INSTITUT DES ACTUAIRES

ISUP

Mémoire présenté le :
pour l'obtention du diplôme
de Statisticien Mention Actuariat
et l'admission à l'Institut des Actuaires

Par : Malcolm-Chris BIKOUMOU

Sujet : Analyse spatiale d'un traité de réassurance Multi-Péril sur récolte

Confidentialité : ☐ NON ☒ OUI (Durée : ☐ 1 an ☒ 2 ans)

Les signataires s'engagent à respecter la confidentialité indiquée ci-dessus.

Membre présents du jury de l'Institut des Actuaires

.....

Membres présents du jury de la filière

Entreprise : Liberty Specialty Markets

Nom :

Signature :

Directeur de mémoire en entreprise :

Nom : Jean-Christophe GARAIX

Signature :

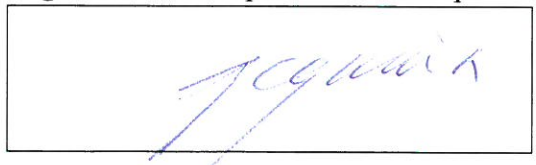
Invité :

Nom :

Signature :

Autorisation de publication et de mise en ligne sur un site de diffusion de documents actuariels *(après expiration de l'éventuel délai de confidentialité)*

Signature du responsable entreprise



Signature du candidat

Secrétariat

Bibliothèque :

Remerciements

Je remercie Jean-Christophe Garaix pour ses conseils prodigués tout au long de mon stage, Abdessamad El Angoudi pour sa disponibilité vis-à-vis de mon sujet de mémoire, Salah Dhouib, Hervé Grenier, Françoise Bollotte, Inès Cloarec pour leur sympathie tout au long de mon stage.

Résumé

En France, le potentiel de sinistres sur les récoltes agricoles dus aux risques climatiques est extrêmement élevé. La sécheresse de l'été 2011 l'a prouvé en provoquant des dommages à hauteur de 242 millions d'euros sur le territoire.

Ce risque de sécheresse, combiné au risque de tempête, de grêle, de gel et d'inondation est couvert par les assurances Multi-Périls sur récoltes. Ce risque Multi-Péril est particulièrement difficile à modéliser car il repose sur plusieurs risques non linéaires. Néanmoins, cette modélisation s'avère nécessaire pour l'assureur et le réassureur pour définir et tarifer la couverture de ce type de risque.

L'objet de ce mémoire est, du point de vue du réassureur, d'étudier la distribution spatiale de ce risque. Nous appréhenderons ce risque en quatre étapes. :

- Tout d'abord, nous définirons les principes de base de la réassurance et de l'assurance multi-périls sur récolte.

- Dans un second temps, nous nous intéresserons à la corrélation spatiale globale entre toutes les communes du portefeuille.

- Ensuite, nous découperons le territoire en zones géographiques homogènes, ce qui nous permettra d'étudier les corrélations spatiales des communes au niveau local.

- Enfin, nous effectuerons un lissage spatial des loss-cost nous permettant de mettre en exergue le risque géographique.

Mots-clés : Assurance Multi-Péril sur Récoltes, Analyse exploratoire spatiale, Indice de Moran, Indice de Geary, Classification ascendante hiérarchique, Regression Locale, autocorrélations spatiales

Abstract

The potential of loss on yields of farm producers in extremely high in France. The drought in summer 2011 proved this fact, by damaging french territory up to €242 billion.

Perils like drought, combined to hail, frost, flood and wind are covered by Multiple Peril Crop Insurance. This Multiple Peril risk is particulary hard to model because it relies on many non-linear risks. Nonetheless, insurers ans reinsurers need to be able to model and price this type of risk.

This paper aims at studying the spatial distribution of this risk.

- The first part is a reminder about the fundamentals of reinsurance and multiple peril crop insurance.

- The second part deals with the spatial autocorrelation of the porfolio

- In the third part, we define homogeneous regions in which we can study the local autocorrelation

- In the fourth part, we use a spatial smoothing on the loss costs in order to highlight the geographic risks.

