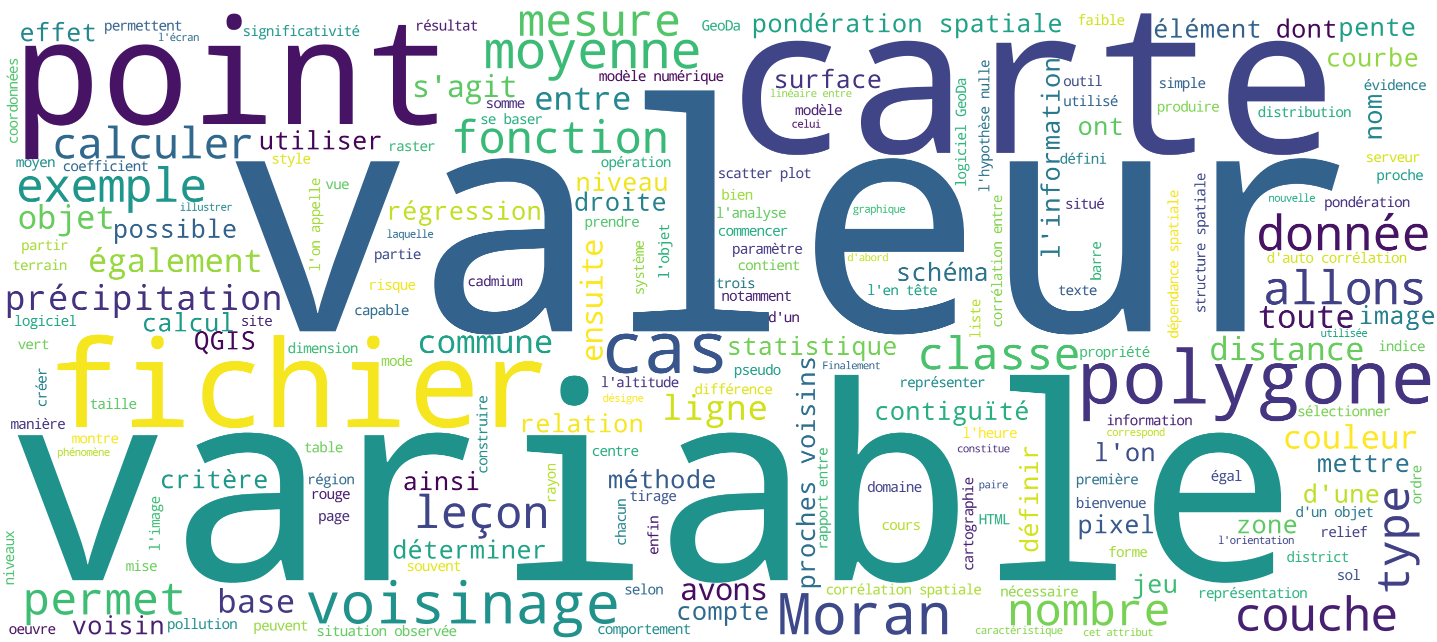


Search MOOC

Video

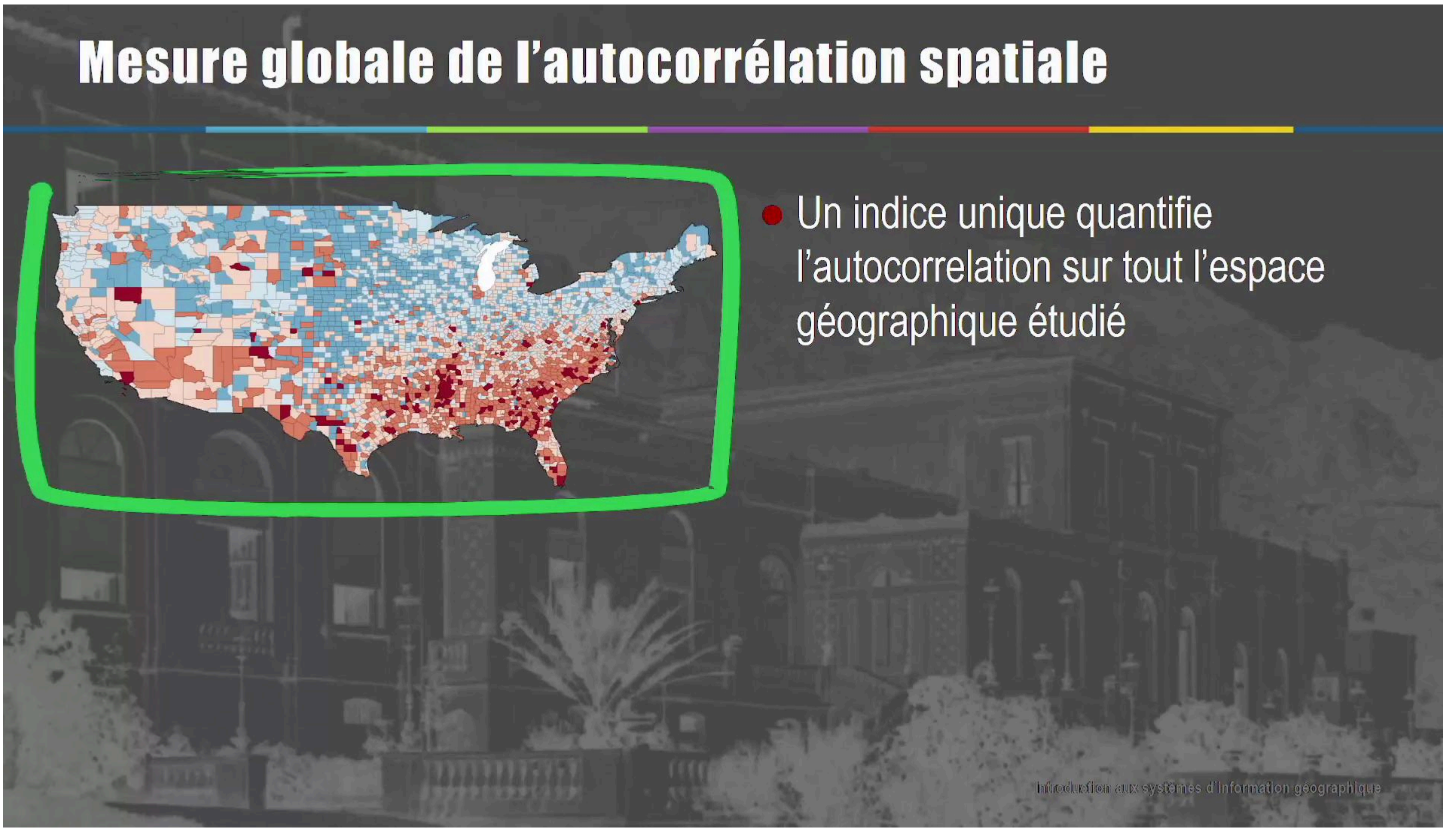
EPFL

Bonjour et bienvenue dans cette leçon qui porte sur une mesure de l'auto-correlation spatiale globale, le I de Moran. Un indice d'auto-correlation permet de quantifier la régularité ou la structure d'un phénomène spatialement distribué et ceci en tenant compte du voisinage de chacun des individus qui constitue le jeu de géodonnées. Nous verrons notamment comment tenir compte de ce voisinage pour élaborer des schémas de pondération spatiale. Ces schémas sont absolument nécessaires pour calculer un indice comme le I de Moran et en même temps estimer la portée de la dépendance spatiale. Vous apprendrez finalement comment évaluer la significativité de cet indice. Les buts de cette leçon sont de transmettre les notions nécessaires au calcul d'un indice d'auto-corrélation spatiale. Nous allons exposer les informations utiles à la définition de schémas de pondération spatiale et expliquer l'indice I de Moran, notamment son interprétation comme coefficient de régression qui fournit une approche intuitive et facile à retenir. Après avoir suivi nos explications, vous devriez être capable de déterminer un mode de pondération spatiale qui soit adapté aux jeux de données analysés. Et vous devriez également être capable de calculer le I de Moran global ainsi que sa significativité pour tout jeu de géodonnées ponctuel ou surfacique.

Notes

Summary

0m 31s

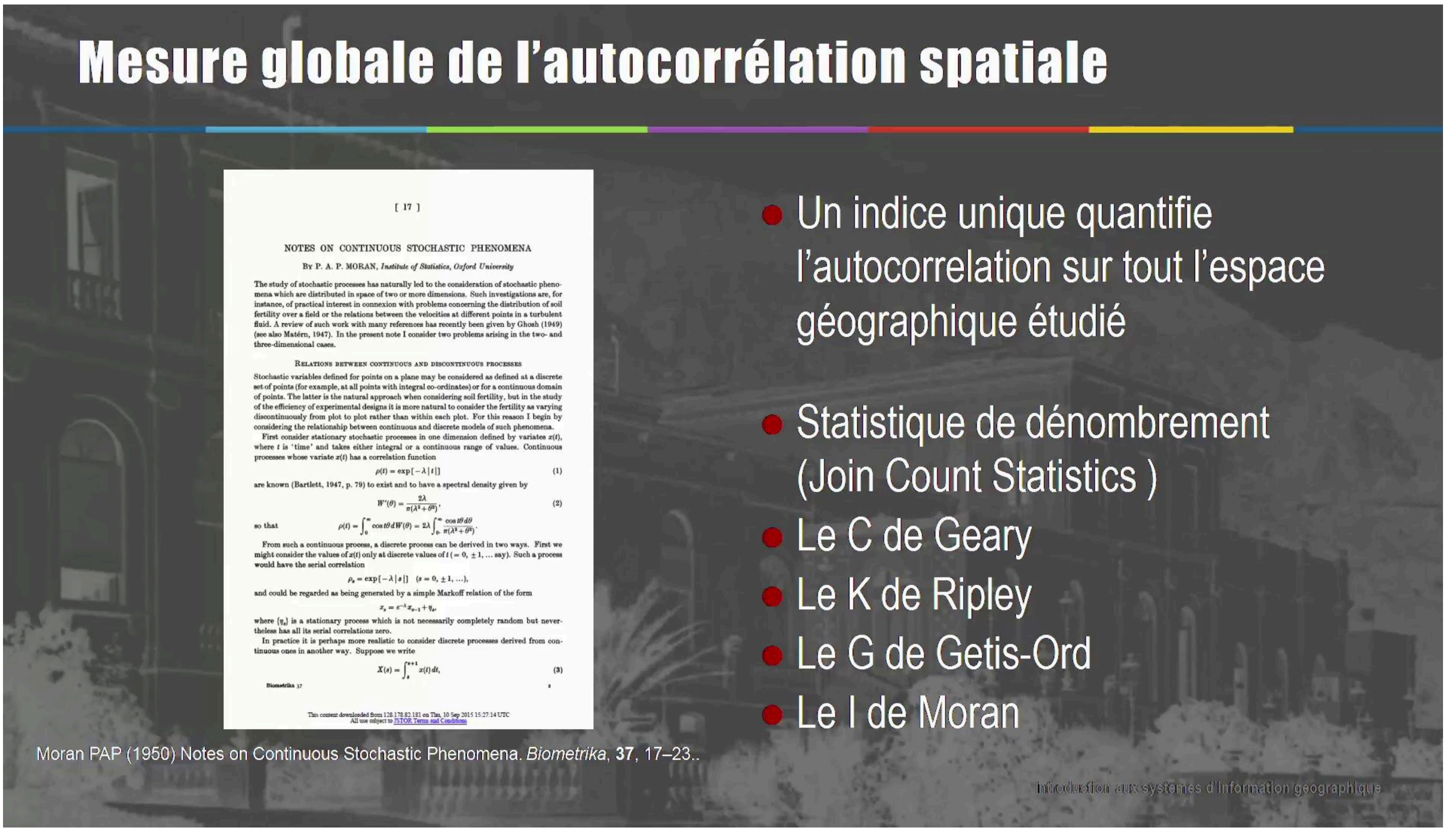


Nous présentons dans cette leçon, les notions requises pour calculer un indicateur global d'auto-correlation. Il s'agit en fait d'un indice unique qui caractérise l'arrangement spatial des unités géographiques selon les valeurs d'un attribut donné et ceci sur l'intégralité d'un territoire analysé.

