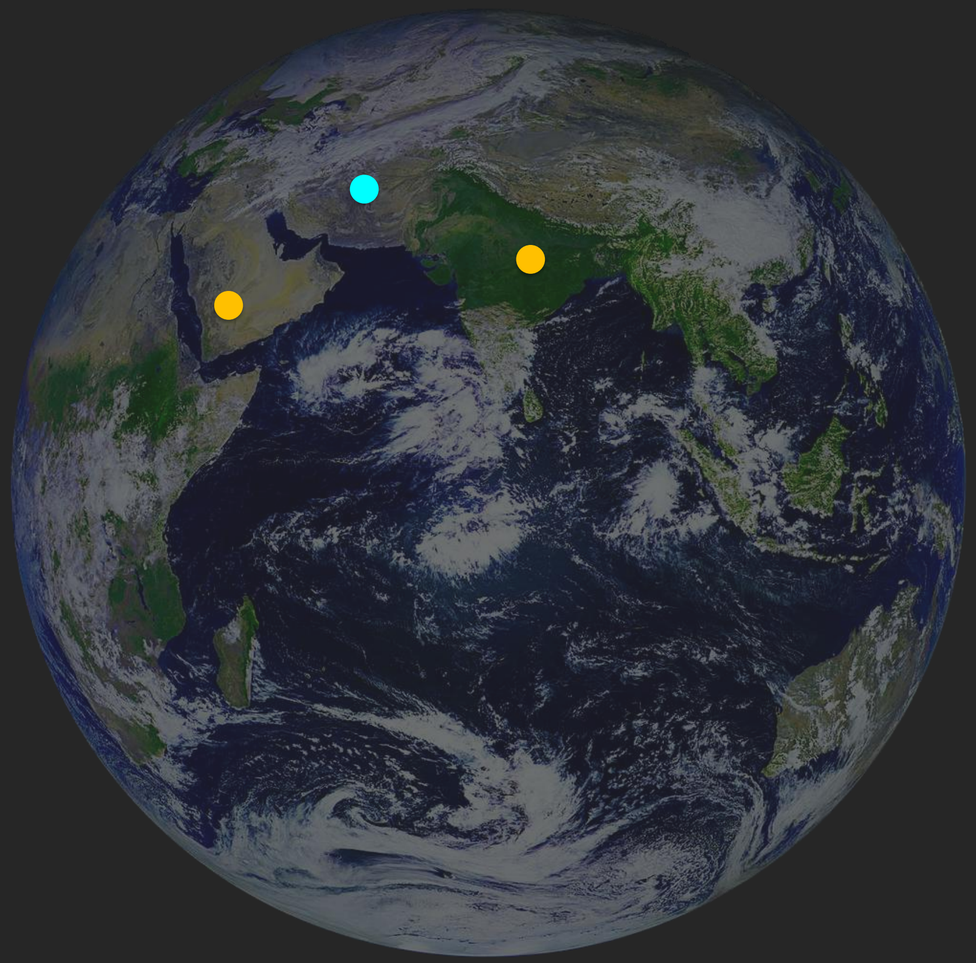
Données: Pinhasi et al. 2005, PLOS Biology | Carte: SRTM, LASIG/EPFL

Mesure de l’autocorrélation spatiale: un paradoxe



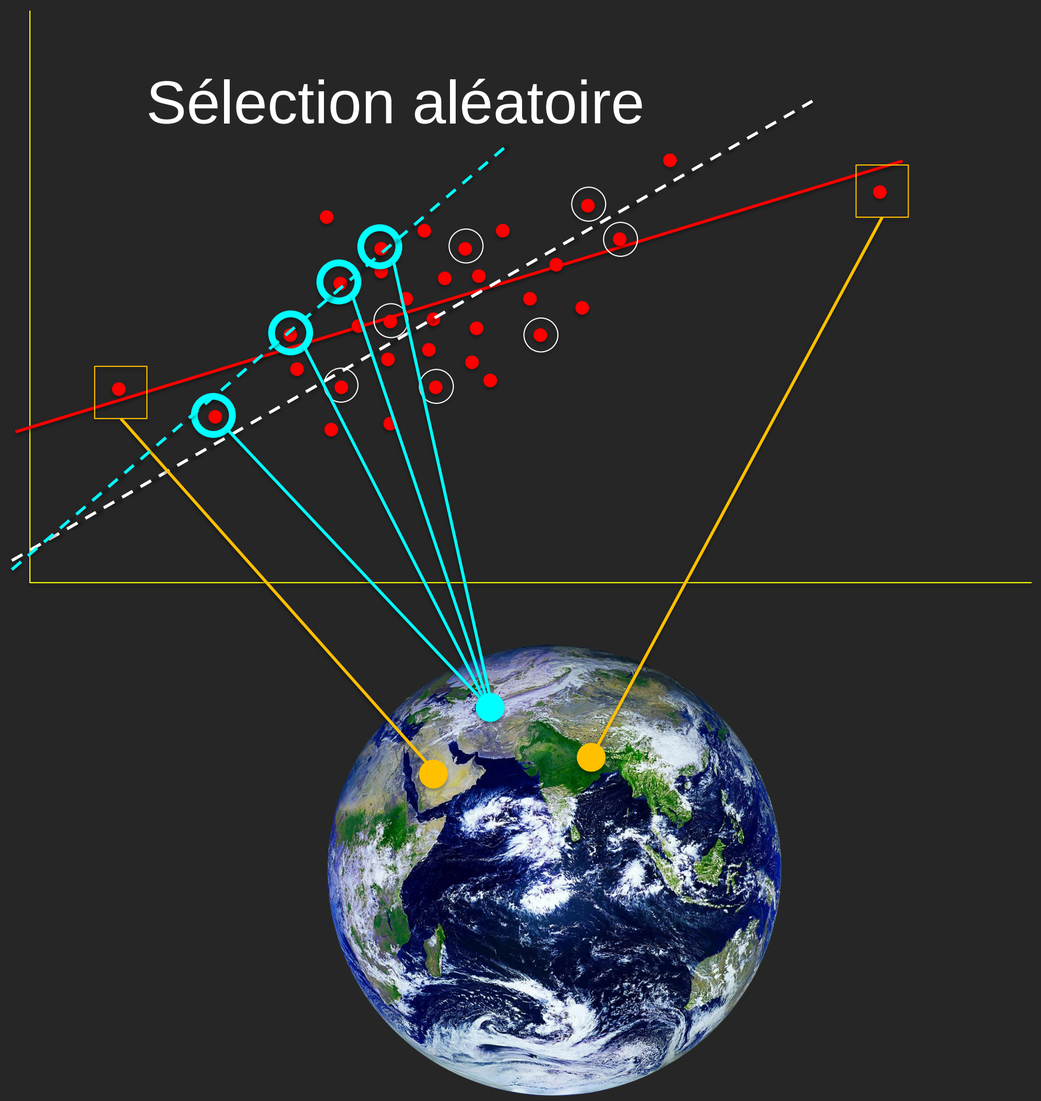
- Selon Tobler et la première loi de la géographie «Tout interagit avec tout, mais les objets proches ont plus de chances de le faire que des objets éloignés»
- Les phénomènes naturels (température de l’air), ou socio-démographiques (densité de population) ne sont pas distribués au hasard dans l’espace géographique
- Pour mesurer la structure spatiale de ces phénomènes, on doit utiliser des outils de la statistique classique qui requièrent l’indépendance entre les échantillons et une distribution aléatoire de ces échantillons

Les outils de la statistique classique…



- Ne sont pas prévus pour être appliqués dans un contexte géospatial
- Leur utilisation est basée sur l'hypothèse selon laquelle l'espace géographique est neutre
- Cet espace géographique constitue le simple support sans friction sur lequel se déroulent les phénomènes étudiés
- Théoriquement, dans ce cadre, la localisation d'observations dans l'espace ne doit pas influencer leurs attributs
- Ce qui n'est pas le cas, donc biais possibles

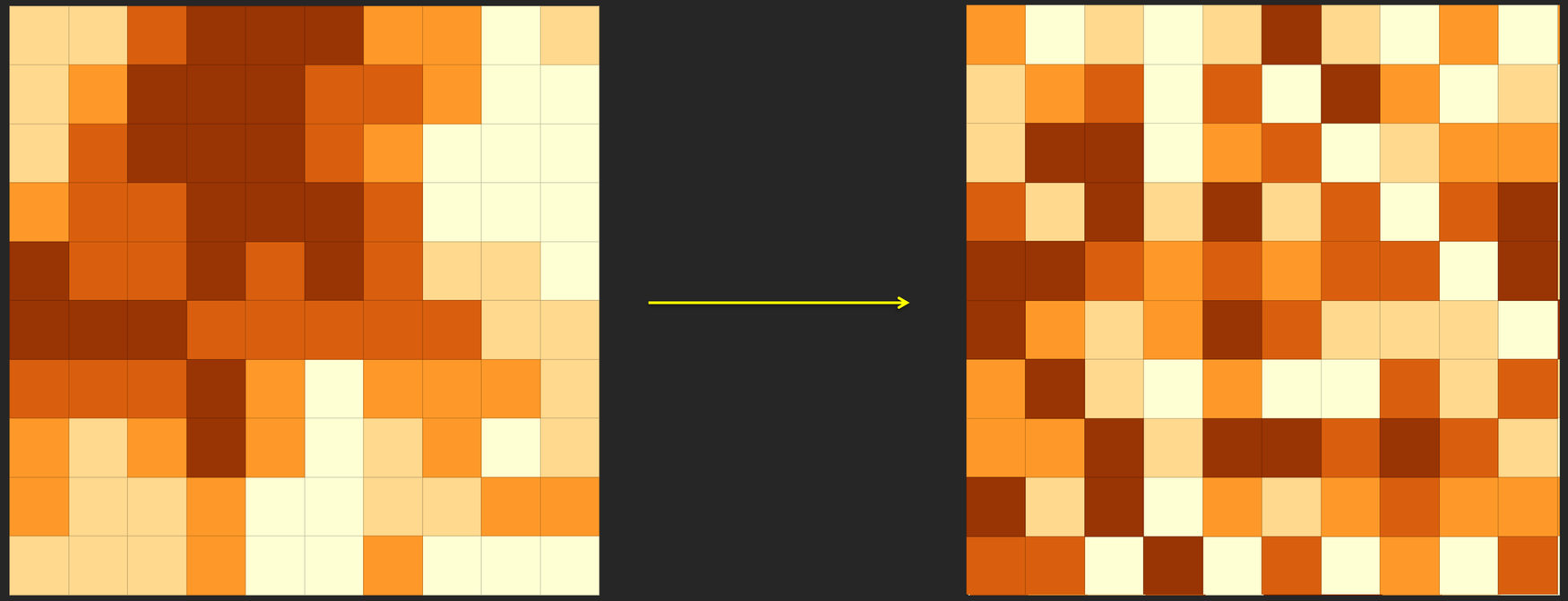
Un exemple de biais: la régression linéaire



- Régression linéaire : devrait être calculée avec des observations sélectionnées selon une procédure aléatoire
- Si les observations sont spatialement dépendantes, les valeurs estimées sont biaisées pour toute la zone d'étude
- Des valeurs exceptionnelles localisées dans des sous-regions particulières influencent les valeurs prédites sur tout le territoire analysé
- Une forte corrélation entre deux attributs d'échantillons situés dans une petite sous-région aura un effet sur toute la zone étudiée

- Des approaches spécifiques ont été développées pour prendre en compte les caractéristiques de l'information géographique (la 1ère loi de la Géographie de Tobler)
- On parle de méthodes spatialement explicites (modèles non-stationnaires vs modèles stationnaires)
- Elles respectent les lois de la statistique théorique

Simuler un espace géographique neutre



- Distinguer la distribution spatiale observée du hasard
- = différencier la distribution spatiale observée d'une distribution aléatoire
- Simuler le hasard par un grand nombre de permutations aléatoires

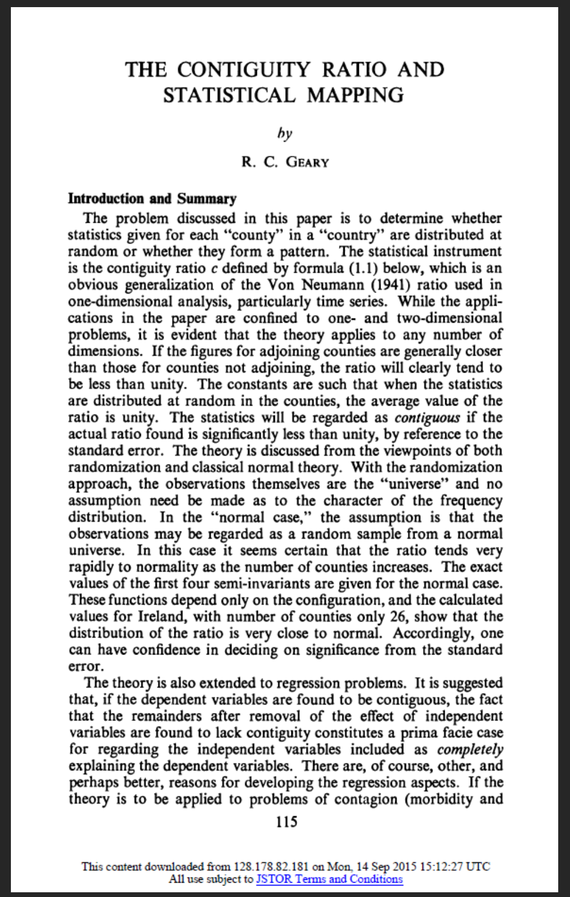
Systèmes d’Information Géographique

https://go.epfl.ch/sig

Autocorrélation spatiale globale

Stéphane Joost, Gabriel Kathari (GEOME-LGB)