Keywords : Multiple Peril Crop Insurance, Spatial Analysis, Moran's I, Geary's C, Hierarchical Clustering, Locally Weighted Regression, Spatial Autocorrelation

Table des matières

I.1. Rappels des principes de réassurance....8
1.Les rôles pour la réassurance....8
2.Les formes juridiques de réassurance....9
3.Les formes techniques de la réassurance....10
I.2 Périls liés aux risques de catastrophes naturelles en France....13
1.Définition de catastrophe naturelle....13
2.Assurance Multi-Périls sur Récoltes....15
3.Problématique de l'étude....16
4.Présentation des données utilisées....17
II.1 Généralités sur la classification des données....20
1.Notion de distance....20
2.Méthode des k-means....21
3. Classification hiérarchique....22
4.Critères d'agrégation....24
II.2 Analyse exploratoire spatiale des données....26
1.Aléa spatial....26
2.Autocorrélation spatiale : Indices de Moran et Geary....26
3.Graphique de dispersion de Moran....28
4.Autocorrélation spatiale locale....29
II.3 Mise en œuvre de la méthode....29
1.Autocorrélation spatiale du portefeuille dans son ensemble....30
2.Test de significativité de la composante spatiale des indices....36
3.Classification ascendante hiérarchique du portefeuille....37
4.Autocorrélations spatiales locales des régions....43
III.1 Présentation du modèle de régression locale....49
1.Généralités théoriques....49
2.Régression locale multivariée....50
III.2 Application du modèle de régression LOESS....51
1.Préparation des données....51
2.Paramètres de la fonction LOESS....52
3.Mise en œuvre de la procédure....52
4.Validation des hypothèses....55
Conclusion……..57
Bibliographie……..58
Annexes……..60
A.Inerties relatives pour la classification ascendante hiérarchique avec les méthodes de Ward et du diamètre....61
B.Carte des climats de la France....63
C.Cartes des LISA pour les régions 1,2,3 et 5....64
D.Carte de significativité des LISA pour les régions 1,2,3 et 5....68

E.Lissages spatiaux des régions 1,2,3 et 5.....72
F.Loss-costs moyens pour les régions 1,2,3 et 5.....76

Partie I
Cadre de l'étude

I.1. Rappels des principes de réassurance

1.Les rôles pour la réassurance

L'article L310-1-1 du Code des Assurances définit la réassurance comme étant « une opération par laquelle une société d'assurance, ou cédante, s'assure elle même auprès d'une autre société dénommée réassureur, ou cessionnaire, pour tout ou partie des risques qu'elle a pris en charge. »

D'abord, la réassurance peut être décrite comme un transfert de risque. Par exemple, lorsqu'un assureur est amené à assurer un risque mal maîtrisé, notamment lors de la tarification d'une nouvelle garantie ou lorsqu'il cherche à écrêter ses risques de pointe. Diviser et partager les risques avec un réassureur permet de réduire le risque de ruine inhérent aux risques cités précédemment. Ces mesures permettent notamment d'augmenter la marge de solvabilité de l'assureur. En France, la Commission de Contrôle des Assurances prend en compte les cessions de réassurance dans le calcul de la solvabilité d'un assureur.

La réassurance permet aussi d'apporter un support financier à l'assureur et d'assurer son développement. Elle permet en effet d'augmenter la rentabilité de ses fonds propres, mais aussi d'augmenter sa capacité de souscription.

Le troisième rôle de la réassurance est d'offrir un support technique auprès des cédantes, en les aidant par exemple à définir leurs besoins de réassurance. Un réassureur peut aussi servir de support dans la tarification de garanties particulières ou de risques rares, ou apporter de l'assistance dans la gestion des grands sinistres.

Enfin un réassureur peut apporter sa compétence à une cédante en matière de formation, quelle soit technique ou informatique.

2.Les formes juridiques de réassurance

Il existe trois formes de réassurance liant juridiquement l'assureur et la cédante.

La réassurance facultative est historiquement la plus ancienne. Elle permet la liberté de décision pour l'assureur de céder ou non ainsi que la liberté d'acceptation ou non pour le réassureur. Elle concerne des risques précis couverts par une police d'assurance spécifique unique. En effet, la possibilité pour le réassureur d'estimer chaque risque séparément permet de tarifer le contrat de manière plus fine par rapport au risques encourus. Elle permet notamment à l'assureur d'avoir une capacité plus grande.

La réassurance obligatoire définit les obligations réciproques de l'assureur et du réassureur au travers d'un traité de réassurance. L'assureur à l'engagement de céder une famille prédéfinie de risques et le réassureur a l'obligation d'accepter tous ces risques.

La réassurance facultative-obligatoire est très peu pratiquée car elle place le réassureur dans un risque d'anti-sélection. En effet, le réassureur a l'obligation d'accepter tous les risques cédés par l'assureur. Ce dernier peut donc sélectionner les risques les plus favorables à conserver et céder au réassureur les risques les plus défavorables.