Notes

Summary

1m 55s

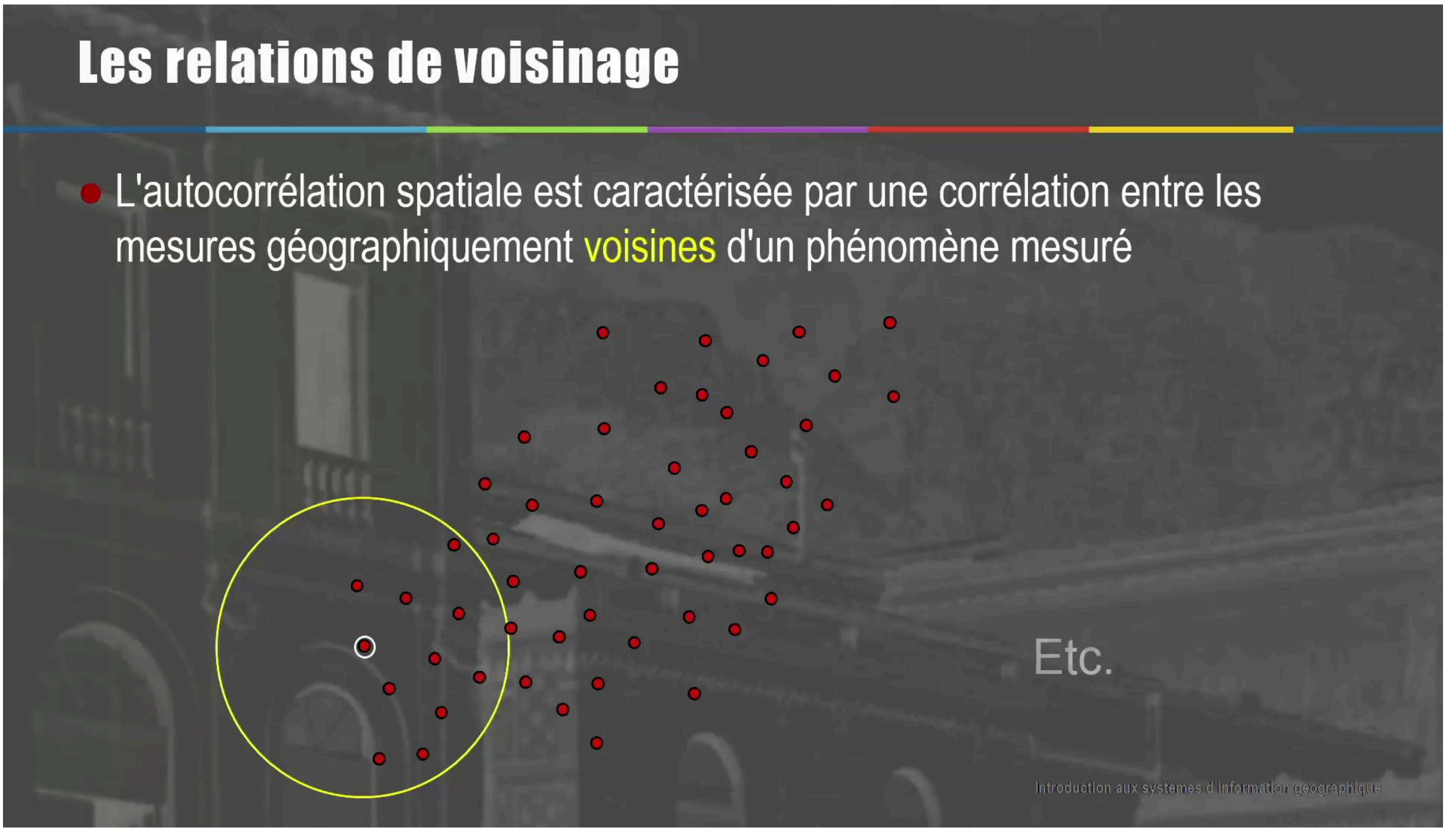


Il existe plusieurs méthodes de mesure : la Join Count Analysis qui est une statistique de dénombrement et qui a la particularité de n'être applicable qu'aux polygones. Et parmi les autres on trouve le C de Geary, le K de Ripley, le G de Getis-Ord et enfin le I de Moran qui est le plus répandu et qui fait l'objet de cette leçon.

Notes

Summary

2m 13s

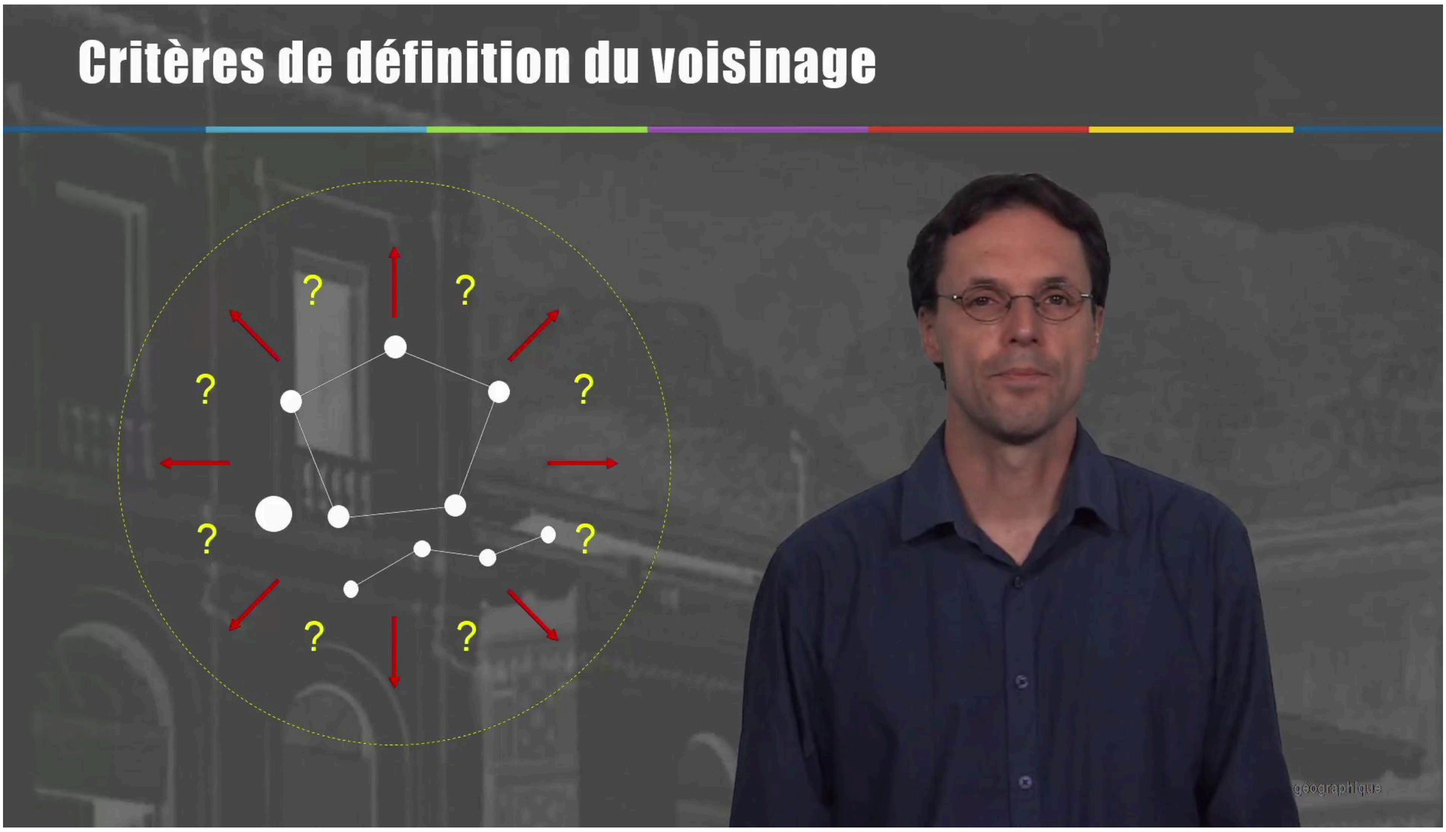


Pour quantifier la dépendance spatiale et produire une mesure d'auto-corrélation spatiale globale, il est nécessaire de tenir compte du voisinage de chacun des objets géographiques considérés. En effet la mesure de l'auto-correlation globale consiste à comparer le comportement d'un objet avec le comportement de ses voisins et ceci sur tout le territoire étudié. La clé étant de définir ce voisinage selon différents critères possibles. Prenons par exemple ce groupe de 54 points qui représente en l'occurrence les centres de masse ou centroïdes de commune pour lequel on décide de définir un voisinage de 5 km. Le cercle blanc identifie l'objet 1 et le cercle jaune délimite un rayon de 5 km autour de ce dernier. Le voisinage est ici défini par ce rayon de 5 km. Il permet ensuite de comparer la valeur d'un attribut A de l'objet 1 avec une statistique de cet attribut A pour les 17 objets mis en évidence en vert et situés dans ce voisinage. Cette statistique peut être la moyenne comme c'est le cas avec le I de Moran. Cette opération est répétée de manière à pouvoir comparer l'attribut A de chaque objet avec l'attribut A moyen de son voisinage, donc ici 54 fois.