Mesure globale de l'autocorrélation spatiale



THE CONTIGUITY RATIO AND STATISTICAL MAPPING

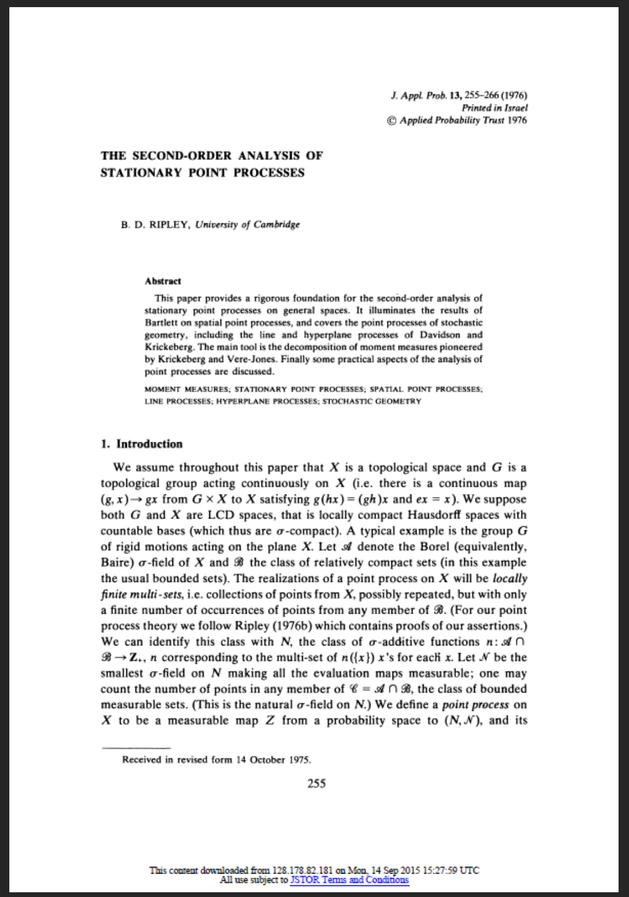
by

R. C. GEARY

Introduction and Summary

The problem discussed in this paper is to determine whether statistics given for each "county" in a "country" are distributed at random or whether they form a pattern. The statistical instrument is the contiguity ratio c defined by formula (1.1) below, which is an obvious generalization of the Von Neumann (1941) ratio used in one-dimensional analysis, particularly time series. While the applications in the paper are confined to one- and two-dimensional problems, it is evident that the theory applies to any number of dimensions. If the figures for adjoining counties are generally closer than those for counties not adjoining, the ratio will clearly tend to be less than unity. The constants are such that when the statistics are distributed at random in the counties, the average value of the ratio is unity. The statistics will be regarded as *contiguous* if the actual ratio found is significantly less than unity, by reference to the standard error. The theory is discussed from the viewpoints of both randomization and classical normal theory. With the randomization approach, the observations themselves are the "universe" and no assumption need be made as to the character of the frequency distribution. In the "normal case," the assumption is that the observations may be regarded as a random sample from a normal universe. In this case it seems certain that the ratio tends very rapidly to normality as the number of counties increases. The exact values of the first four semi-invariants are given for the normal case. These functions depend only on the configuration, and the calculated values for Ireland, with number of counties only 26, show that the distribution of the ratio is very close to normal. Accordingly, one can have confidence in deciding on significance from the standard error.

The theory is also extended to regression problems. It is suggested that, if the dependent variables are found to be contiguous, the fact that the remainders after removal of the effect of independent variables are found to lack contiguity constitutes a prima facie case for regarding the independent variables included as *completely* explaining the dependent variables. There are, of course, other, and perhaps better, reasons for developing the regression aspects. If the theory is to be applied to problems of contagion (morbidity and



J. Appl. Prob. **13**, 255–266 (1976)
Printed in Israel

THE SECOND-ORDER ANALYSIS OF STATIONARY POINT PROCESSES

B. D. RIPLEY, *University of Cambridge*

Abstract

This paper provides a rigorous foundation for the second-order analysis of stationary point processes on general spaces. It illuminates the results of Bartlett on spatial point processes, and covers the point processes of stochastic geometry, including the line and hyperplane processes of Davidson and Krickeberg. The main tool is the decomposition of moment measures pioneered by Krickeberg and Vere-Jones. Finally some practical aspects of the analysis of point processes are discussed.

MOMENT MEASURES; STATIONARY POINT PROCESSES; SPATIAL POINT PROCESSES; LINE PROCESSES; HYPERPLANE PROCESSES; STOCHASTIC GEOMETRY

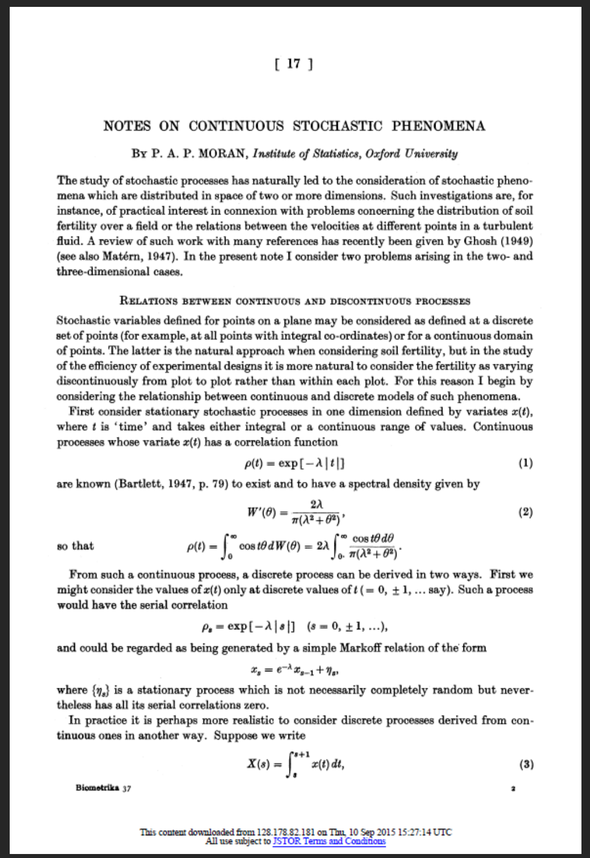
1. Introduction

We assume throughout this paper that X is a topological space and G is a topological group acting continuously on X (i.e. there is a continuous map $(g, x) \to gx$ from $G \times X$ to X satisfying $g(hx) = (gh)x$ and $ex = x$). We suppose both G and X are LCD spaces, that is locally compact Hausdorff spaces with countable bases (which thus are σ-compact). A typical example is the group G of rigid motions acting on the plane X. Let $\mathscr{A}$ denote the Borel (equivalently, Baire) σ-field of X and $\mathscr{B}$ the class of relatively compact sets (in this example the usual bounded sets). The realizations of a point process on X will be *locally finite multi-sets*, i.e. collections of points from X, possibly repeated, but with only a finite number of occurrences of points from any member of $\mathscr{B}$. (For our point process theory we follow Ripley (1976b) which contains proofs of our assertions.) We can identify this class with N, the class of σ-additive functions $n: \mathscr{A} \cap \mathscr{B} \to \mathbf{Z}_+$, n corresponding to the multi-set of $n(\{x\})$ x's for each x. Let $\mathscr{N}$ be the smallest σ-field on N making all the evaluation maps measurable; one may count the number of points in any member of $\mathscr{C} = \mathscr{A} \cap \mathscr{B}$, the class of bounded measurable sets. (This is the natural σ-field on N.) We define a *point process* on X to be a measurable map Z from a probability space to $(N, \mathscr{N})$, and its

Received in revised form 14 October 1975.

- Un indice unique quantifie l'autocorrelation sur tout l'espace géographique étudié
- Statistique de dénombrement (Join Count Statistics)
- Le C de Geary
- Le K de Ripley

Mesure globale de l'autocorrélation spatiale



[17]

NOTES ON CONTINUOUS STOCHASTIC PHENOMENA

By P. A. P. MORAN, *Institute of Statistics, Oxford University*

The study of stochastic processes has naturally led to the consideration of stochastic phenomena which are distributed in space of two or more dimensions. Such investigations are, for instance, of practical interest in connexion with problems concerning the distribution of soil fertility over a field or the relations between the velocities at different points in a turbulent fluid. A review of such work with many references has recently been given by Ghosh (1949) (see also Matérn, 1947). In the present note I consider two problems arising in the two- and three-dimensional cases.

RELATIONS BETWEEN CONTINUOUS AND DISCONTINUOUS PROCESSES

Stochastic variables defined for points on a plane may be considered as defined at a discrete set of points (for example, at all points with integral co-ordinates) or for a continuous domain of points. The latter is the natural approach when considering soil fertility, but in the study of the efficiency of experimental designs it is more natural to consider the fertility as varying discontinuously from plot to plot rather than within each plot. For this reason I begin by considering the relationship between continuous and discrete models of such phenomena.

First consider stationary stochastic processes in one dimension defined by variates $x(t)$, where t is 'time' and takes either integral or a continuous range of values. Continuous processes whose variate $x(t)$ has a correlation function

$$\rho(t) = \exp[-\lambda|t|] \tag{1}$$

are known (Bartlett, 1947, p. 79) to exist and to have a spectral density given by

$$W'(\theta) = \frac{2\lambda}{\pi(\lambda^2+\theta^2)}, \tag{2}$$

so that $$\rho(t) = \int_0^\infty \cos t\theta \, dW(\theta) = 2\lambda \int_0^\infty \frac{\cos t\theta \, d\theta}{\pi(\lambda^2+\theta^2)}.$$

From such a continuous process, a discrete process can be derived in two ways. First we might consider the values of $x(t)$ only at discrete values of t ($= 0, \pm 1, \ldots$ say). Such a process would have the serial correlation

$$\rho_s = \exp[-\lambda|s|] \quad (s = 0, \pm 1, \ldots),$$

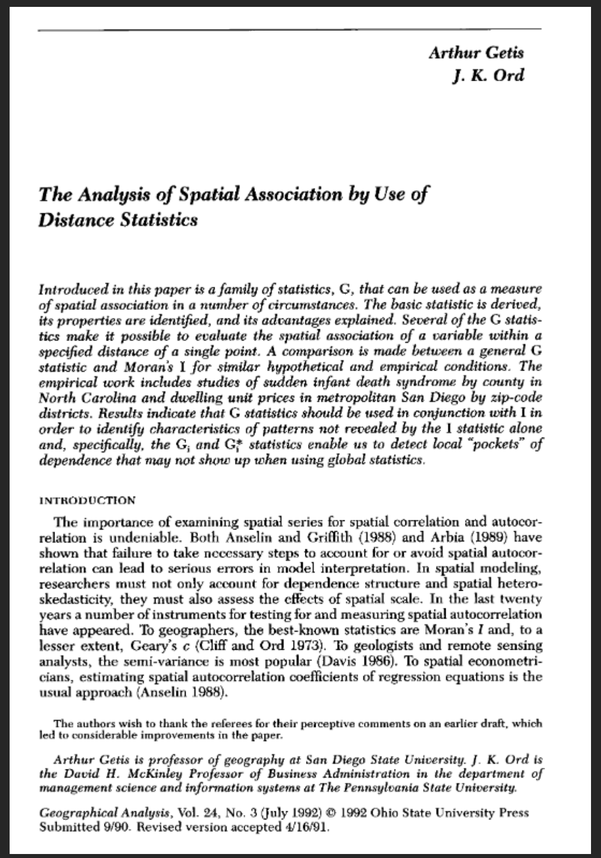
and could be regarded as being generated by a simple Markoff relation of the form

$$x_s = e^{-\lambda} x_{s-1} + \eta_s,$$

where $\{\eta_s\}$ is a stationary process which is not necessarily completely random but nevertheless has all its serial correlations zero.

In practice it is perhaps more realistic to consider discrete processes derived from continuous ones in another way. Suppose we write

$$X(s) = \int_s^{s+1} x(t)\, dt, \tag{3}$$



Arthur Getis
J. K. Ord

The Analysis of Spatial Association by Use of Distance Statistics

Introduced in this paper is a family of statistics, G, that can be used as a measure of spatial association in a number of circumstances. The basic statistic is derived, its properties are identified, and its advantages explained. Several of the G statistics make it possible to evaluate the spatial association of a variable within a specified distance of a single point. A comparison is made between a general G statistic and Moran's I for similar hypothetical and empirical conditions. The empirical work includes studies of sudden infant death syndrome by county in North Carolina and dwelling unit prices in metropolitan San Diego by zip-code districts. Results indicate that G statistics should be used in conjunction with I in order to identify characteristics of patterns not revealed by the I statistic alone and, specifically, the G_i and G_i^* statistics enable us to detect local "pockets" of dependence that may not show up when using global statistics.

INTRODUCTION

The importance of examining spatial series for spatial correlation and autocorrelation is undeniable. Both Anselin and Griffith (1988) and Arbia (1989) have shown that failure to take necessary steps to account for or avoid spatial autocorrelation can lead to serious errors in model interpretation. In spatial modeling, researchers must not only account for dependence structure and spatial heteroskedasticity, they must also assess the effects of spatial scale. In the last twenty years a number of instruments for testing for and measuring spatial autocorrelation have appeared. To geographers, the best-known statistics are Moran's *I* and, to a lesser extent, Geary's *c* (Cliff and Ord 1973). To geologists and remote sensing analysts, the semi-variance is most popular (Davis 1986). To spatial econometricians, estimating spatial autocorrelation coefficients of regression equations is the usual approach (Anselin 1988).