Notes

Summary

2m 44s

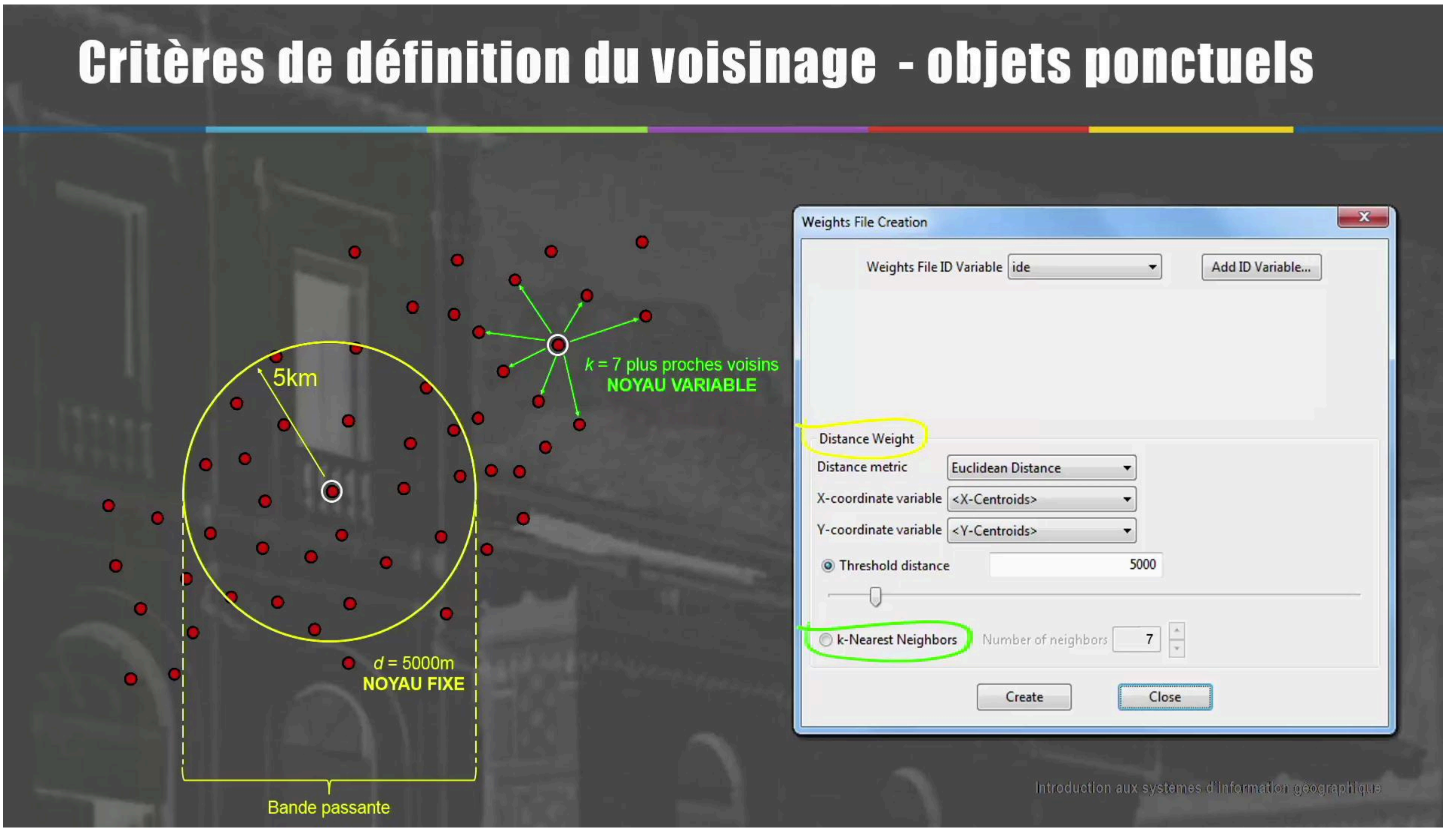


Il existe plusieurs critères sur lesquels se baser pour définir le voisinage d'un objet géographique. Et nous utiliserons ici l'interface du logiciel GeoDa pour illustrer la mise en oeuvre de ces critères.

Notes

Summary

3m 59s

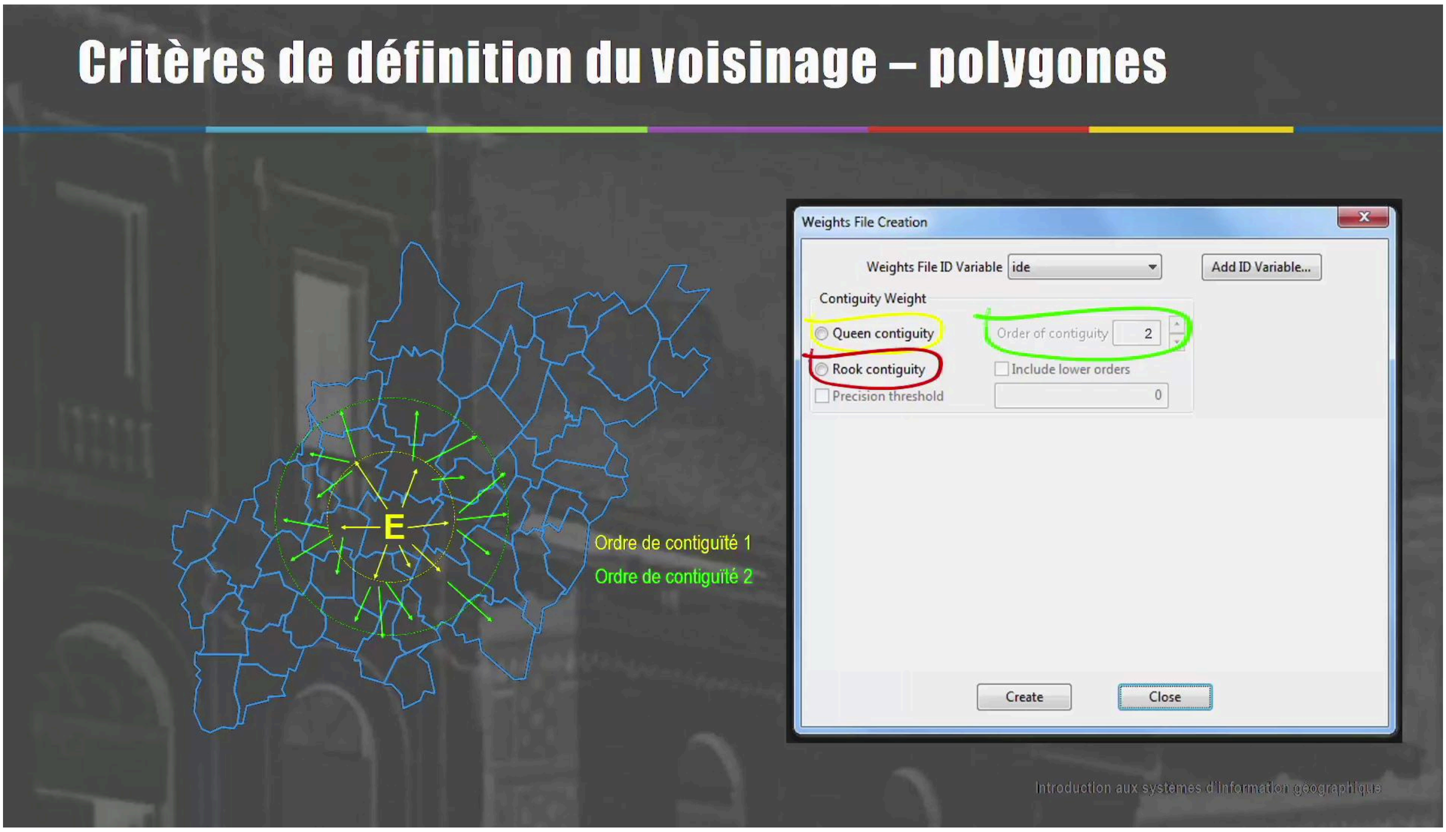


Les options à disposition pour définir un critère dépendent en partie du type d'objet considéré. Lorsque l'on désire définir le voisinage d'objets de type ponctuel, les critères les plus fréquemment utilisés sont de se baser soit sur une distance, soit sur un critère de proximité qui consiste à identifier les K plus proches voisins. En ce qui concerne la première option, la distance ou le rayon utilisé ici est de 5000 m et reprend l'exemple utilisé tout à l'heure. Dans ce cas, le voisinage n'est défini que sur la base de la distance entre le point et ses voisins et constitue ce que l'on appelle un noyau fixe. La distance D de 5000 m est ce que l'on appelle la bande passante. Dans le cas des plus proches voisins, ici K est égal à 7, le voisinage est ajusté en fonction de la densité d'objets aux alentours du point. Et on dit alors que la fonction associée utilise un noyau variable.

Notes

Summary

4m 12s

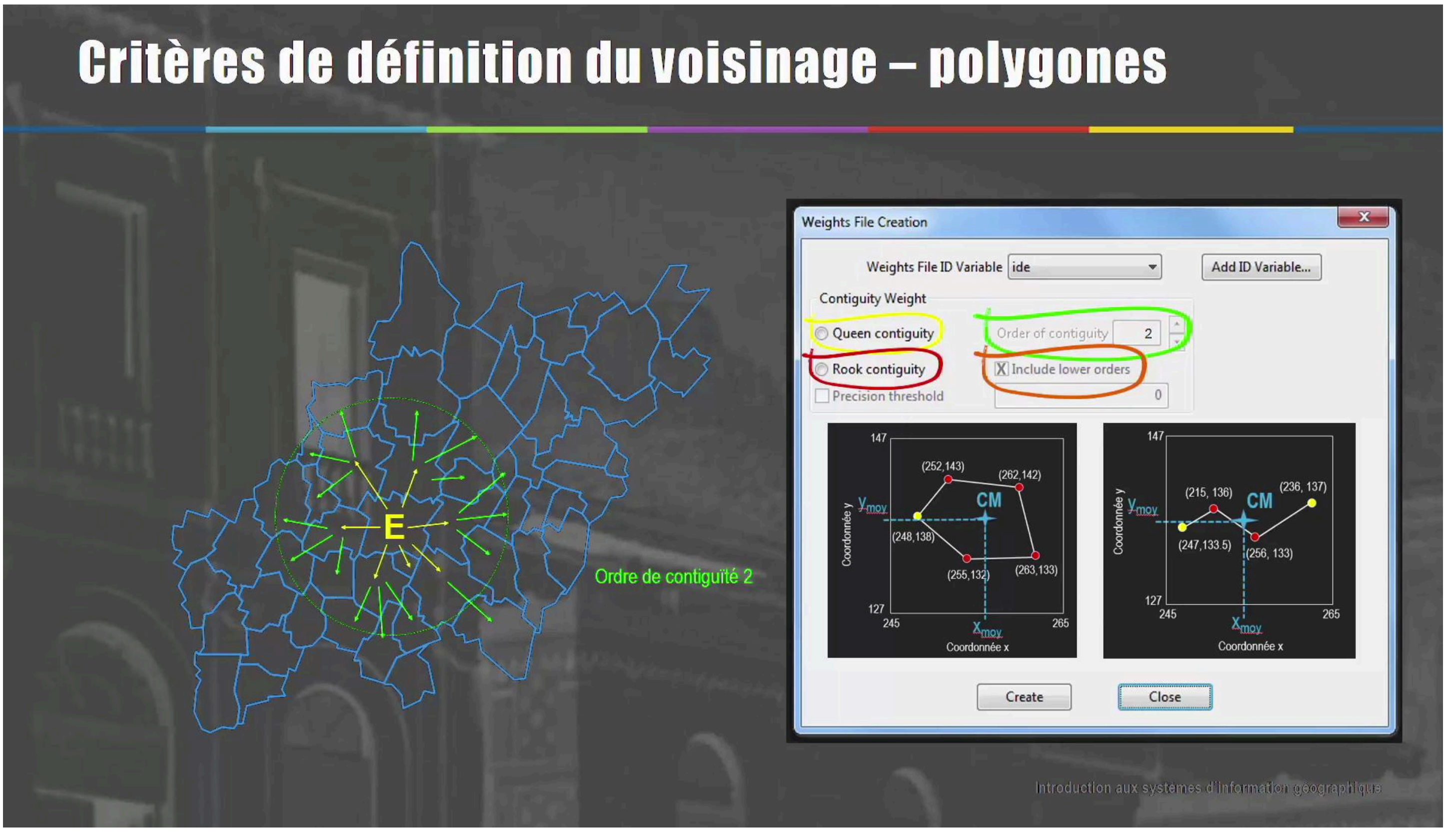


Les objets de type surfacique ou polygones permettent en plus de jouer sur les relations d'adjacence ou de contiguïté. Il y a 2 types de relation d'adjacence : les types Queen et Rook qui font référence aux déplacements des pièces dans un jeu d'échec. Le type Queen, soit la reine qui se déplace dans toutes les directions, correspond à un voisinage qui inclut tous les polygones qui touchent le polygone d'intérêt, il faut au moins une paire de coordonnées en commun. Le type Rook, soit la tour qui se déplace uniquement au nord, au sud, à l'est et à l'ouest, correspond à un voisinage qui inclut tous les polygones qui ont au moins 1 côté commun avec le polygone d'intérêt. Ce type de relation d'adjacence est principalement utilisé dans le cas où les unités géographiques sont des polygones orthogonaux comme beaucoup de comtés aux Etats-Unis par exemple. Il y a en outre un paramètre qui est l'ordre de contiguïté. En effet, on peut prendre en compte les voisins immédiatement contigus comme c'est ici le cas en jaune autour du polygone E ce qui correspond à un ordre de contiguïté 1, mais on peut également définir un voisinage qui repose sur un ordre de contiguïté 2, ici désigné par les flèches vertes, c'est à dire d'inclure les polygones qui, en fonction du type Queen ou Rook, sont voisins des polygones d'ordre 1.

Notes

Summary

5m 10s

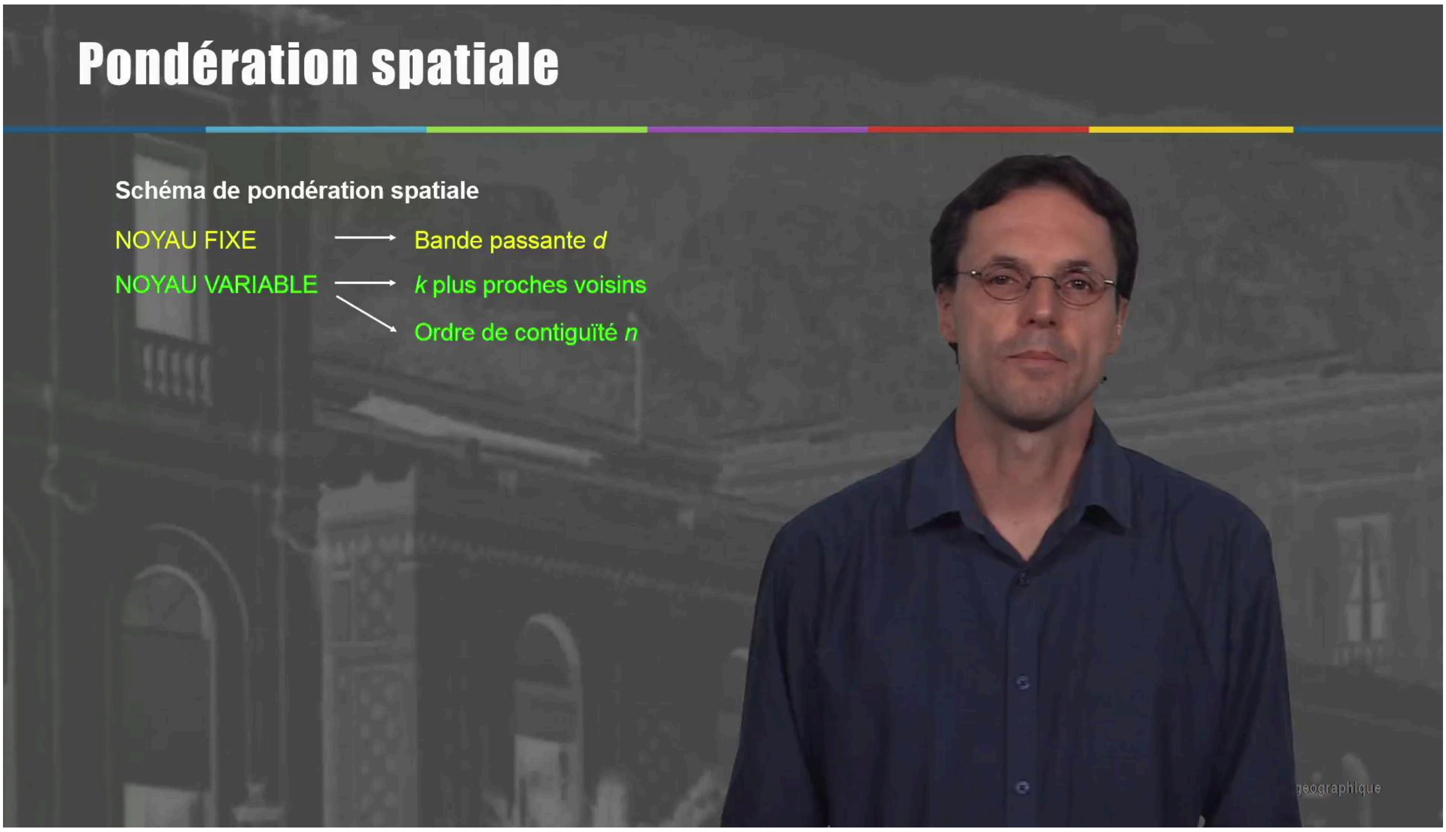


Finalement, selon les caractéristiques de l'analyse en cours, il est possible d'inclure ou non dans le voisinage les ordres de contiguïté de rang inférieur. Les noyaux fixes comme la distance, les noyaux variables comme les K plus proches voisins sont applicables aux polygones, mais ce sont dans ce cas les coordonnées géographiques du centre de gravité géométrique de l'objet qui sont pris en compte pour effectuer les calculs. C'est également le cas quand on travaille avec des objets de type linéaire.

Notes

Summary

6m 29s

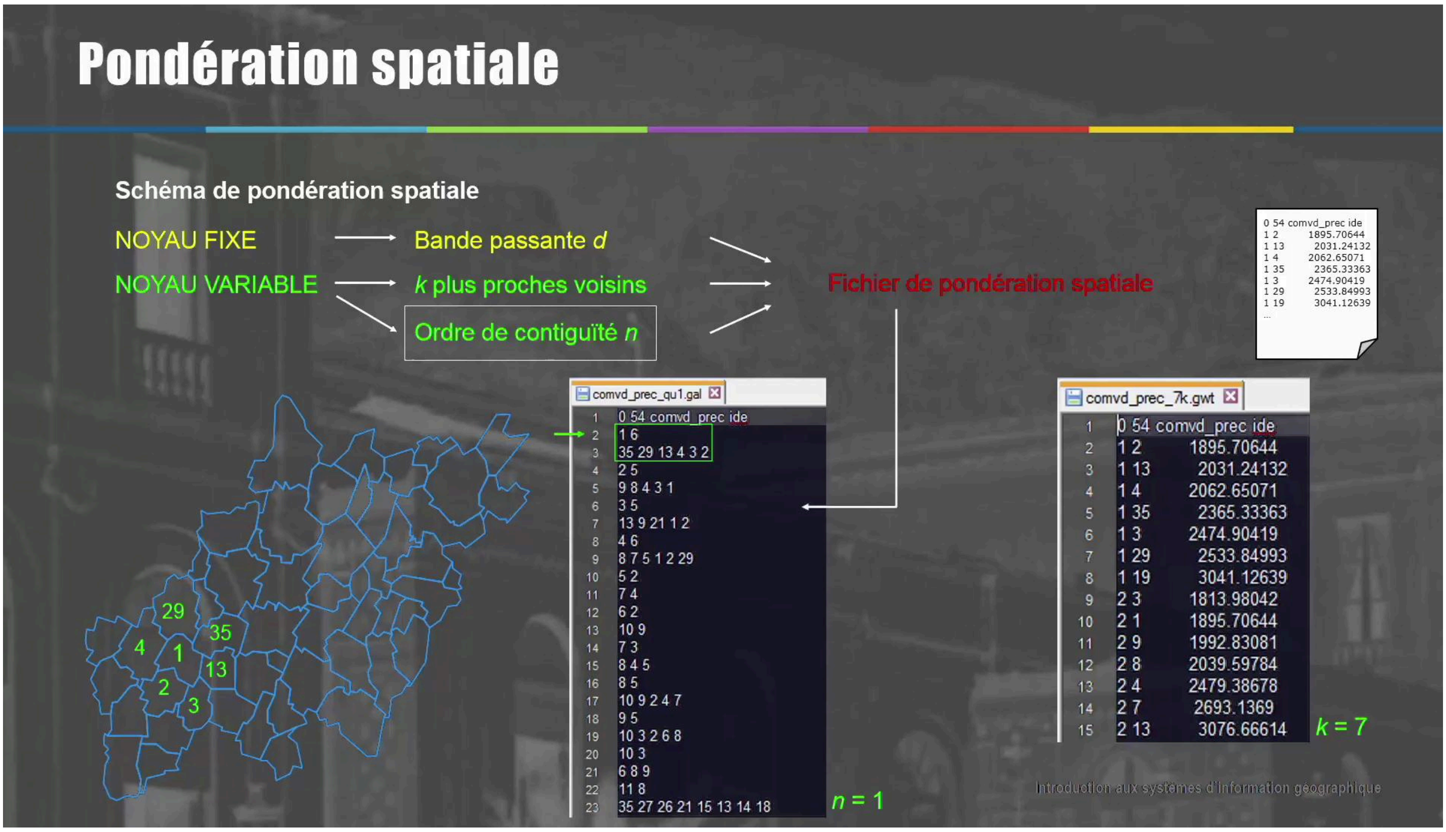


Les relations de voisinage permettent de déterminer le schéma de pondération spatiale que l'on souhaite appliquer à un jeu de données. Ce schéma de pondération sera appliqué pour créer un fichier contenant pour chaque objet la liste de ses voisins.

Notes

Summary

7m 07s



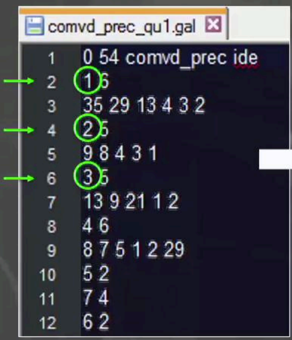
Il s'agit du fichier de pondération spatiale. Voilà ce que l'on peut lire quand on édite ce fichier. Nous reprenons ici l'exemple des 54 communes utilisé précédemment. La première ligne constitue l'en-tête et contient 4 éléments séparés par un espace. Le 0 qui n'a pas de fonction dans la version actuelle de GeoDa, 54 qui désigne le nombre de polygones dans le jeu de données, le nom du fichier et le nom de l'identifiant unique. Le reste du fichier établit la liste des voisins de chaque polygone. A la ligne 2, le chiffre 1 désigne le polygone numéro 1 et le chiffre 6 indique que ce polygone a 6 voisins selon le critère indiqué et que ce sont les polygones numéro 35 - 29 - 13 - 4 - 3 et 2 indiqués à la ligne 3. Et ainsi de suite, à la ligne 4, le polygone 2 a 5 voisins qui sont les polygones 9 - 8 - 4 - 3 et 1. Avec le critère des K plus proches voisins illustré sur la droite, après l'en-tête, le fichier contient les 7 plus proches voisins du polygone 1 avec sur la droite la distance qui sépare les centroïdes puis dès la ligne 9, les 7 plus proches voisins du polygone 2, etc.

Notes

Summary

7m 22s

Corrélation entre mesures spatialement voisines…



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

Introduction aux systèmes d'information géographique

L'auto-corrélation spatiale est quantifiée par le calcul d'une corrélation entre les mesures géographiquement voisines d'un phénomène mesuré. Avec le I de Moran, on calcule la corrélation entre l'attribut analysé d'un objet géographique et la moyenne de cet attribut calculée au sein du voisinage défini. On va donc utiliser le fichier de pondération spatiale de contiguïté décrit il y a un moment pour calculer la moyenne de cet attribut dans le voisinage de chacun des objets constituant le jeu de données. La variable notée Z que nous utilisons dans cet exemple représente la moyenne mensuelle de la somme des précipitations exprimée en dixième de millimètre par commune. Cette variable Z est la troisième colonne du tableau après l'identifiant IDE et la liste des polygones adjacents qui permet de calculer la moyenne de Z qui figure dans la dernière colonne. Les voisins listés dans la deuxième colonne permettent de déterminer ω, soit le poids attribué aux polygones dans le calcul de la moyenne de Z dans le voisinage. Dans le cas du critère de contiguïté, ce poids est de 1 si un polygone est adjacent et de 0 s'il ne l'est pas. Le coefficient d'auto-corrélation I de Moran est une extension du coefficient de corrélation de Bravais-Pearson.