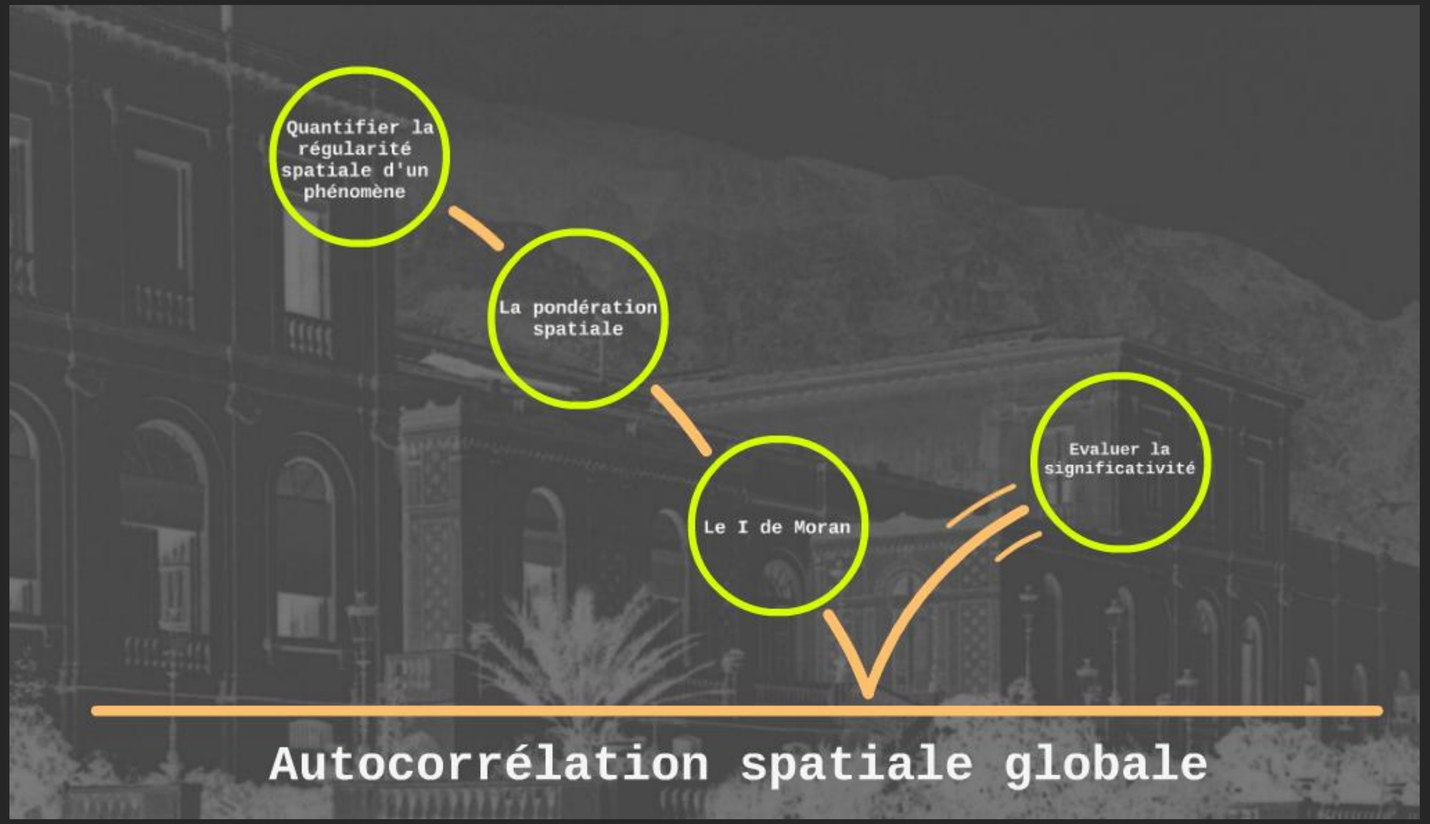
The authors wish to thank the referees for their perceptive comments on an earlier draft, which led to considerable improvements in the paper.

Arthur Getis is professor of geography at San Diego State University. J. K. Ord is the David H. McKinley Professor of Business Administration in the department of management science and information systems at The Pennsylvania State University.

Geographical Analysis, Vol. 24, No. 3 (July 1992) © 1992 Ohio State University Press
Submitted 9/90. Revised version accepted 4/16/91.

- Un indice unique quantifie l'autocorrelation sur tout l'espace géographique étudié
- Le G de Getis-Ord
- Le I de Moran

Autocorrélation spatiale globale



- Les relations de voisinage
- La pondération spatiale
- Le I de Moran
- Significativité du I de Moran

Principe

- But: quantifier la ressemblance entre n objets dispersés sur un territoire donné pour un attribut donné.
- Par exemple: la quantité de précipitations qui tombe dans 54 communes du Gros-de-Vaud dépend-elle de la localisation de ces communes ?

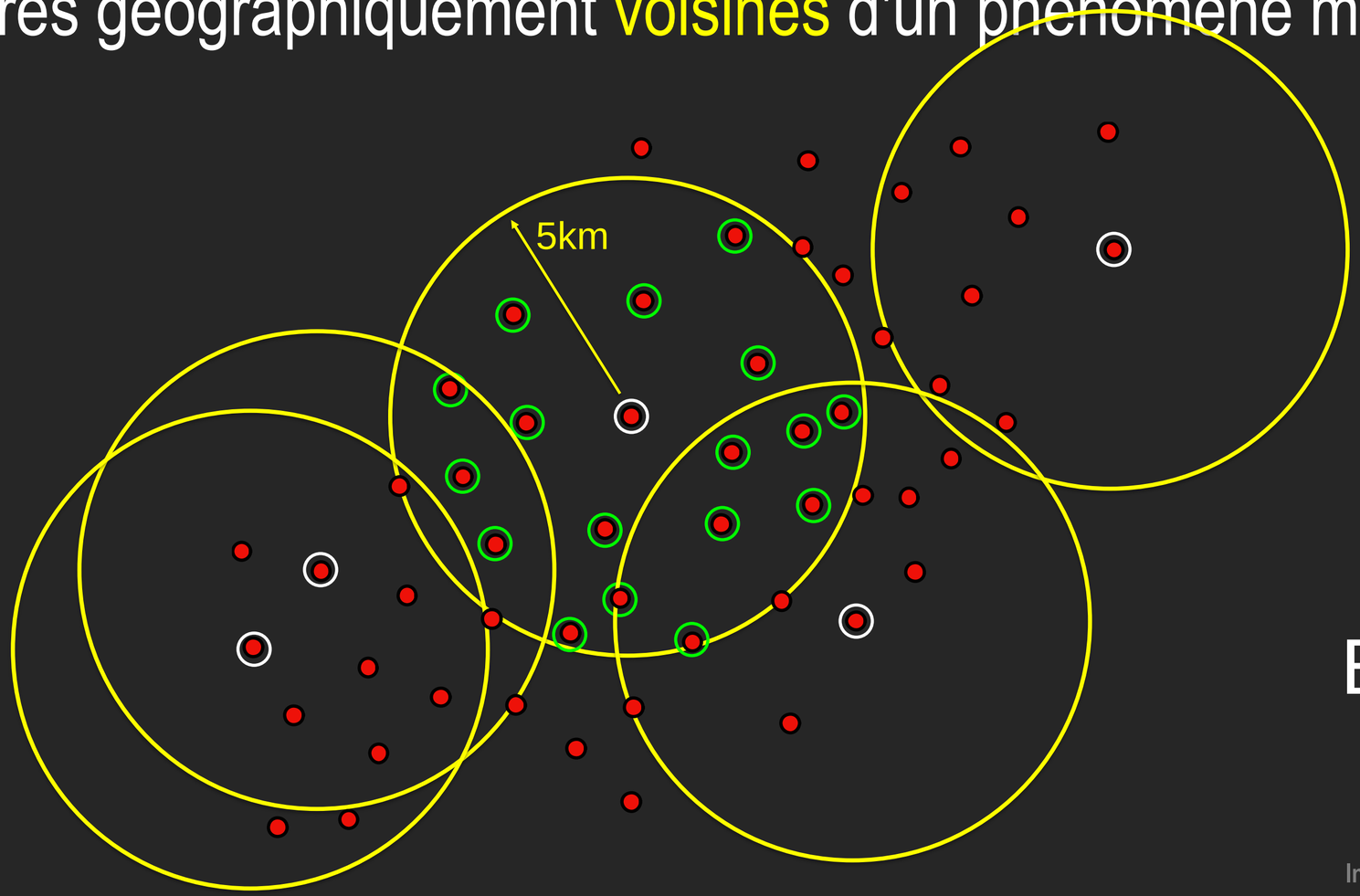
Si oui → haut coefficient de ressemblance,
Si pas de relation = 0,
Si relation inverse (les valeurs voisines ne se ressemblent pas), coefficient = -1

Principe - approche

1. Définir un voisinage de référence autour de chaque unité géographique
2. Calculer la valeur pondérée de chaque unité (la valeur moyenne de l'attribut z dans le voisinage)
3. Mesurer la ressemblance entre les unités géographiques et leur voisinage
4. Comparer la situation observée à des situations aléatoires (permutations aléatoires) pour tester la significativité de la mesure de ressemblance

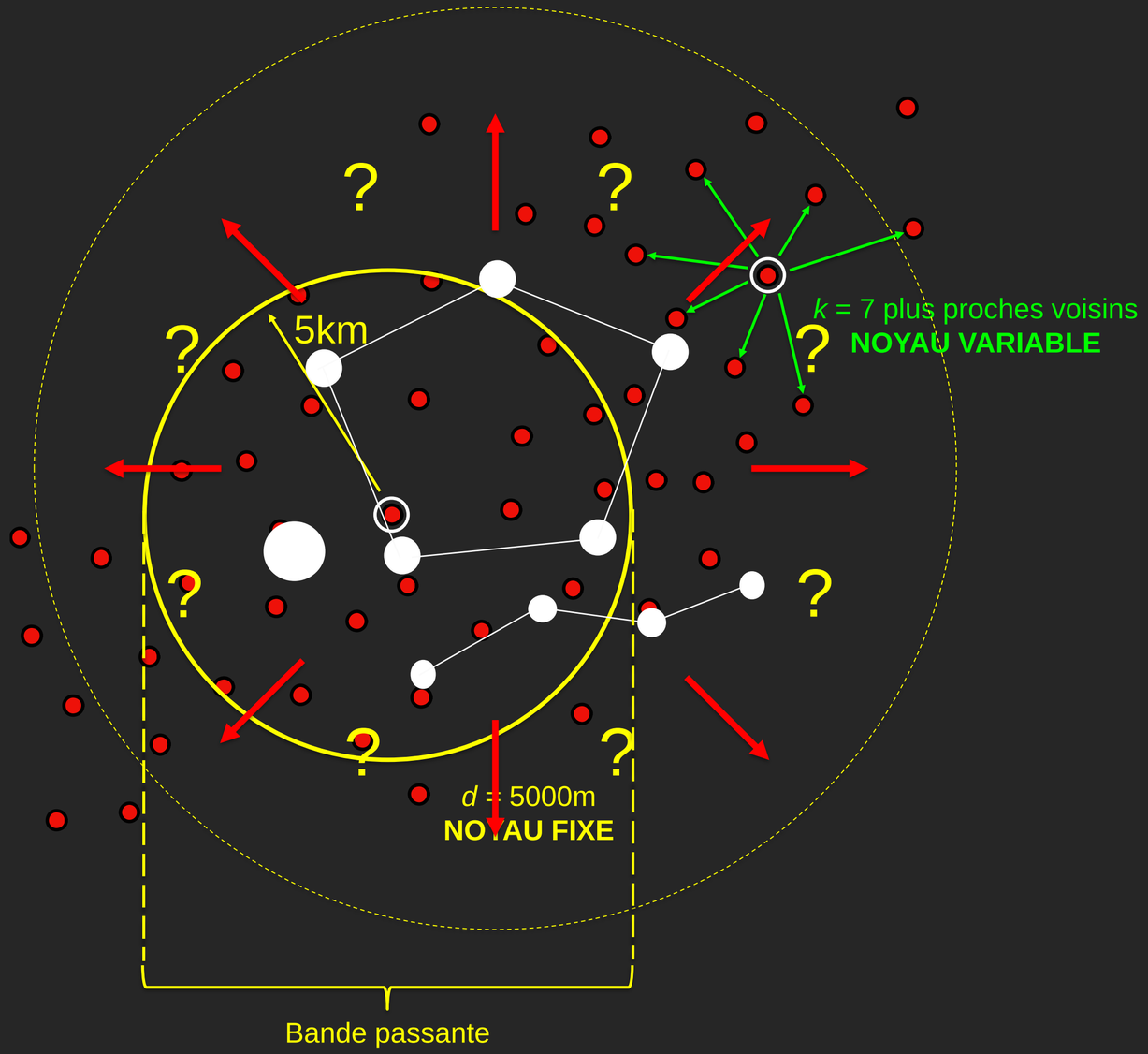
Les relations de voisinage

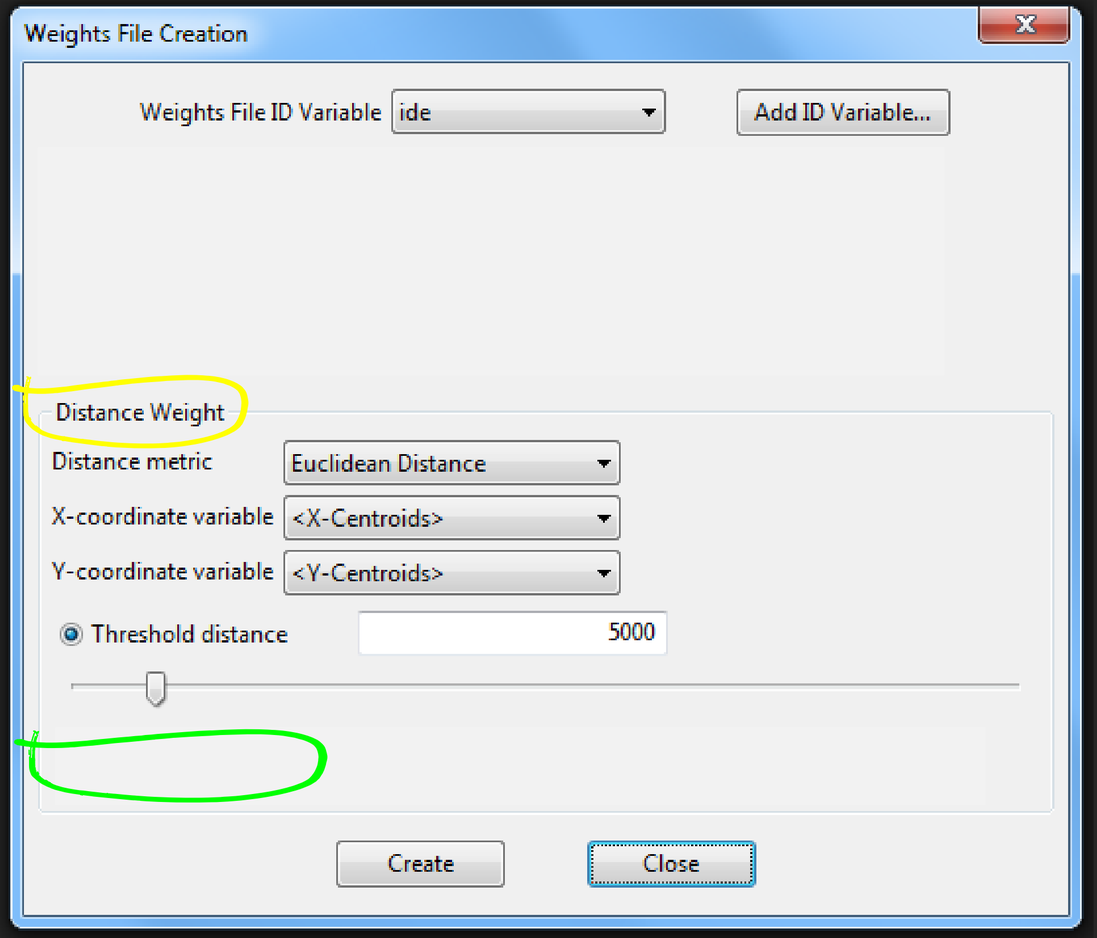
- L'autocorrélation spatiale est caractérisée par une corrélation entre les mesures géographiquement voisines d'un phénomène mesuré



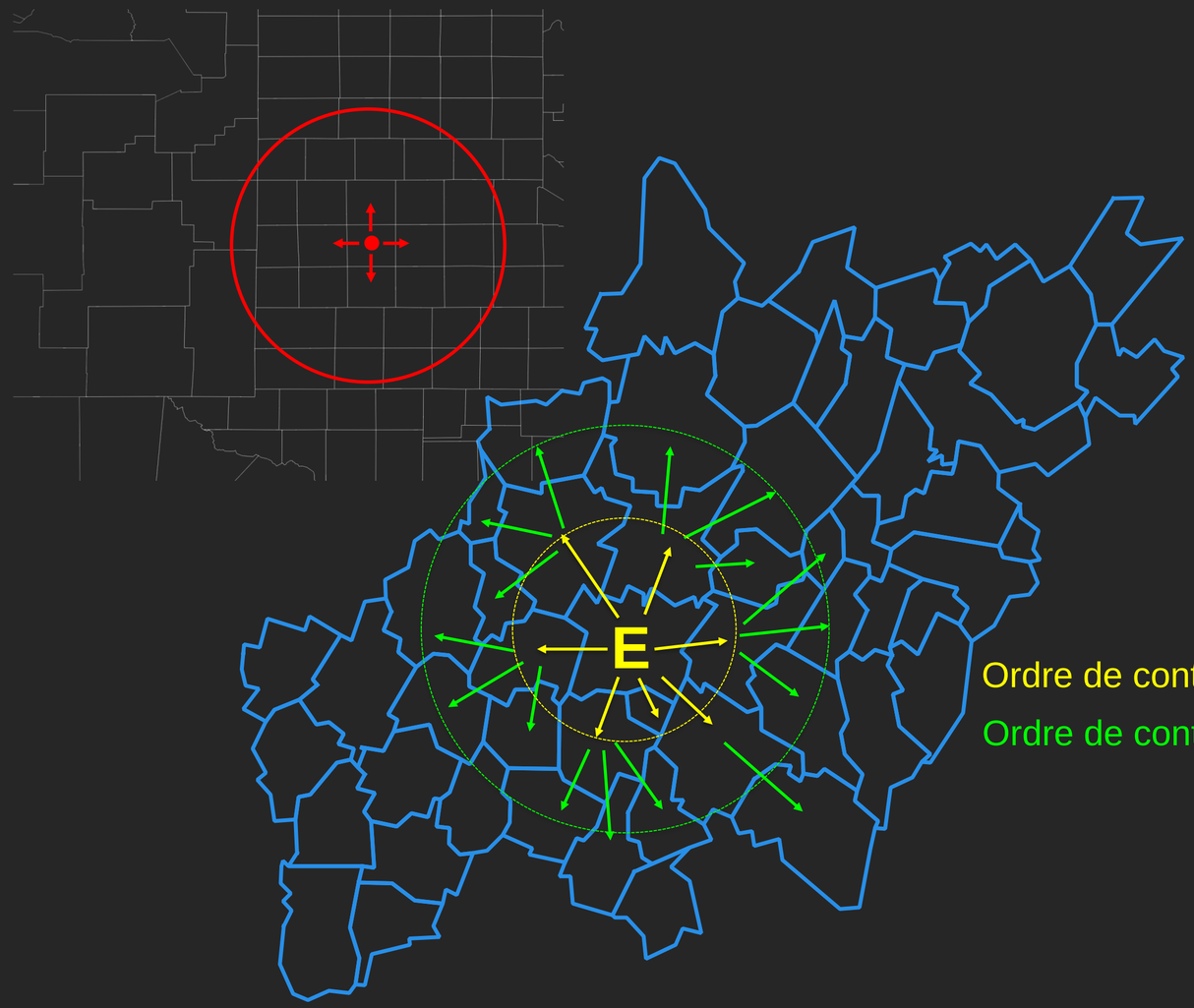
Etc.

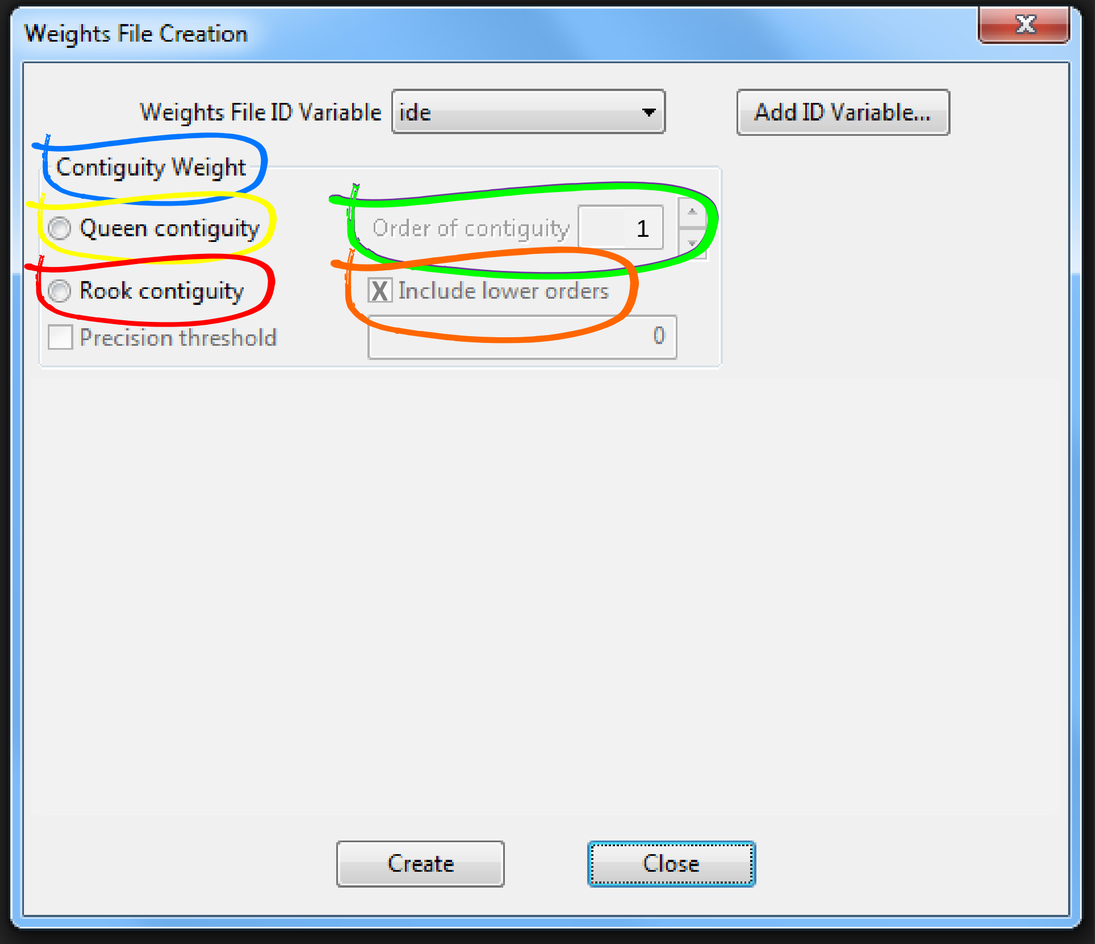
Critères de définition du voisinage - objets ponctuels





Critères de définition du voisinage – polygones





Pondération spatiale

Schéma de pondération spatiale

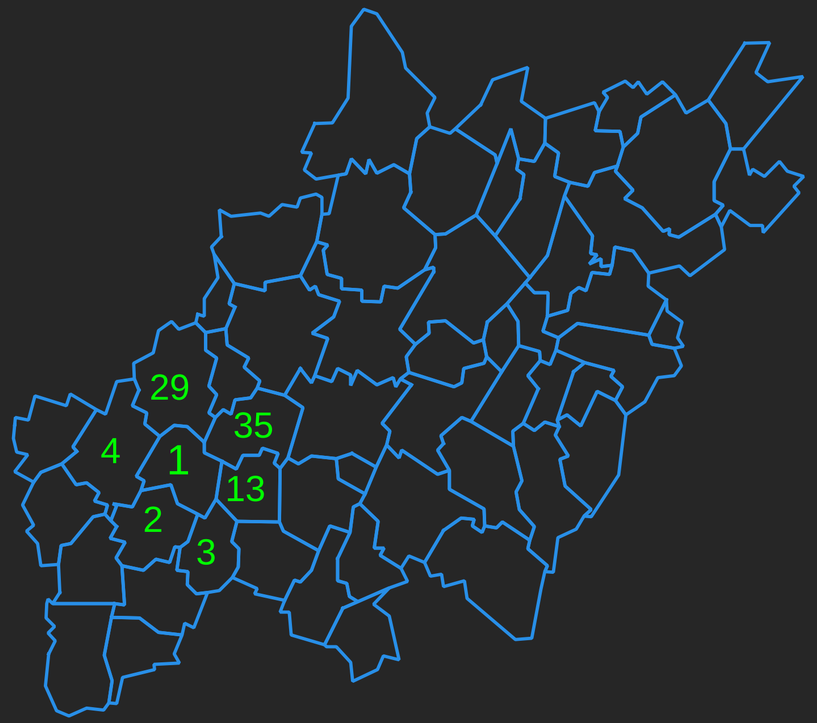
NOYAU FIXE → Bande passante *d*

NOYAU VARIABLE → *k* plus proches voisins

→ Ordre de contiguïté *n*

→ Fichier de pondération spatiale

```
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
...
```



comvd_prec_qu1.gal

```
1   0 54 comvd_prec ide
2   1 6
3   35 29 13 4 3 2
4   2 5
5   9 8 4 3 1
6   3 5
7   13 9 21 1 2
8   4 6
9   8 7 5 1 2 29
10  5 2
11  7 4
12  6 2
13  10 9
14  7 3
15  8 4 5
16  8 5
17  10 9 2 4 7
18  9 5
19  10 3 2 6 8
20  10 3
21  6 8 9
22  11 8
23  35 27 26 21 15 13 14 18
```