Notes

Summary

8m 41s

Corrélation entre mesures spatialement voisines…

```
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
```

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*

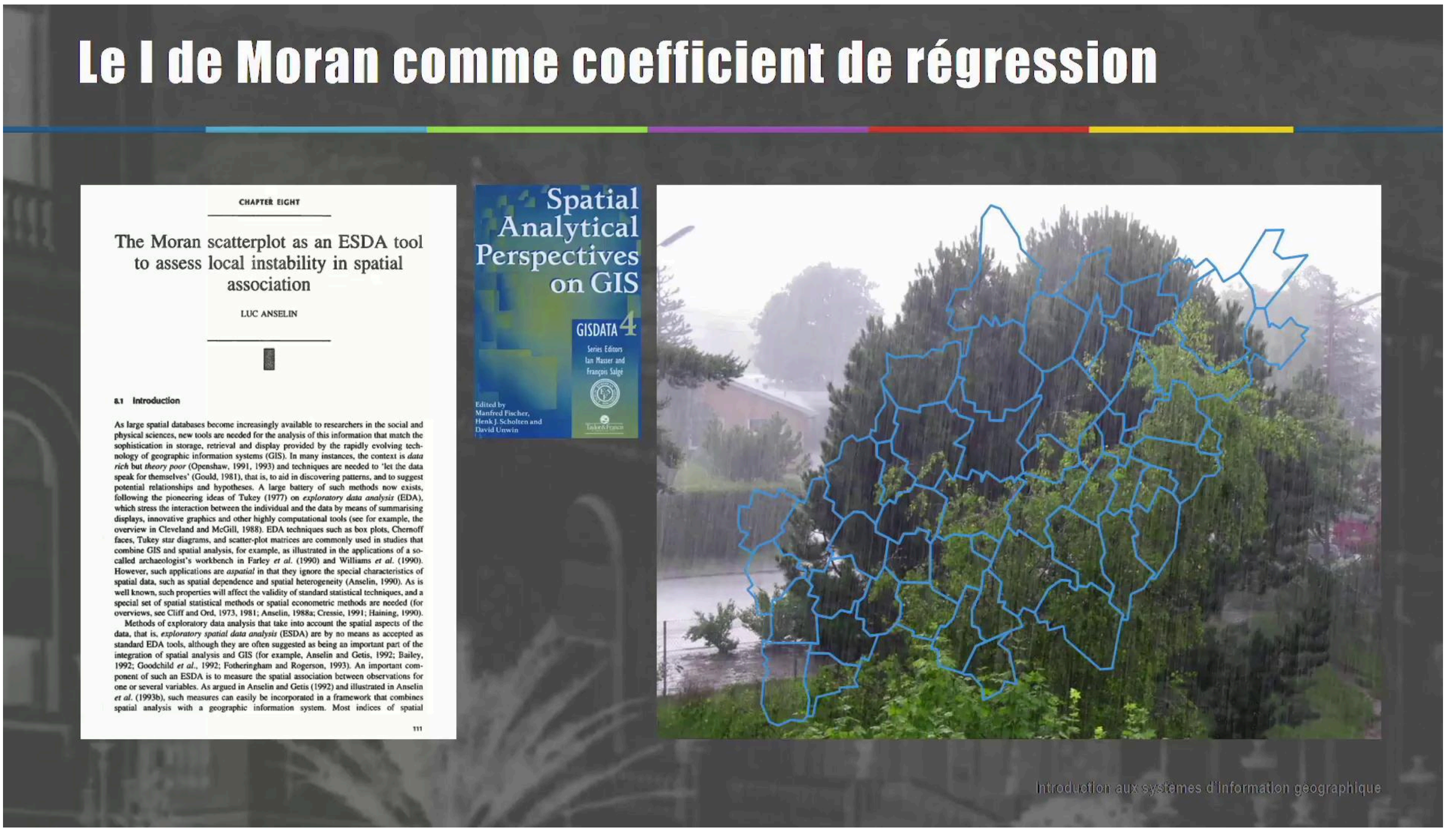
Introduction aux systèmes d'information géographique

Il exprime l'importance de la différence des valeurs d'une variable entre toutes les paires d'objets géographiques voisins. Il est défini comme on le voit ici dans la formule par le rapport entre la covariance d'une variable, par rapport à la moyenne de cette variable et sa variance sur toute la zone étudiée. Il est une mesure de la corrélation linéaire entre 2 variables, soit ici Z et Z barre, et produit une valeur comprise entre +1 et -1 où 1 signifie qu'il y a corrélation positive totale, 0 pas de corrélation et -1 corrélation totale négative ou inverse.

Notes

Summary

9m 57s



En 1996, Luc Anselin propose une interprétation du I de Moran comme coefficient de régression. Cette interprétation permet de bien comprendre le calcul du I de Moran via les outils de statistique exploratoire implémentés dans le logiciel GeoDa. Nous allons la mettre en oeuvre en calculant le I de Moran sur la variable de précipitation qui caractérise les 54 communes de notre jeu de données.

Notes

Summary

10m 35s

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

...

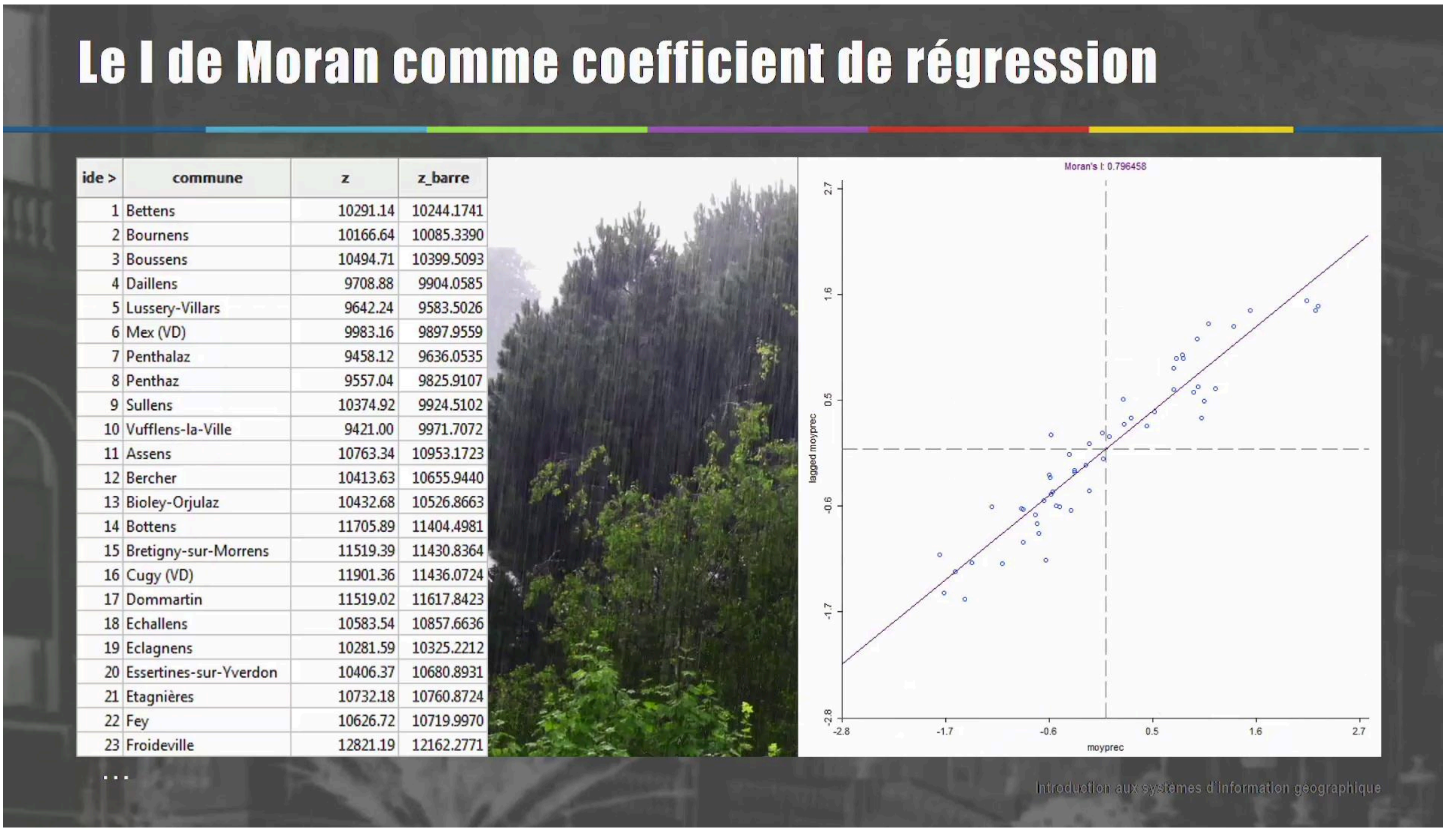
Introduction aux systèmes d'information géographique

Nous allons utiliser le même schéma de pondération soit un noyau variable avec un ordre de contiguïté de Queen de 1.

Notes

Summary

11m 00s



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

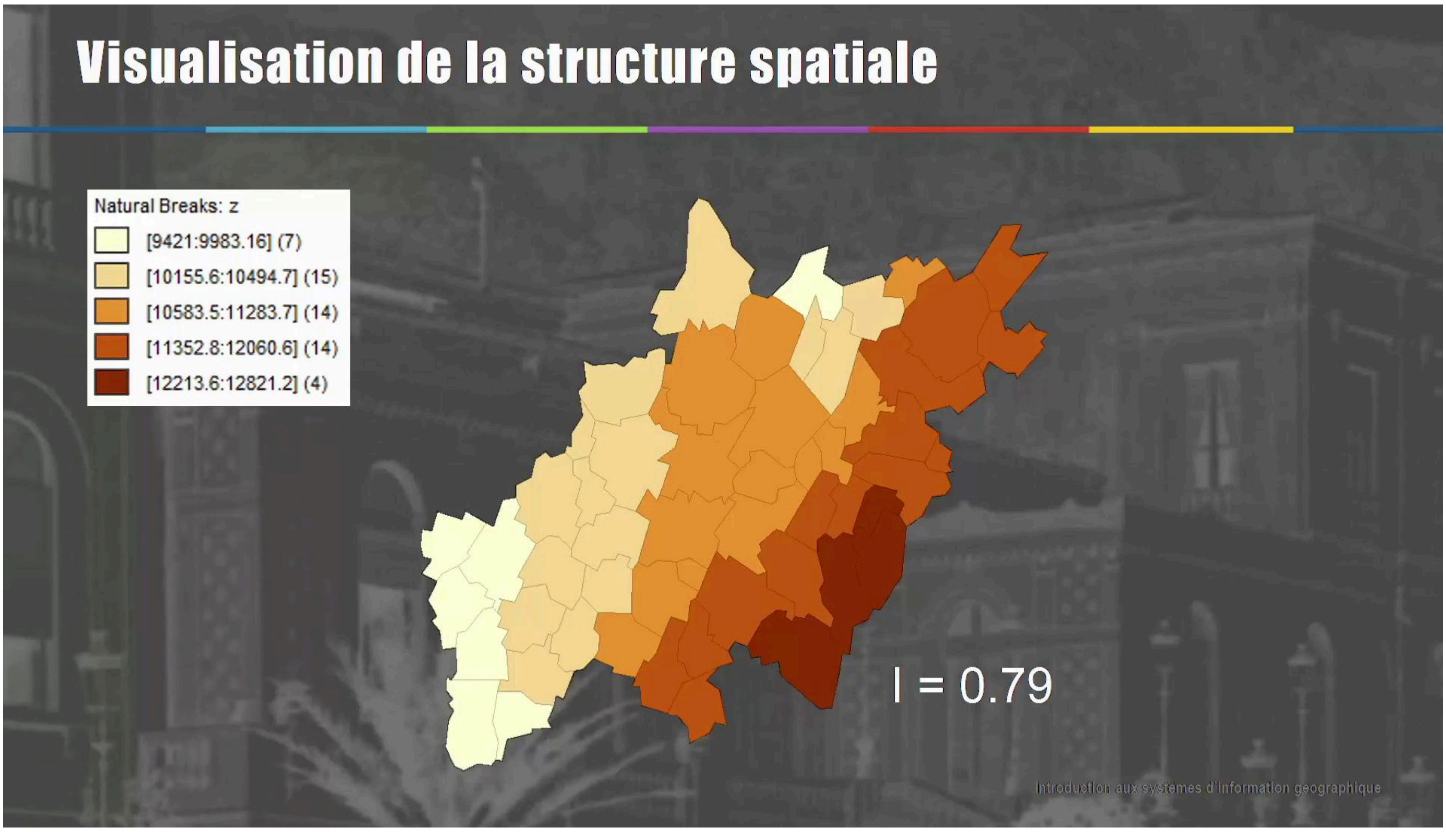
...