n = 1

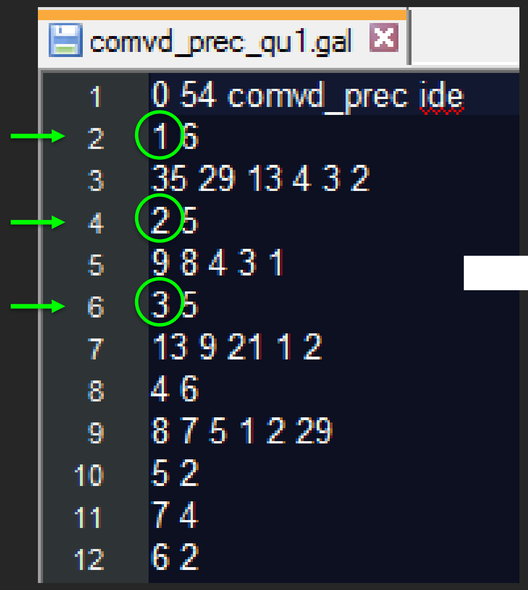
comvd_prec_7k.gwt

1	0 54 comvd_prec ide	
2	1 2	1895.70644
3	1 13	2031.24132
4	1 4	2062.65071
5	1 35	2365.33363
6	1 3	2474.90419
7	1 29	2533.84993
8	1 19	3041.12639
9	2 3	1813.98042
10	2 1	1895.70644
11	2 9	1992.83081
12	2 8	2039.59784
13	2 4	2479.38678
14	2 7	2693.1369
15	2 13	3076.66614

k = 7

Corrélation entre mesures spatialement voisines…

Contiguïté ordre 1



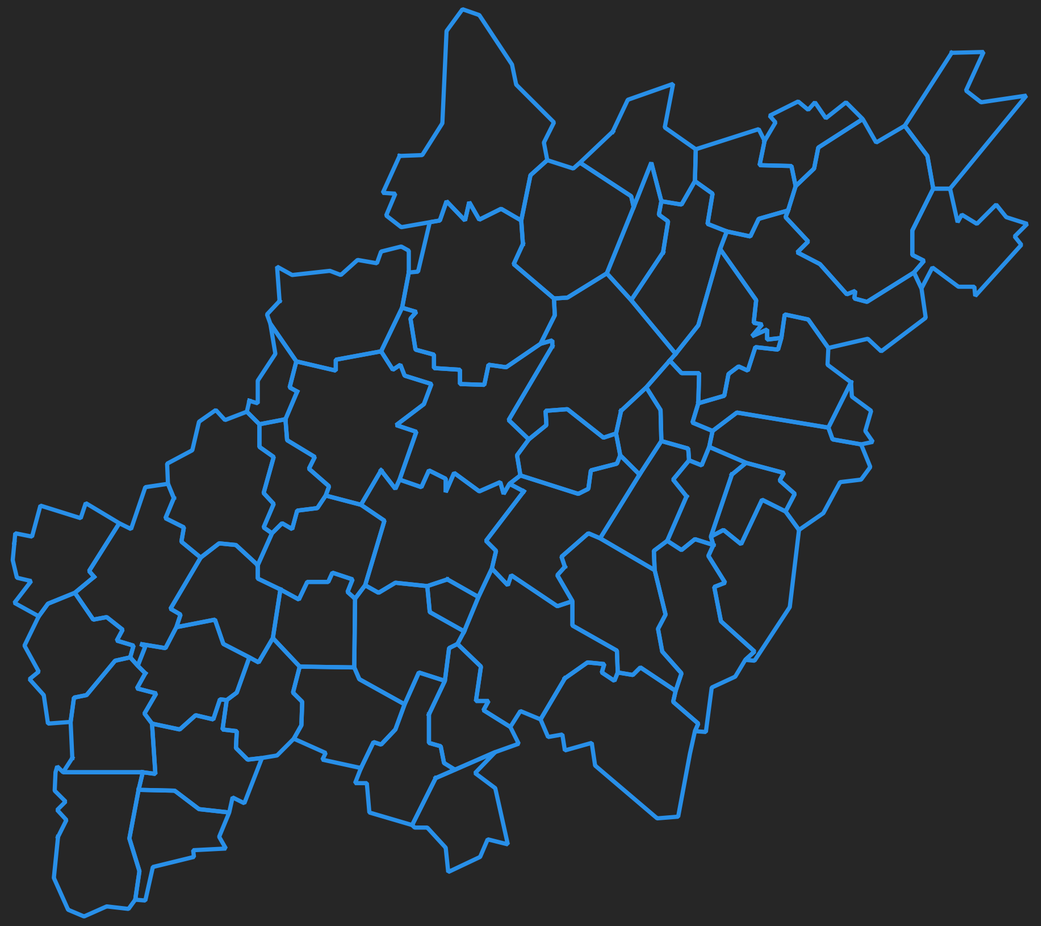
ide	voisins ($\rightarrow\omega$)	C	z	$\bar{z}$
1	35, 29, 13, 4, 3, 2	6	10291.14	10244.17
2	9, 8, 4, 3 , 1	5	10166.64	10085.33
3	13, 9, 21, 1, 2	5	10494.71	10334.34
…	…	…	…	…

$$I_m = \frac{Covariance}{Variance} = \frac{1/C \sum \omega_{i,j}(z_i - \bar{z})(z_j - \bar{z})}{1/n \sum (z_i - \bar{z})^2} = \frac{n \sum (z_i - \bar{z})(z_j - \bar{z})}{C \sum (z_i - \bar{z})^2}$$

Où n: nombre d'unités spatiales; C: nombre de voisins ou de connexions; z_i: valeur de la variable pour l'unité i; z_j: valeur de la variable pour l'unité i; ω_{ij}: poids de la connexion, 1 si adjacent, 0 autrement

La valeur de I varie entre +1 (corrélation positive totale) et -1 (corrélation négative totale); 0 signifie *absence d'autocorrélation*, *pas de dépendance spatiale*, ou encore *espace géographique neutre*

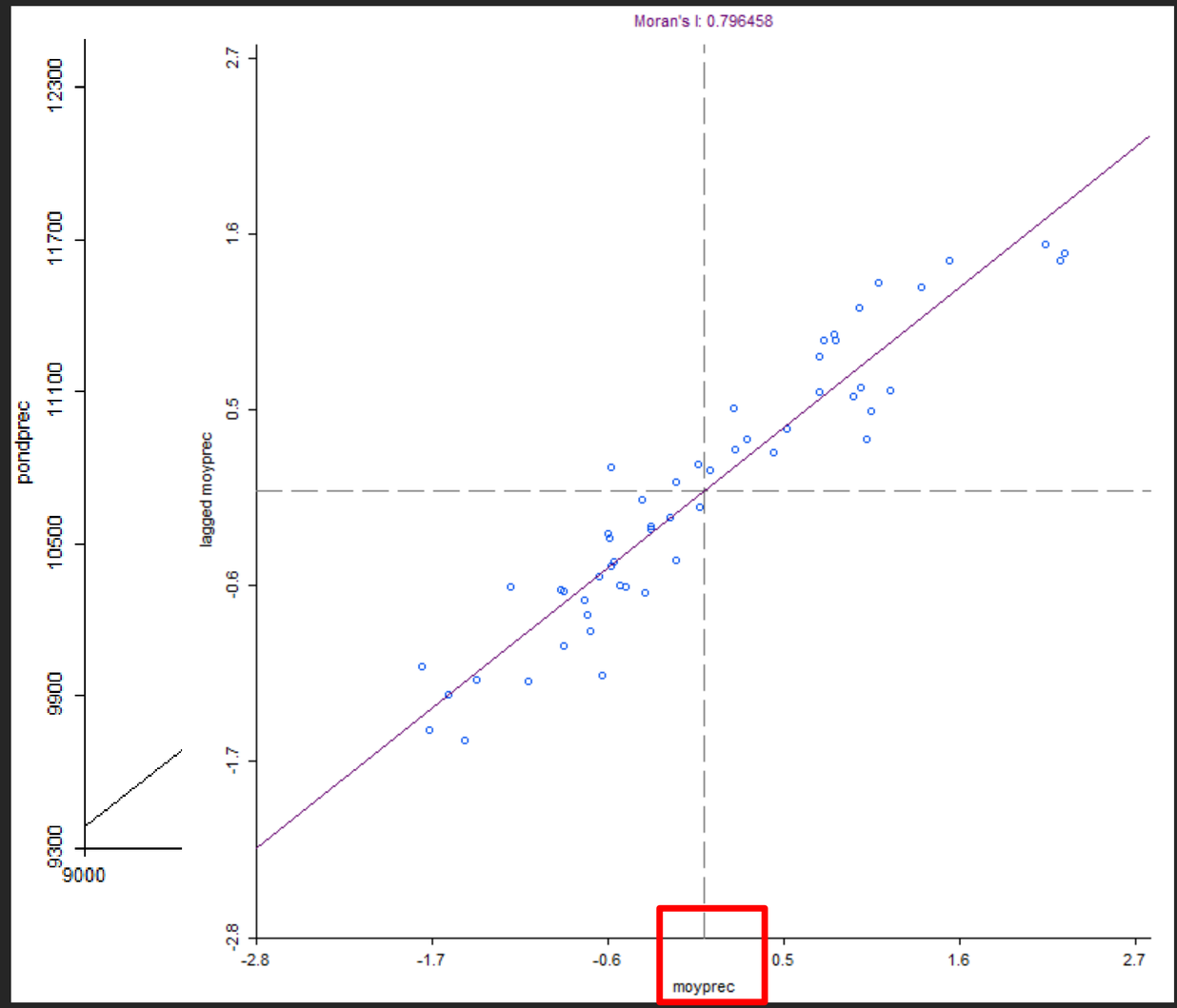
Le I de Moran comme coefficient de régression



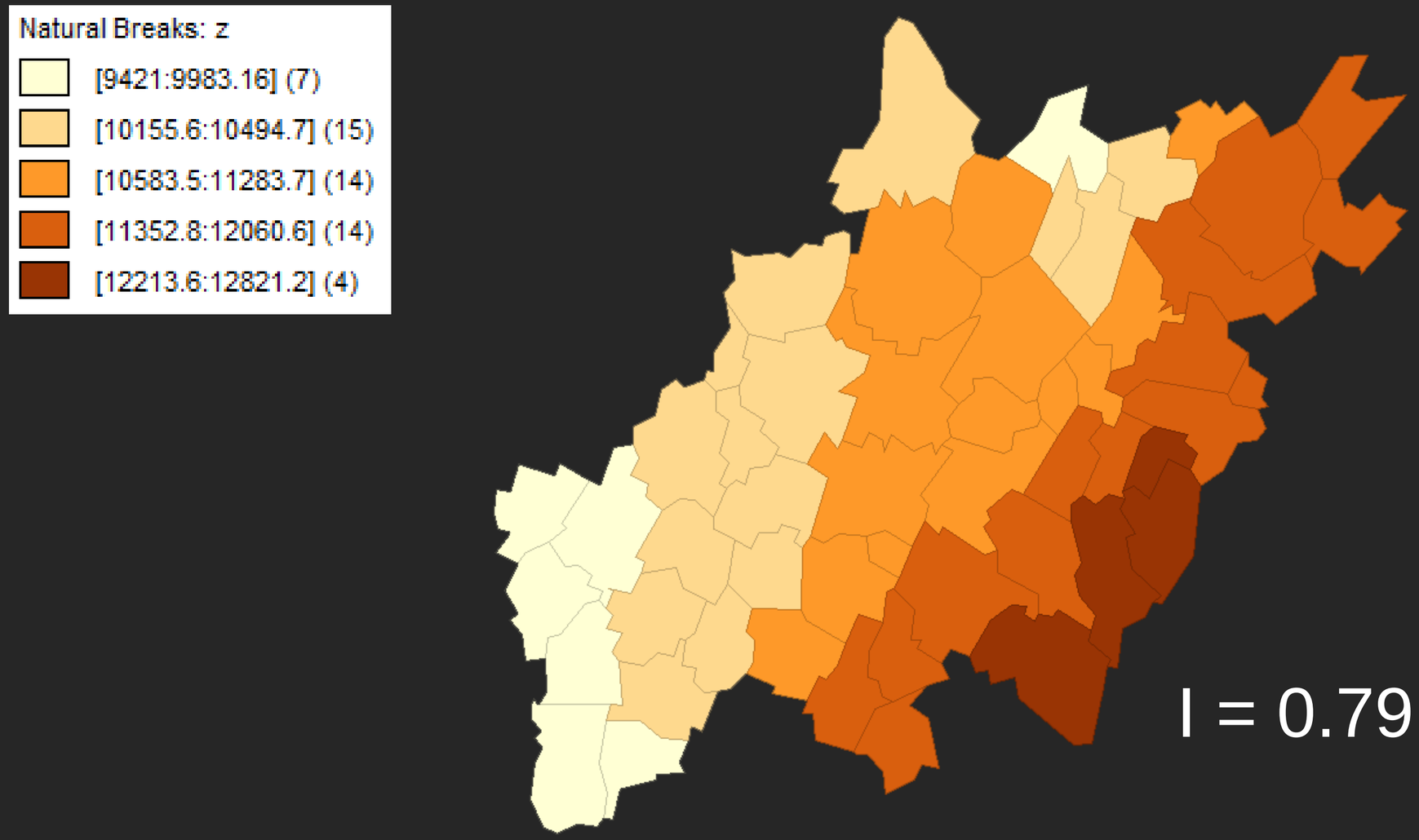
Le I de Moran comme coefficient de régression

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

...



Visualisation de la structure spatiale



Significativité du I de Moran et permutations aléatoires

Situation observée : $I \rightarrow I_0$

Tirage 1: $I = I_1$

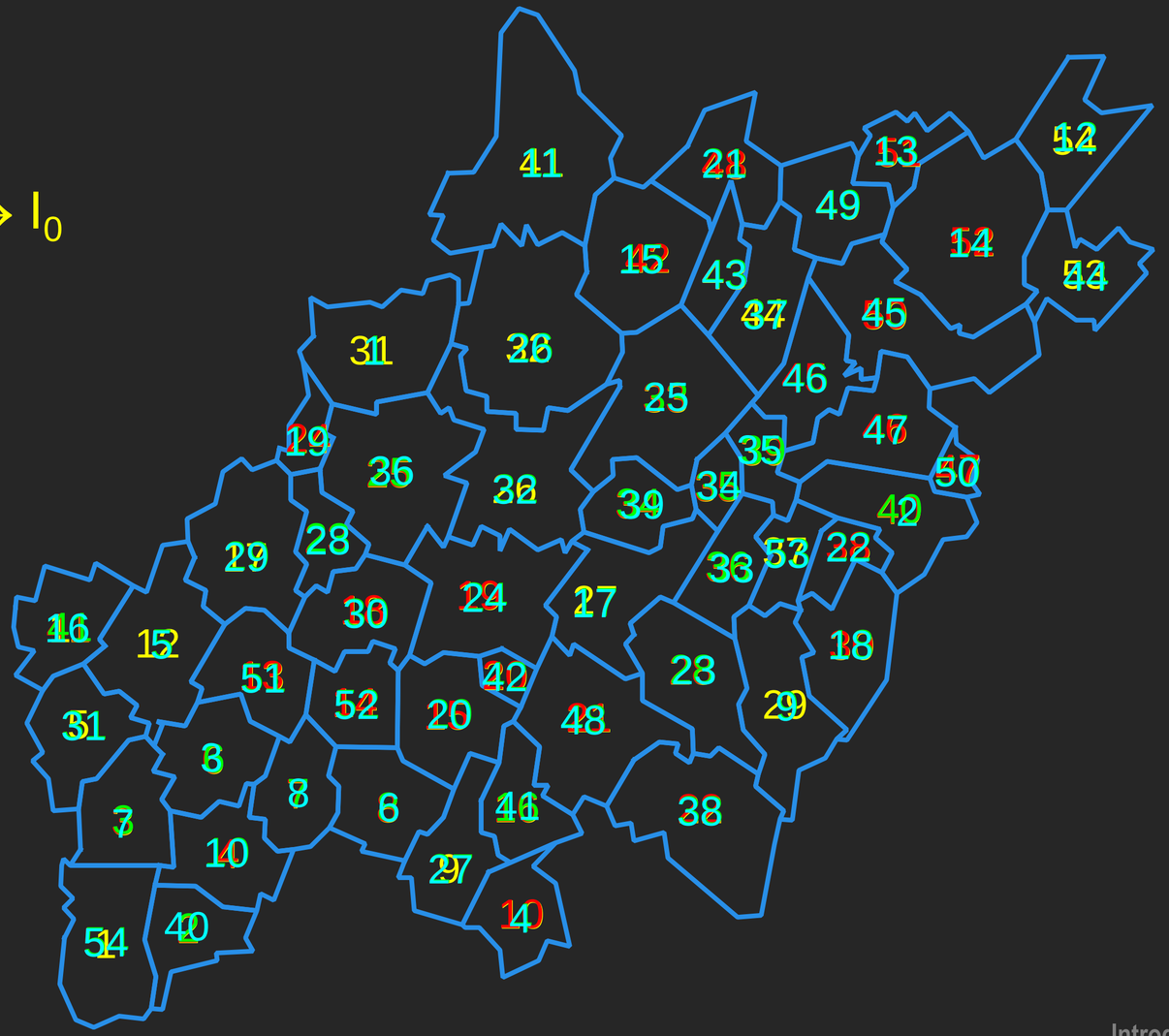
Tirage 2: $I = I_2$

Tirage 3: $I = I_3$

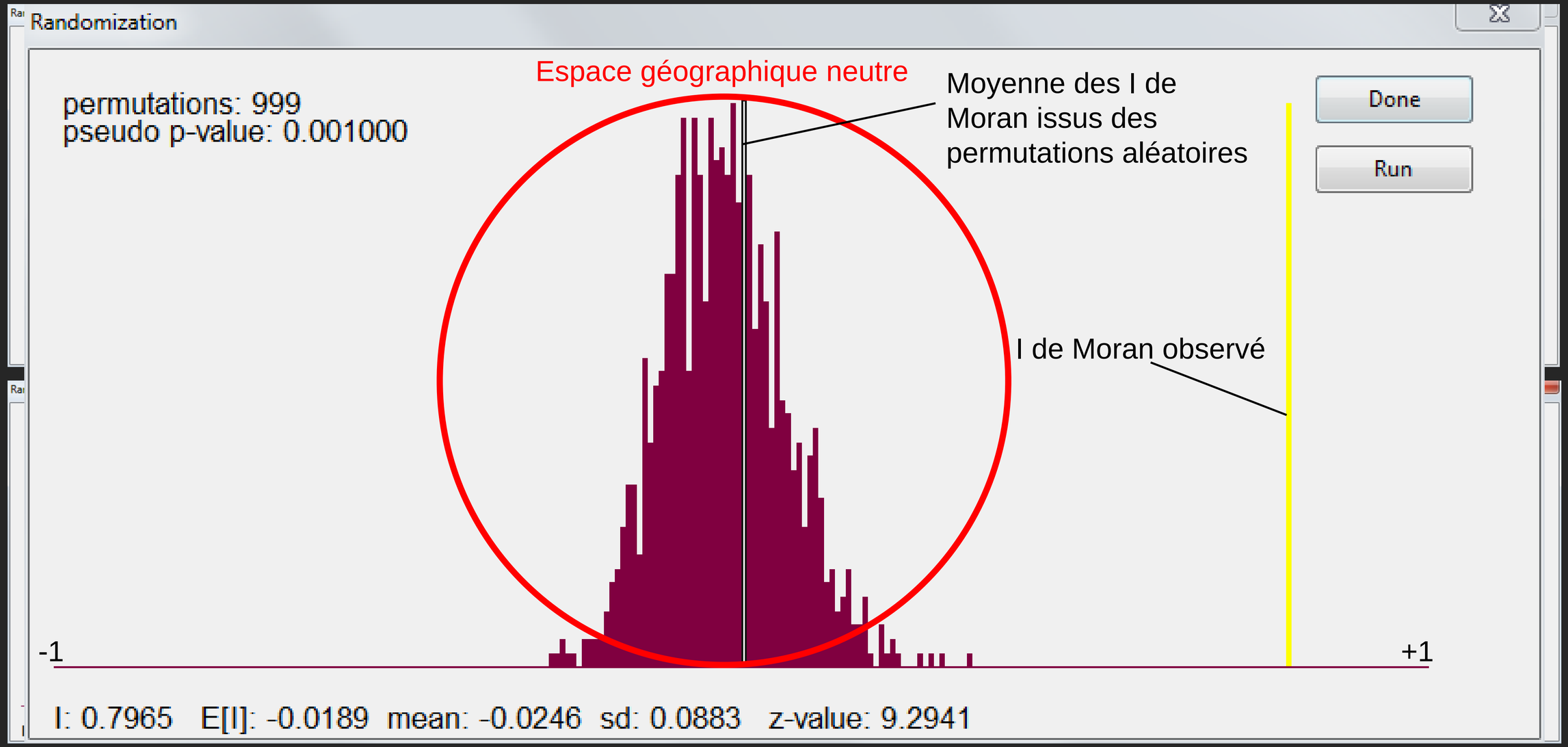
...

...

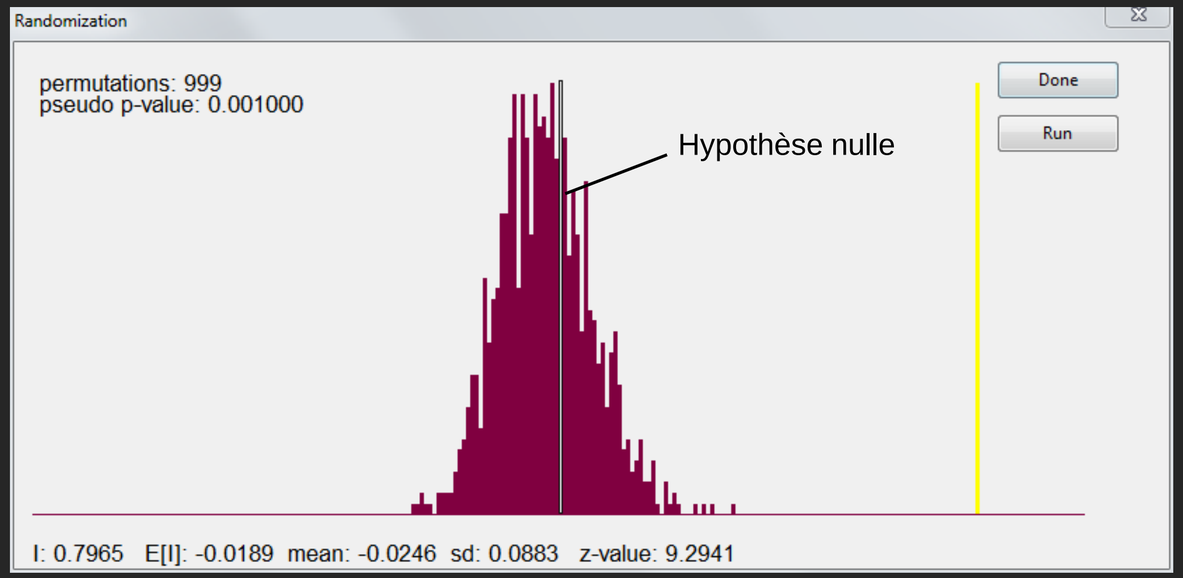
54! configurations possibles



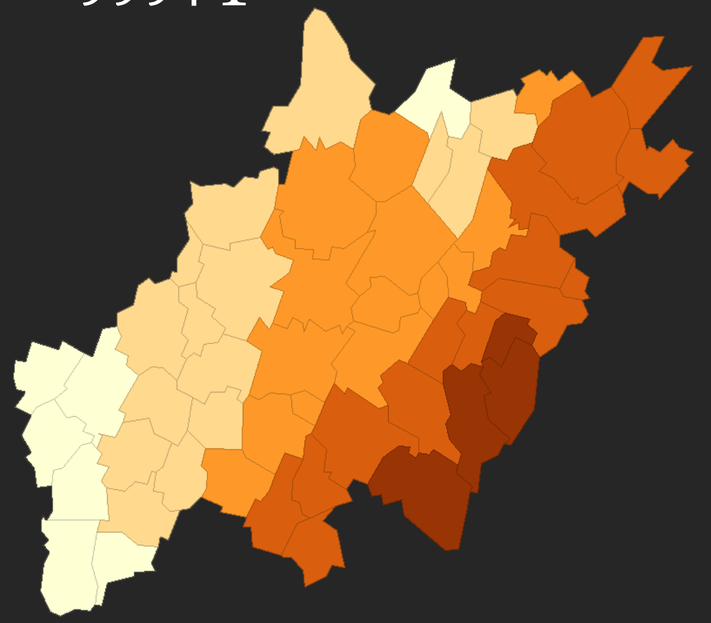
Histogramme des permutations et p-valeur



Calcul de la significativité et pseudo p-valeur



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



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

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

Le I de Moran de 0.79 traduit une structure spatiale significativement différente d'une distribution spatiale aléatoire

Systèmes d’Information Géographique

https://go.epfl.ch/sig

Autocorrélation spatiale locale – Le I de Moran local

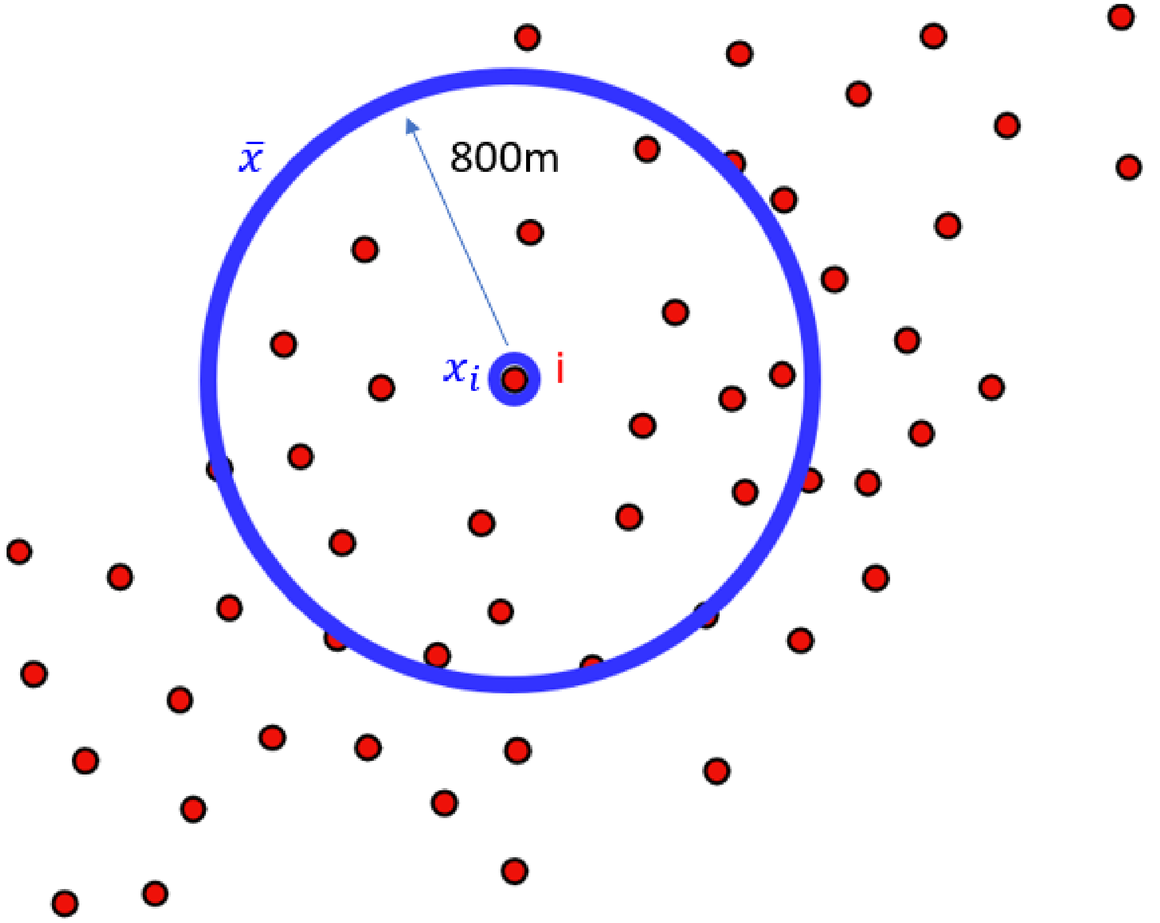
Stéphane Joost, Gabriel Kathari (GEOME-LGB)