La formule du I de Moran revient à calculer la corrélation entre la valeur de précipitation pour chaque commune, c'est la variable z dans la table, et la moyenne des précipitations pour les voisins de chaque commune, c'est la variable z barre dans la table. En appliquant une régression de la variable indépendante Z sur la variable dépendante Z barre la pente de la droite de régression est égale au I de Moran. Par l'intermédiaire d'un scatter plot bivarié, cette interprétation fournit un moyen de visualiser la relation linéaire entre la variable étudiée et la valeur moyenne de cette variable dans son voisinage. Ici, le paramètre bêta, soit la pente, est égal à 0,79. Le scatter plot de Moran montre la même relation mais sur les valeurs centrées et réduites. Ce I de Moran élevé montre que les valeurs de précipitation sur les communes de ce territoire ressemblent à la moyenne de leurs voisines et qu'il y a donc auto-corrélation spatiale, que la variable de précipitation utilisée est donc spatialement dépendante.

Notes

Summary

11m 14s



Cette carte thématique de la variable de précipitation en 5 places met en évidence la structure spatiale révélée par le I de Moran. En effet, l'intensité des précipitations décline progressivement de l'est vers l'ouest.

Notes

Summary

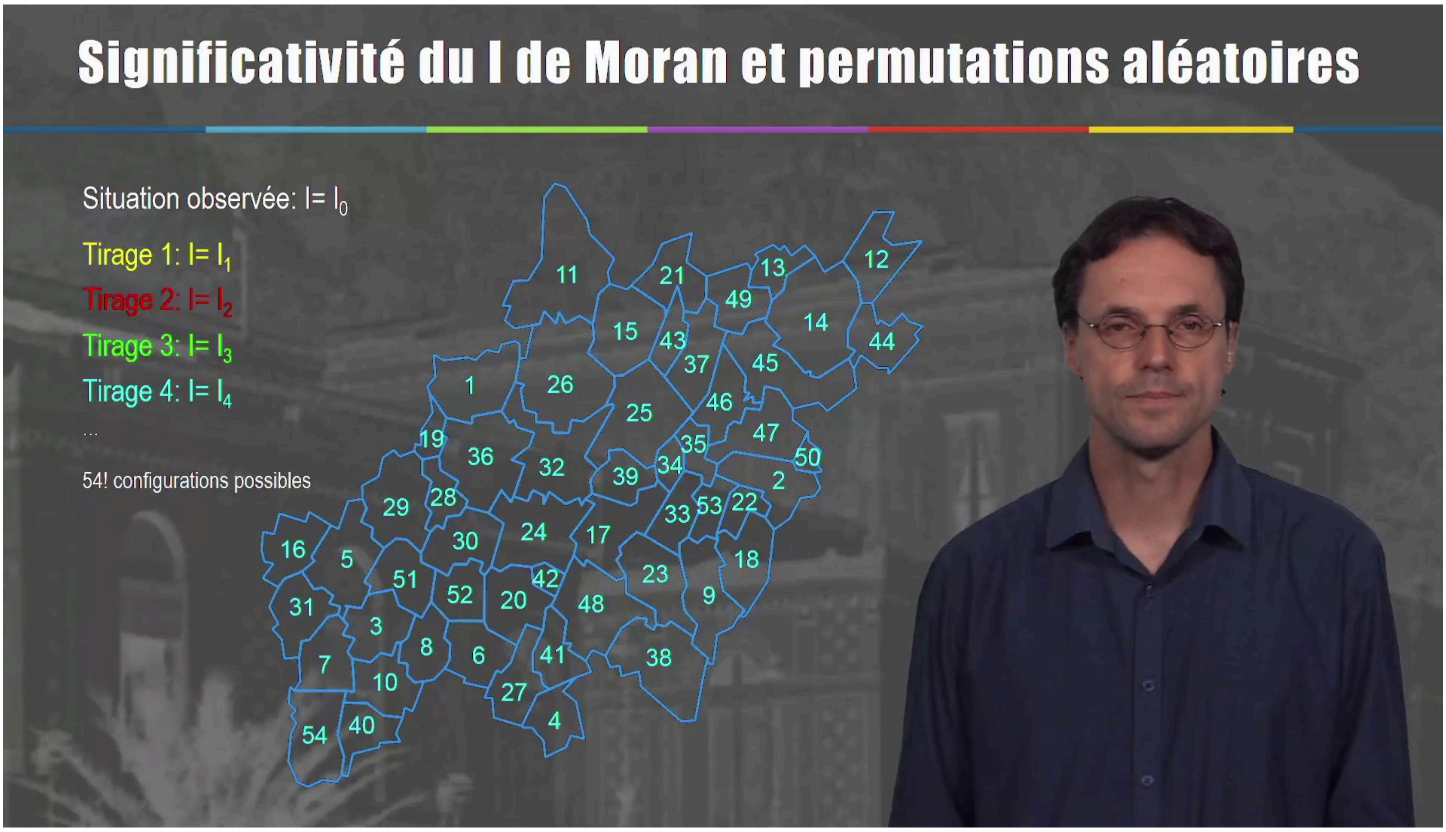
12m 12s

Mais cette valeur d'auto-corrélation spatiale est-elle statistiquement significative? En effet, dans quelle mesure l'agencement spatial de la variable de précipitation n'est-il pas dû au simple hasard?

Notes

Summary

12m 34s

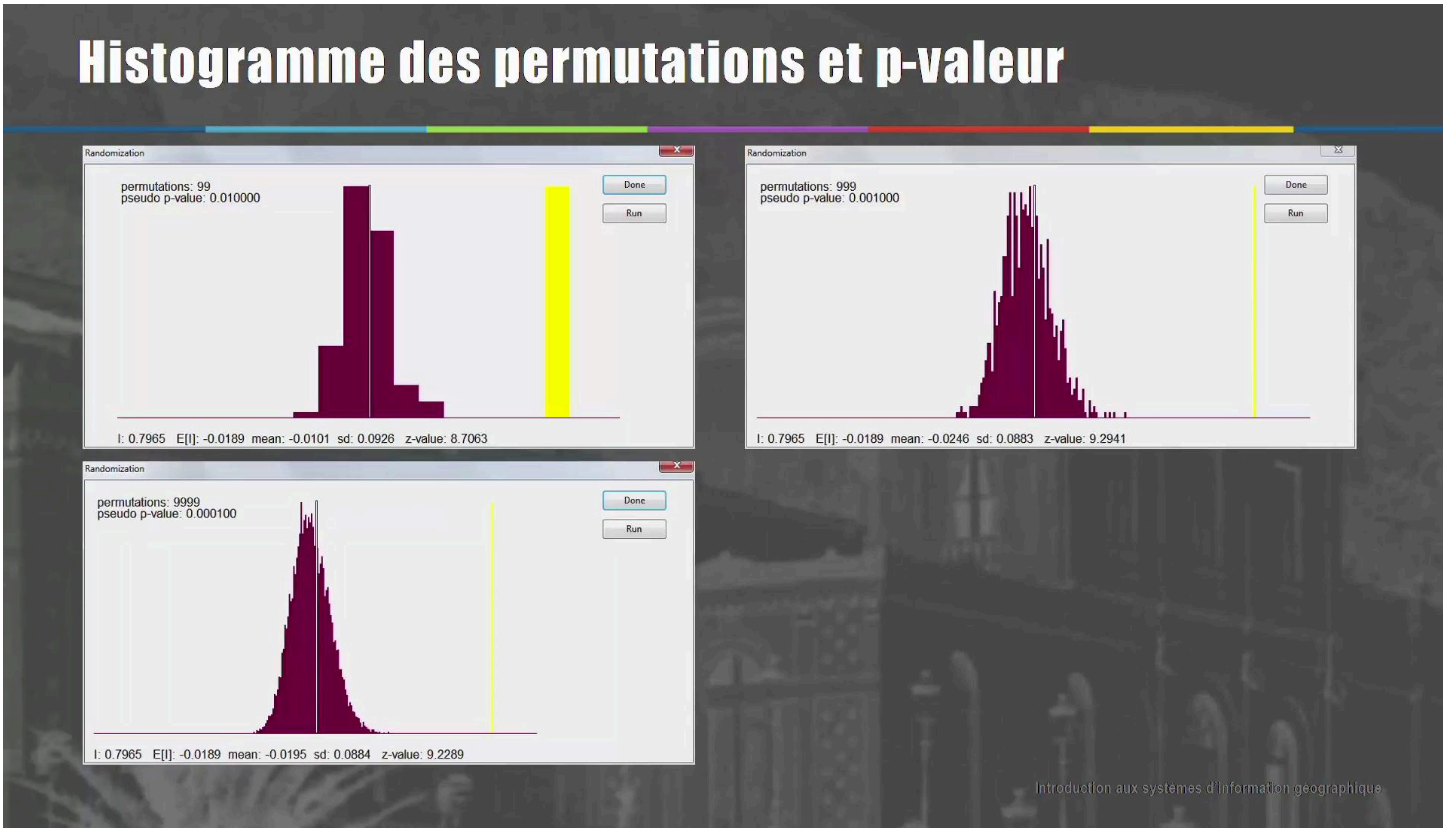


Comment est-ce que la situation observée, soit la structure spatiale révélée par le I de Moran de 0,79, se comporte si on la compare, si ce n'est à toutes les configurations possibles, en tout cas à une grande partie d'entre elles ? Par configuration possible, on pense à permuter aléatoirement les valeurs, on a utilisé ici une variable fictive avec des valeurs comprises entre 1 et 54 pour illustrer la méthode, entre toutes les autres localisations possibles dans le jeu de données analysé. Chaque configuration correspond à un tirage aléatoire. Et on procède jusqu'à plusieurs milliers de permutations dans le cas de la méthode dite de Monte Carlo. A chaque tirage, on calcule le I de Moran que l'on compare avec le I de Moran de la situation observée. Après 999 tirages, on a 999 configurations et autant de valeurs de I de Moran à comparer avec la situation observée, ce qui permet de savoir si cette dernière ressemble aux configurations aléatoires ou si au contraire, elle est clairement différente.