Du I de Moran global au I de Moran local

- Le I de Moran global est une somme des produits croisés entre une valeur observée en un point *i* et sa moyenne dans un voisinage déterminé (spatial lag)
- Pour calculer un I de Moran local, on tire parti du fait que le I de Moran global est une somme de produits croisés individuels
- On peut évaluer la similarité entre les unités spatiales en calculant un I de Moran local pour chacune d'entre elles et en évaluant la significativité statistique pour chaque I obtenu
- La somme de tous les I de Moran locaux est égale au I de Moran global

Différence à la moyenne locale

$z_i = x_i - \bar{x}$ où $\bar{x}$ est la moyenne de la variable x dans le voisinage déterminé



Calcul du I de Moran local

Somme de produits croisés individuels

45	44	44
43	42	39
38	32	34

Pondération spatiale de Rook

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

Moyenne = 40.1

Variance (S^2) = 21.861

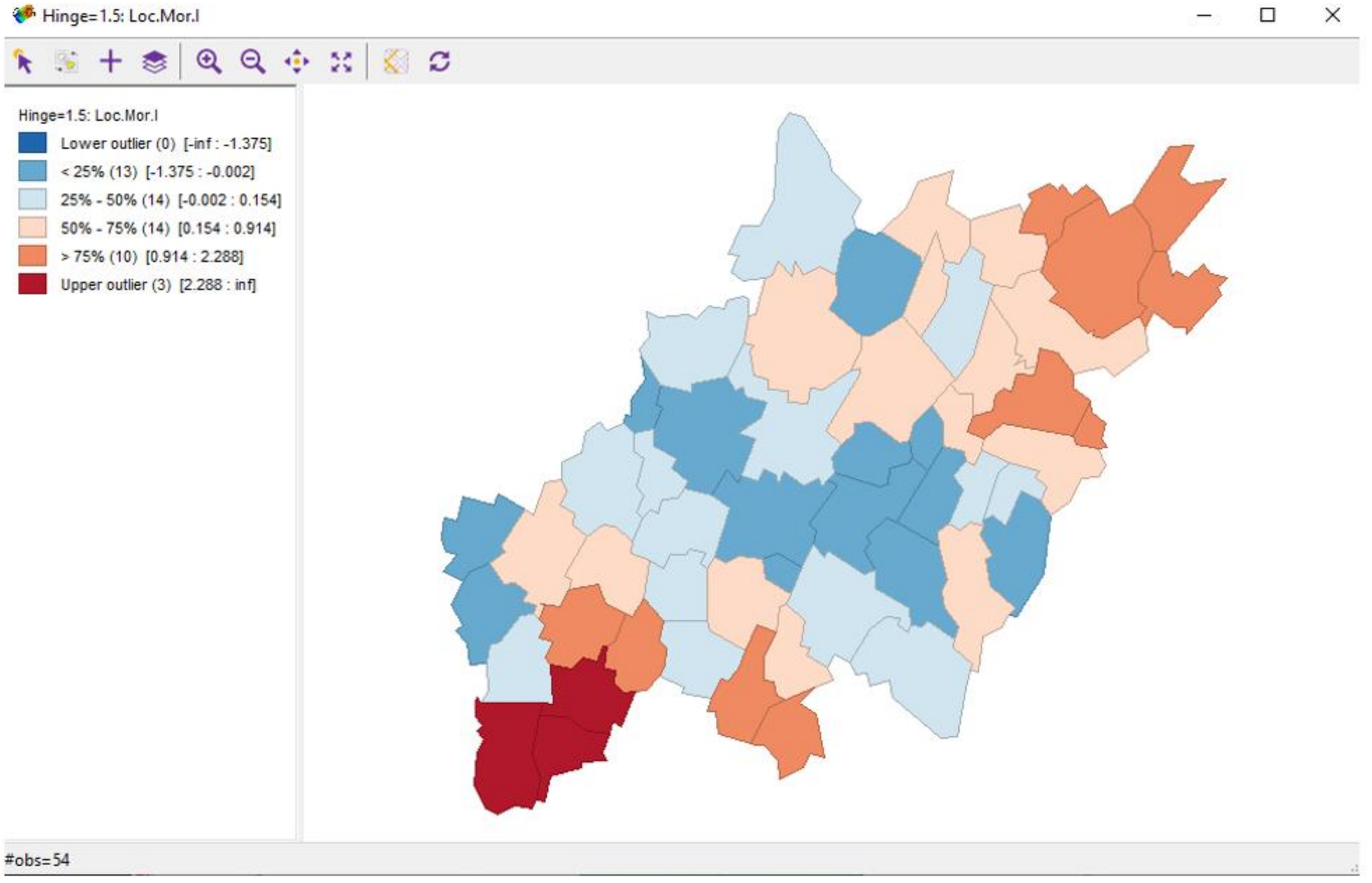
$$\frac{\sum (x - \bar{x})^2}{(n-1)}$$

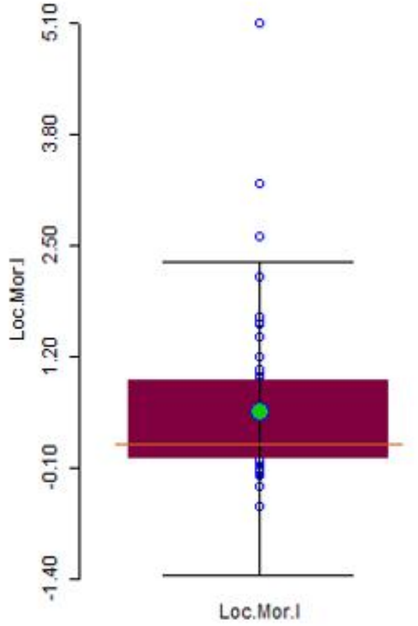
y_i	z_i	ω_{ij}	$w_{ij}z_j$
45	4.889	0	0
43	2.889	0.25	0.722
38	-2.111	0	0
44	3.889	0.25	0.972
42	1.889	0	0
32	-8.111	0.25	-2.028
44	3.889	0	0
39	-1.111	0.25	-0.278
34	-6.111	0	0
		1	-0.611

y = valeur de l'attribut
z = différence à la moyenne
ω = pondération

- z_i = 42 - 40.111 = 1.889
- La somme des poids multipliée par les différences à la moyenne = -0.611
- I_i = (1.889/21.861) · -0.611 = -0.053

Cartographie du I de Moran local



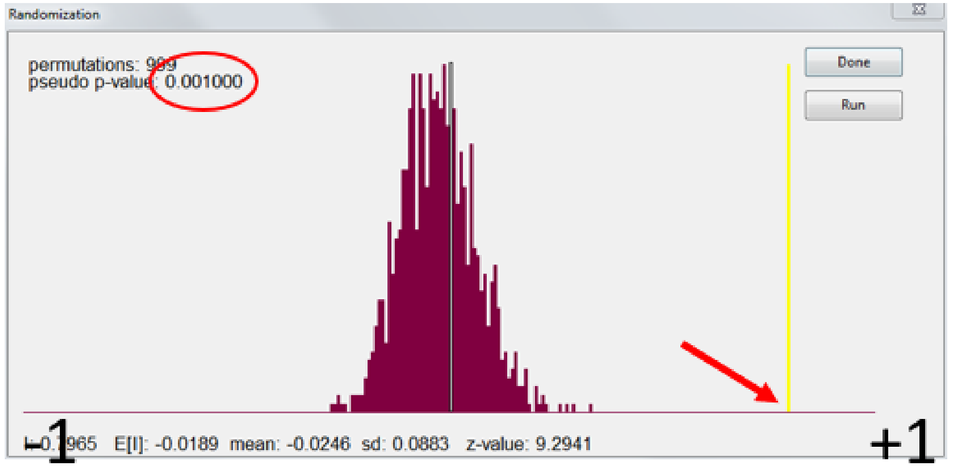


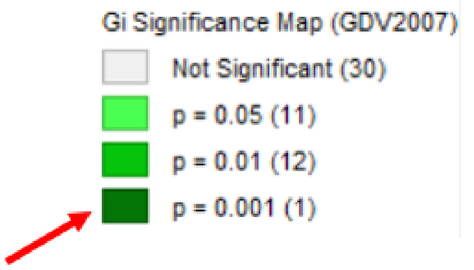
	Loc.Mor.I
min	-0.5625
max	5.0930
Q1	-0.0016
median	0.1541
Q3	0.9141
IQR	0.9157
mean	0.5430
s.d.	0.9792

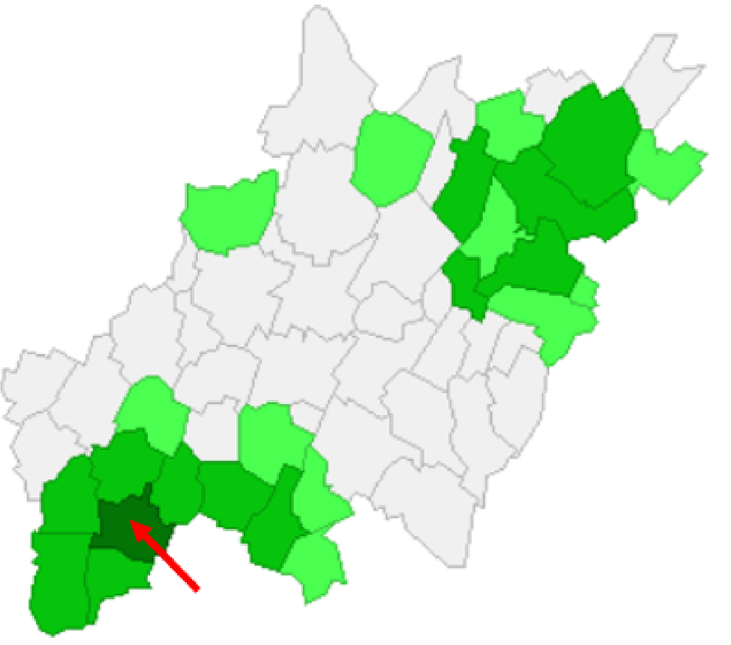
EPFL

Significativité statistique locale

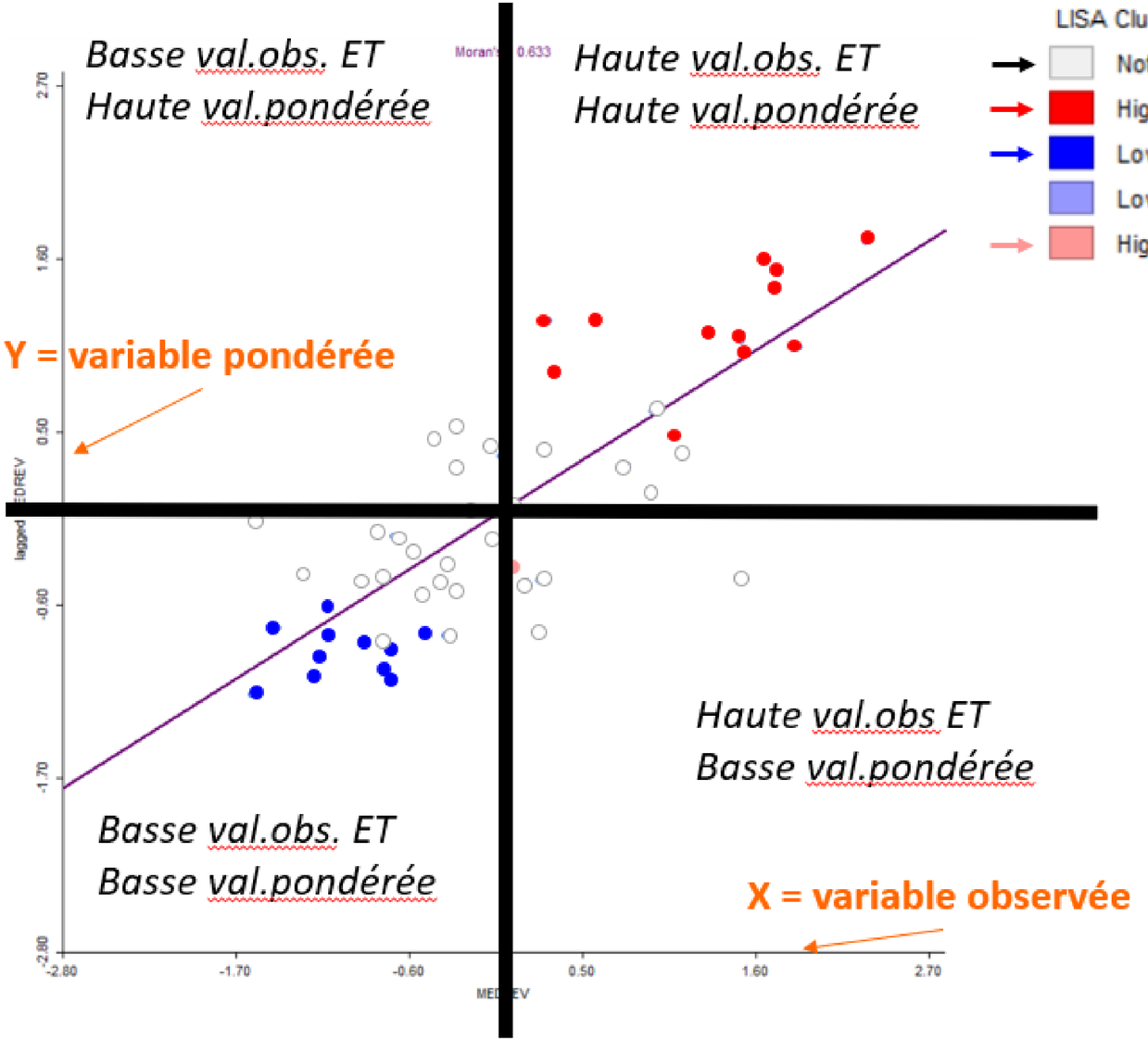
- Permutations aléatoires (méthode Monte-Carlo)
- A chaque permutation, chaque z_i est maintenue fixe, et le reste des z-valeurs sont permutées aléatoirement un grand nombre de fois
- Production d'une distribution statistique de référence

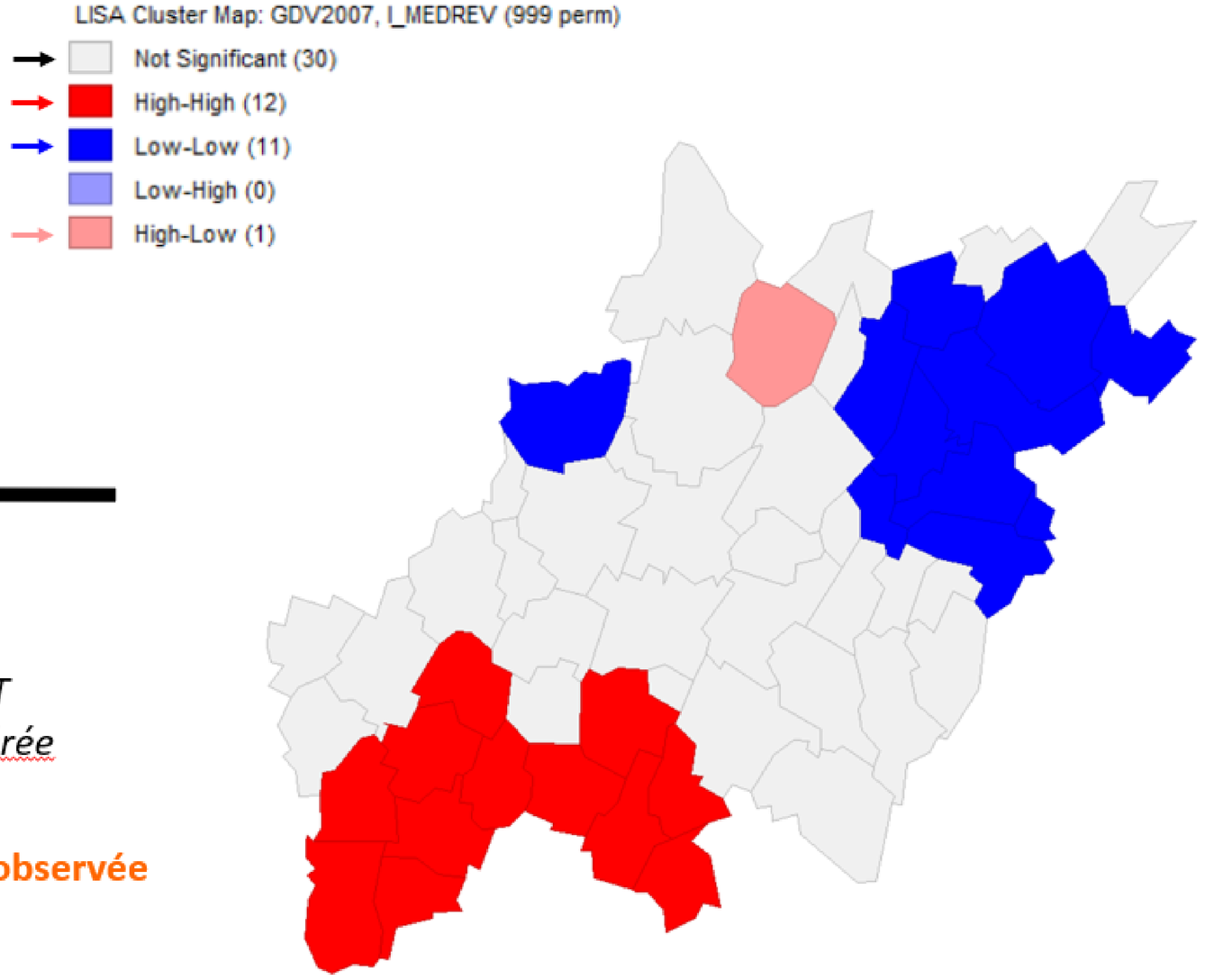




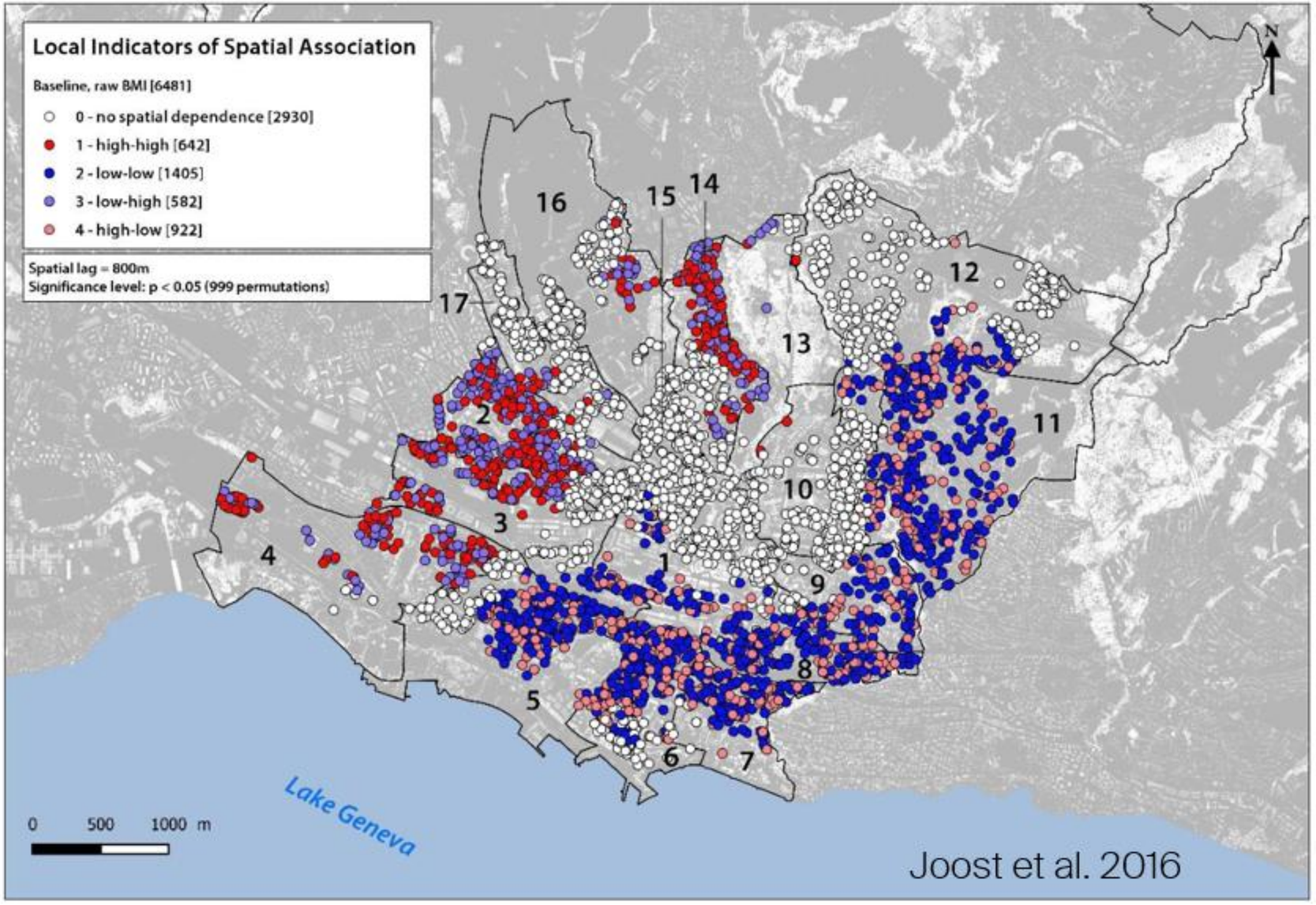


Scatterplot de Moran pour classification

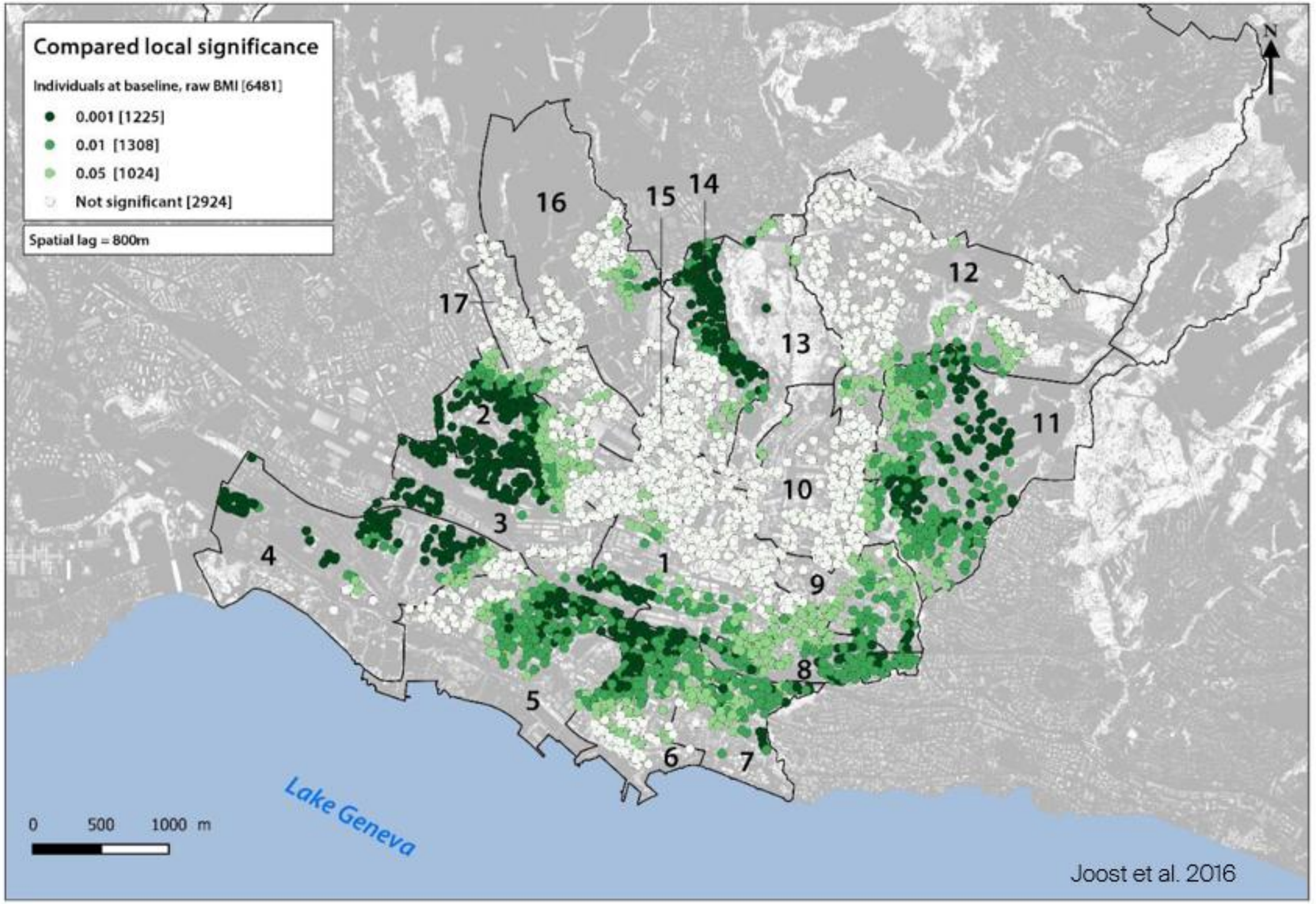




Dépendance spatiale de l'IMC



Variation de la significativité locale



Conclusion

- Mesure de la dépendance spatiale
- Statistiques basées sur une loi simple “Everything is related to everything else, near things are more related than distant things”
- Outil de la statistique confirmatoire (processus stationnaires) adapté au context spatial (processus non-stationnaires) grâce aux permutations aléatoires
- Outil puissant pour la détection de structures spatiales globales et locales
- Analyse exploratoire des données (EDA)


Merci pour votre attention !

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

$$I_i = \frac{(x_i - \bar{x})}{m_2} \sum_{j} w_{ij} (x_j - \bar{x})$$