Notes

Summary

12m 47s

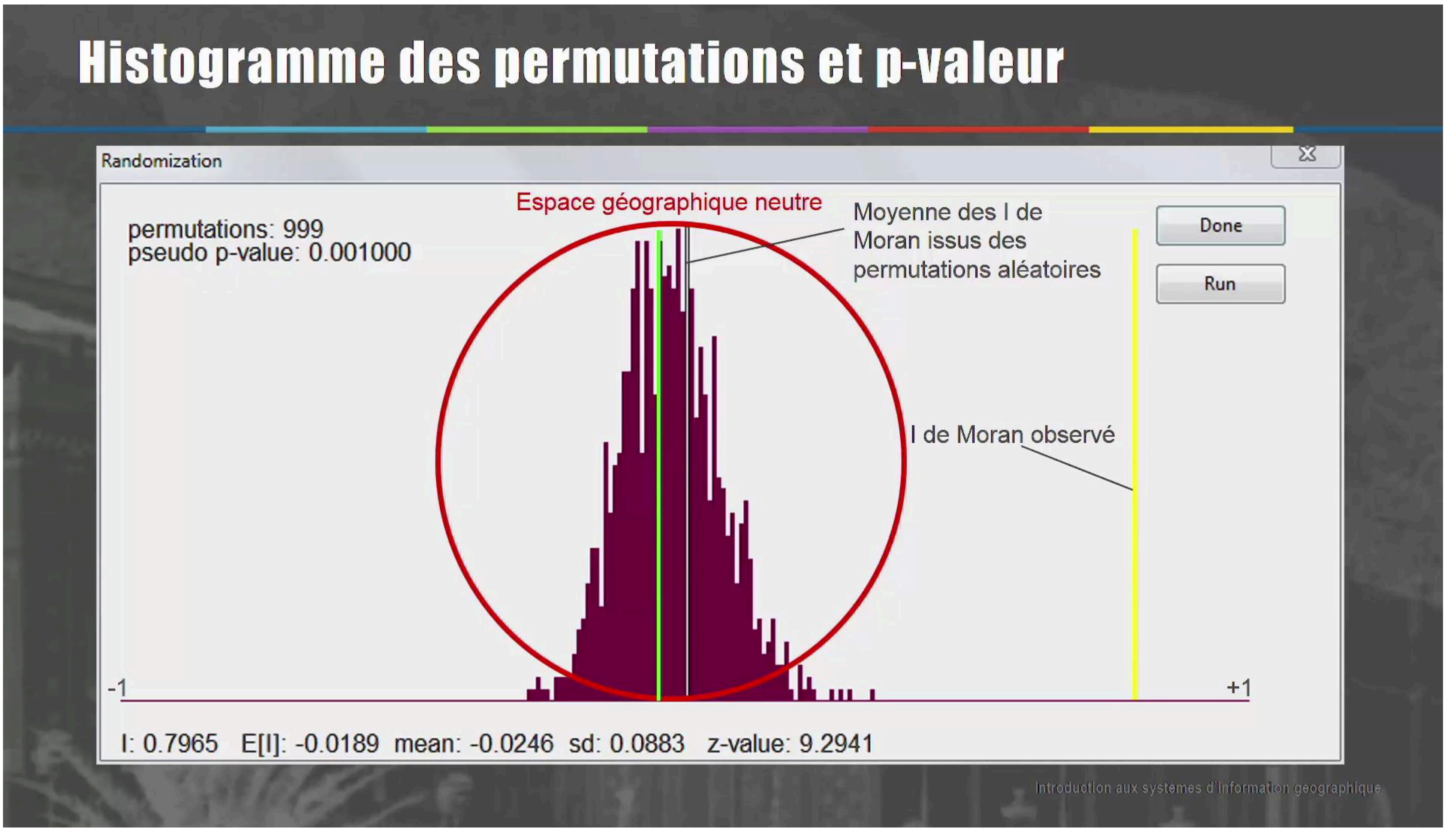


Le logiciel Geoda va stocker les I de Moran correspondants à toutes les configurations aléatoires générées et les utiliser pour construire un histogramme. Sur la base des données réelles de précipitations par commune, voici l'histogramme généré sur la base de 99 tirages, puis de 999, de 9'999 et enfin de 99'999 tirages aléatoires sur un total possible de 54 factorielles.

Notes

Summary

13m 48s

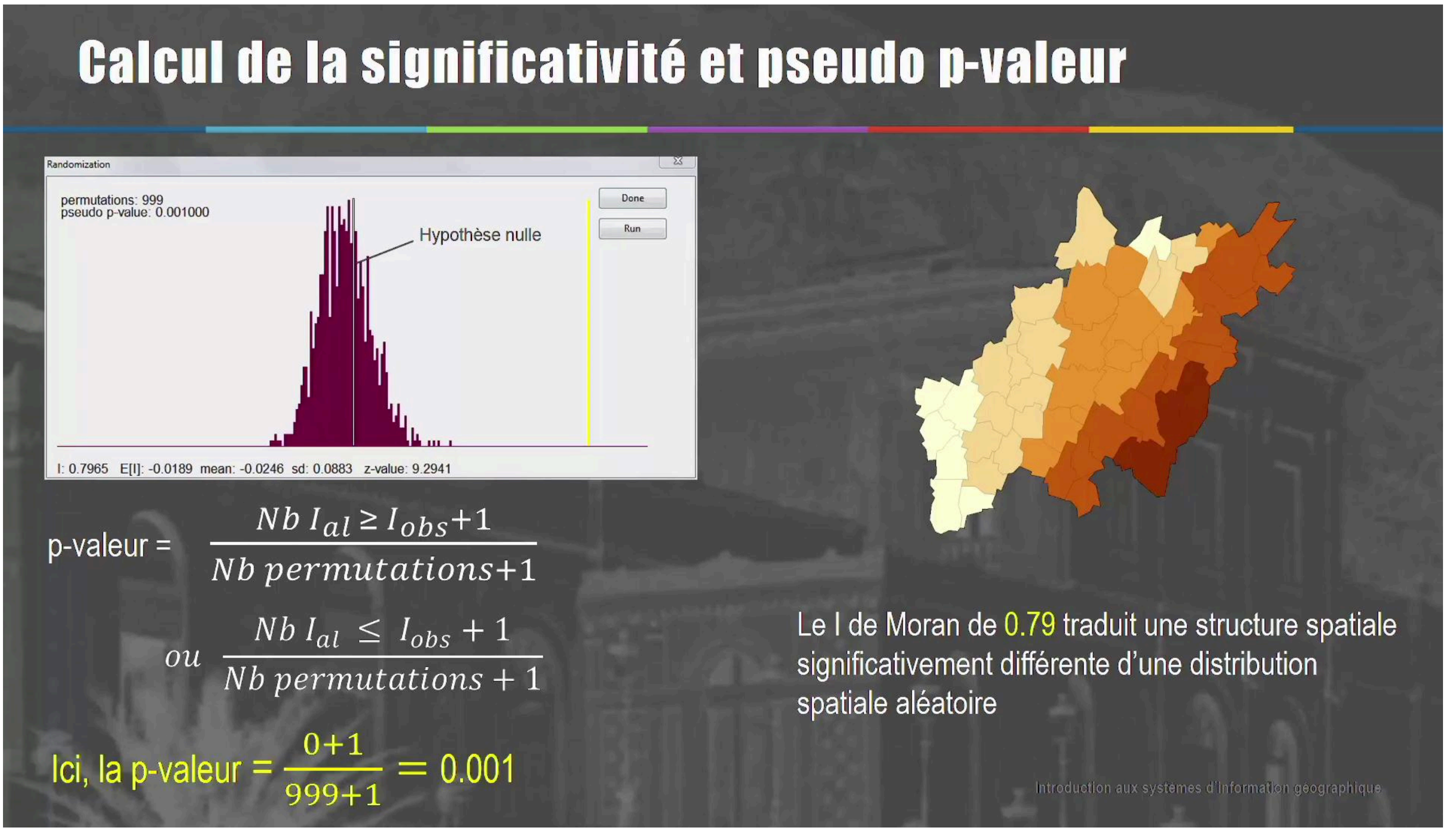


Plus le nombre de tirages est élevé, plus la distribution apparaît normale et plus l'écart-type et la moyenne sont censés s'approcher de leurs valeurs théoriques. Regardons maintenant en détail l'histogramme généré sur la base de 999 permutations. La distribution aléatoire représente ce que l'on peut appeler l'espace neutre. Autour de la moyenne, qui est indiquée ici et qui est proche de 0, les valeurs de précipitations des communes ne ressemblent pas à la moyenne de leurs voisines. Et l'on voit que la valeur du I de Moran pour la situation observée ici indiquée par le trait jaune, et ce trait jaune n'est pas une barre d'histogramme, se distingue clairement du reste de la distribution. Elle est donc apparemment significative. Une valeur non significative apparaîtrait comme ici en vert par exemple, soit au milieu de la distribution neutre, où il n'y a pas de dépendance spatiale.

Notes

Summary

14m 14s

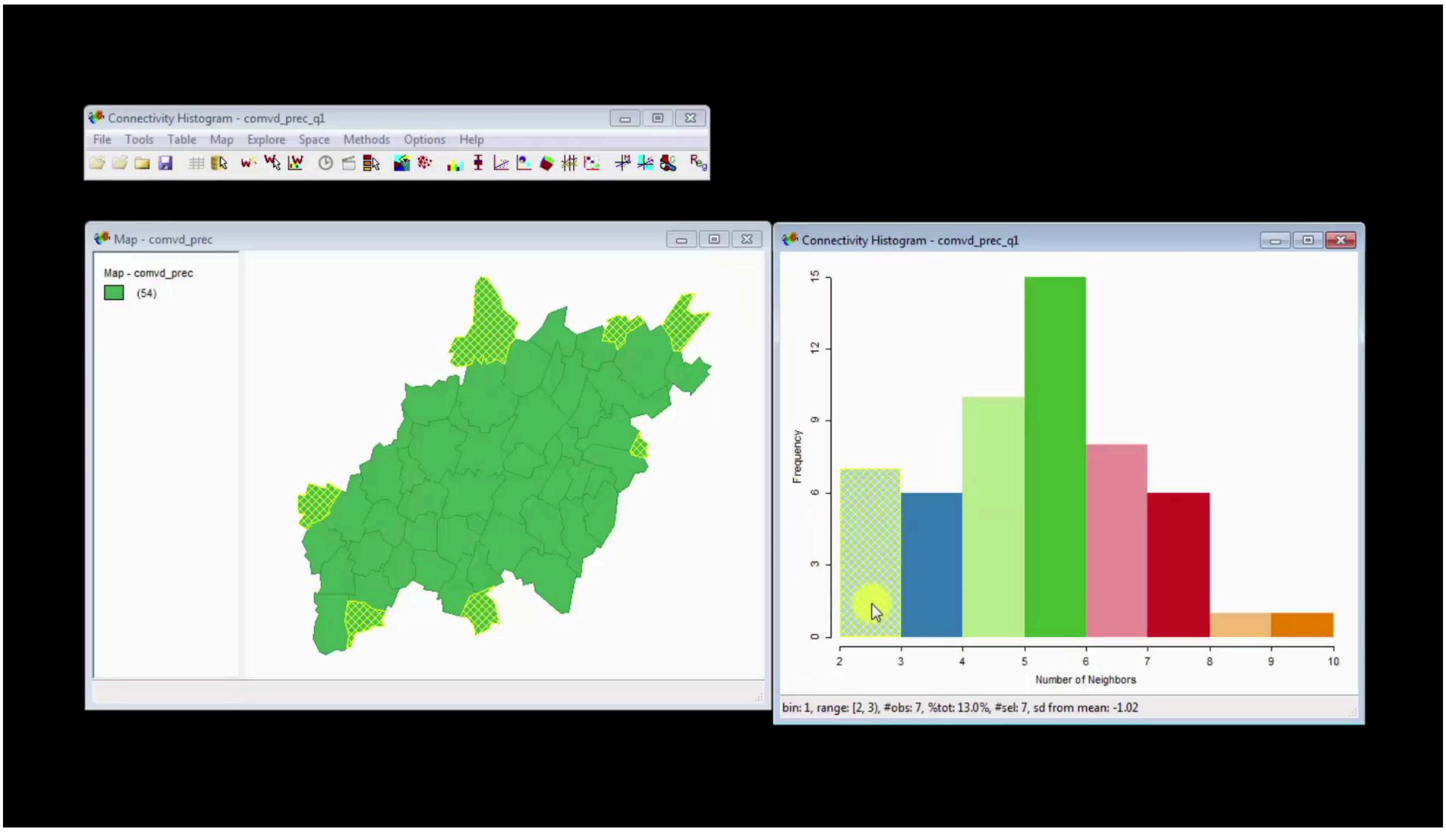


Cette significativité est traduite numériquement par une probabilité de rejeter l'hypothèse nulle. La p-valeur est une valeur limite de rejet de l'hypothèse nulle. Et l'hypothèse nulle ici, c'est que la valeur observée est le fait du hasard et qu'elle ressemble aux autres valeurs générées par permutation aléatoire. Plus la p-valeur est faible, plus le risque de se tromper en rejetant cette hypothèse nulle est faible, donc, plus la valeur observée est significativement différente d'une distribution aléatoire. Cette p-valeur, c'est le rapport entre le nombre de I de Moran générés aléatoirement qui sont plus grands ou égaux que les I de Moran observés plus 1 et le nombre total de permutations aléatoires plus 1. Comme le I de Moran peut également être négatif, la détermination de la p-valeur va dans ce cas se baser sur le nombre de I de Moran générés aléatoirement qui sont plus petits ou égaux que les I de Moran observés. On parle dans le cas de ce test de pseudo p-valeur car le seuil de significativité dépend du nombre de permutations. Dans le cas présent, on peut conclure que la variable de précipitation analysée est spatialement ou globalement auto-corrélée de manière significative.

Notes

Summary

15m 07s

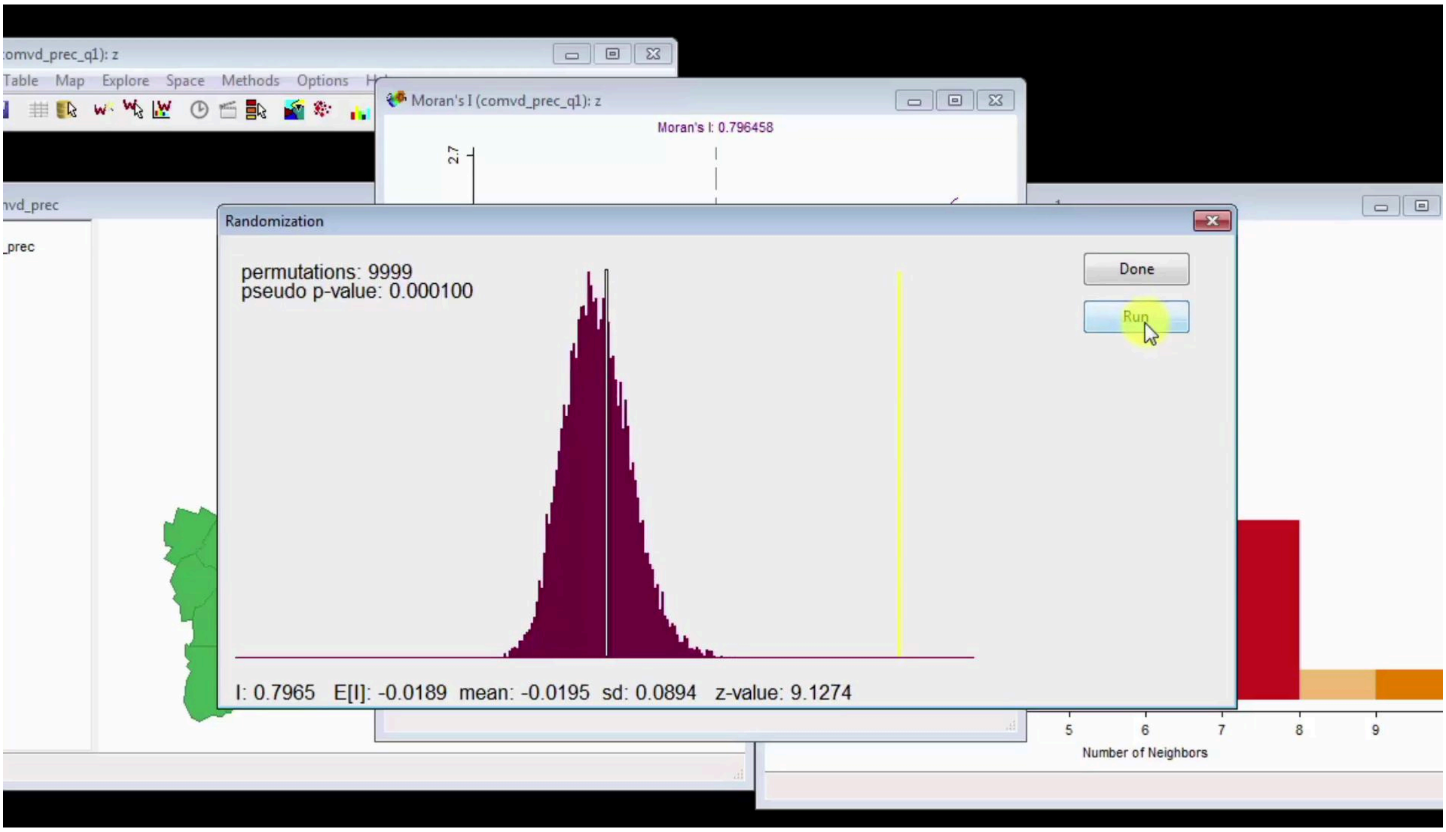


Nous allons voir maintenant comment générer un fichier de pondération spatiale, calculer le I de Moran global et évaluer sa significativité avec GeoDa. Tout d'abord, il faut ouvrir un fichier au format shape qui contient les polygones des communes et leurs attributs dont notre variable de précipitation. Ensuite, il faut créer le fichier de pondération spatiale en cliquant sur le bouton create weights. Et dans la boîte de dialogue correspondante, il faut d'abord sélectionner un identifiant unique, ici IDE, puis sélectionner le schéma de pondération désiré, soit un critère de contiguïté de Queen d'ordre 1. Après avoir cliqué sur le bouton create, il faut choisir l'emplacement où stocker ce fichier de pondération et lui donner un nom qui permette de l'identifier. Ensuite, on va inspecter la connectivité. On va utiliser le bouton connectivity histogram qui nous permet d'inspecter l'histogramme de connectivité, soit de nous donner une indication de la fréquence de polygones ayant le nombre de voisins indiqué en abscisse. Ainsi, il est possible de mettre en évidence la commune la mieux connectée ici ou alors les sept communes les moins bien connectées qui sont situées dans les bords de la zone.

Notes

Summary

16m 31s

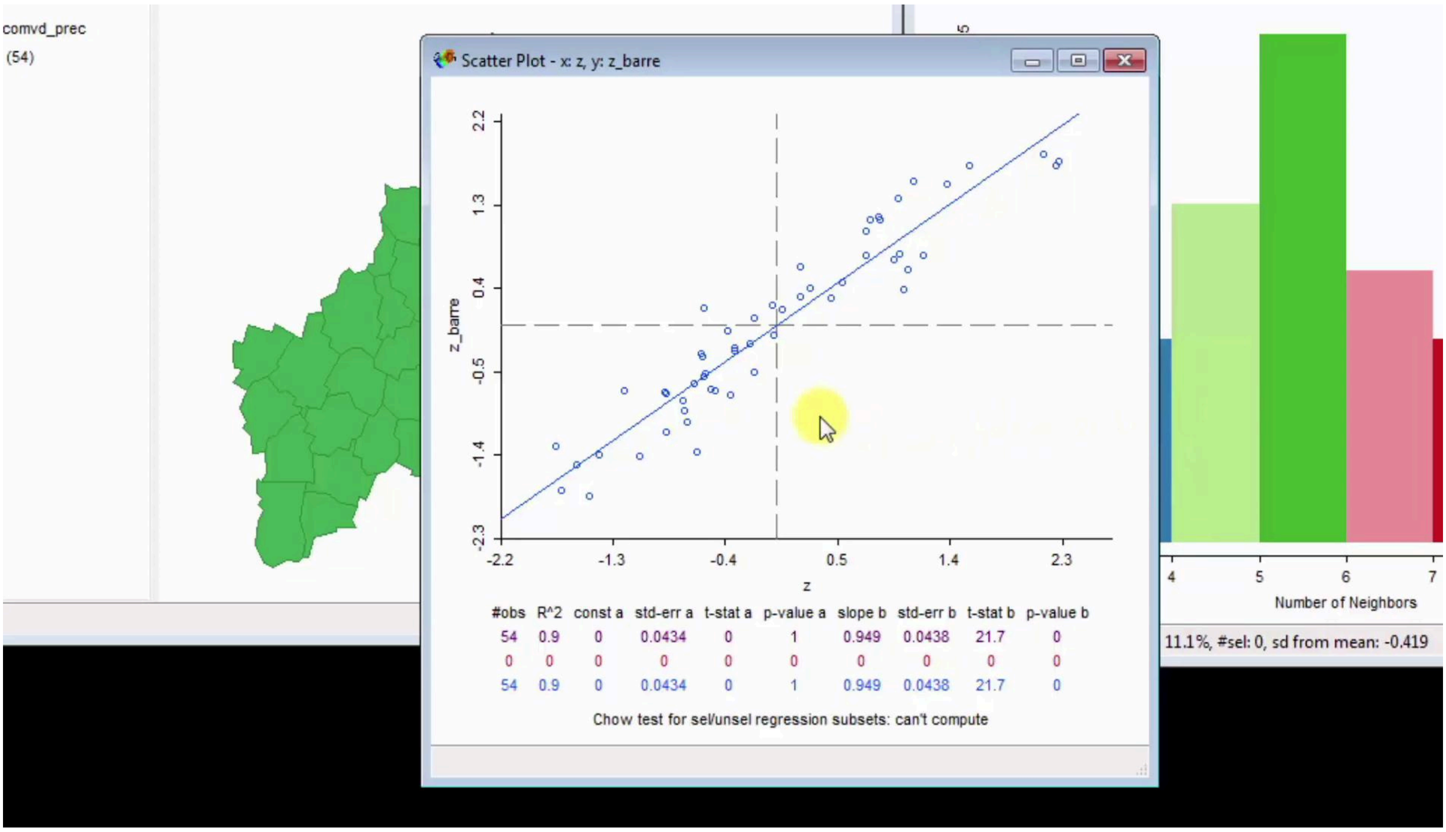


Jetons maintenant un coup d'œil dans le fichier de pondération. Alors ce fichier de pondération, on peut l'ouvrir avec un éditeur de texte et on retrouve la structure présentée tout à l'heure avec l'en-tête sur la première ligne puis la description du voisinage du polygone 1 qui compte 6 voisins, puis du polygone 2 qui en compte 5, et du polygone 3 qui en compte 5 aussi, etc. Nous allons maintenant lancer le calcul du I de Moran proprement dit. Il faut sélectionner la variable d'intérêt pour commencer et il s'agit de Z qui désigne les précipitations mensuelles. Le scatter plot de Moran est généré immédiatement puisqu'on avait déjà généré le fichier de pondération tout à l'heure et il était chargé. Il nous montre le nuage de valeurs centrées et réduites et en haut un I de Moran de 0,79. Ensuite, il est rapidement possible de tester la significativité en faisant un clic droit sur le scatter plot et en cliquant sur randomisation. Et nous allons générer 9'999 configurations spatiales aléatoires. L'histogramme correspondant nous montre ici une pseudo p-valeur très petite qui signifie qu'il est peu probable de se tromper en rejetant l'hypothèse nulle et donc, que la structure spatiale observée et caractérisée par un I de Moran de 0,79 est significativement différente d'une distribution spatiale aléatoire.

Notes

Summary

17m 57s



Chaque fois que l'on clique sur le bouton run, on régénère une série de 9'999 permutations aléatoires avec des statistiques différentes. Mais on s'aperçoit que la pseudo p-valeur reste très petite. Pour terminer, nous allons calculer le I de Moran une nouvelle fois mais selon l'interprétation de Luc Anselin, soit de construire une régression linéaire entre la variable de précipitations Z en abscisse et la variable de précipitations pondérée Z barre en ordonnée. Et on constate qu'effectivement, le coefficient bêta qui donne la pente de la droite de régression est égal à 0,79 et qu'en standardisant les 2 distributions de précipitations exprimées en dixième de millimètres d'eau, on tombe sur le scatter plot de Moran obtenu directement tout à l'heure.

Notes

Summary

19m 29s