3.Les formes techniques de la réassurance

Il existe deux formes de réassurance.

La réassurance proportionnelle est basée sur la partage proportionnel de risque entre d'assureur et le réassureur. Il s'agit d'une répartition de la somme assurée, de la prime et des sinistres. Elle se décline en deux modes, le quote-part et l'excédent de plein.

Dans un **quote-part**, l'assureur cède au réassureur un pourcentage fixe des risques qu'il souscrit. Le réassureur prendra le même pourcentage de sinistres à sa charge et recevra le même pourcentage de primes. Pour l'assureur, le quote-part est très simple à gérer, ce qui est un avantage. Mais pour cela, il doit céder un montant important de son portefeuille, y compris des risques mineurs qu'il aurait pu conserver. De plus, ce type de contrat ne nivelle pas les risques. En conséquence, un mauvais résultat réassuré restera mauvais. C'est la raison pour laquelle ce type de traité n'est plus populaire et tend à être de moins en moins fréquent. Plus l'assureur conserve le risque , plus il montre qu'il a confiance dans le risque qu'il a souscrit. Ainsi, le réassureur sera plus à même de lui apporter son soutien car il sait que le partage du risque sera équitable.

Dans un **excédent de plein**, l'assureur conserve une somme assurée maximum sur chaque risque, appelée plein de conservation et cède la partie qui dépasse ce plein au réassureur, et ce, pour chaque risque.

Appelons :

- K le capital associé au risque i
- C la capacité de l'excédent de plein
- R le plein de conservation
- t_i le taux de cession

Alors : $t_i = Min\left(Max\left(\frac{K_i - R}{K_i}; 0 \right); \frac{C}{K_i} \right)$

Du point de vue de l'assureur, l'excédent de plein a de nombreux avantages. Il permet en effet de niveler ses risques, d'augmenter sa capacité, de conserver les petits et moyens risques. La gestion s'effectue risques par risques. Cependant, il y a un déséquilibre avec le réassureur car il conserve les risques les plus dangereux. L'exposition au risque est bien plus importante qu'un quote-part, pour le réassureur.

Dans la **réassurance non proportionnelle**, la répartition entre l'assureur et le réassureur se fait uniquement sur les sinistres. Le réassureur va couvrir à concurrence d'un certain montant (la portée ou la capacité) l'excédent de sinistre qui dépasse priorité correspondant au montant de sinistre conservé par l'assureur.

Pour un traité en **excédent de sinistre**, l'assureur détermine le montant maximum qu'il peut conserver sur un seul sinistre et le réassureur prend en charge l'excédent de ce sinistre.

Appelons :

- a la priorité de l'assureur
- b la portée du réassureur
- X_i le i ème sinistre de l'assureur

Alors le montant global à charge du réassureur est $\sum_{X_i=1}^{N} Min\left(Max\left(X_i - a\,;0\right);b\right)$

Du point de vue du réassureur, l'excédent de sinistres présente des difficultés. En effet, il doit disposer de statistiques fiables pour pouvoir coter ce type de traité. Et de plus, en cas de sinistres, la prime qu'il reçoit est très faible. On distinguera l'excédent de sinistre par risque et l'excédent de sinistre par événement.

Dans un excédent de sinistre par risque, le réassureur s'engage à payer un montant inférieur ou égale à la portée à chaque fois qu'une police de l'assureur est sinistrée.

Dans un excédent de sinistre par événement, le sinistre correspond à une cause impactant plusieurs polices d'une même branche.

Pour un traité en **excédent de pertes**, les sinistres sont agrégées sur l'année. La priorité de l'assureur est déterminée par le stop-loss, c'est à dire le rapport sinistres/primes.

Appelons :

- a la priorité de l'assureur
- b la portée du réassureur
- P la prime de l'assureur
- $\left(X_i\right)_{i=1..N}$ les sinistres de la cédante durant l'année
- S les sinistres à la charge du réassureur

Alors $$S = Min\left(Max\left(\frac{\sum_{X_i=1}^{N} X_i}{P} - a; 0\right); b\right) \quad P$$

Un traité en excédent de pertes peut également comporter un montant maximum de sinistres à partir duquel celui-ci n'interviendra plus, ou un montant minimum de primes encaissées en deçà duquel le réassureur n'interviendra pas.

I.2 Périls liés aux risques de catastrophes naturelles en France

1.Définition de catastrophe naturelle

La loi du 13 juillet 1982 définit les catastrophes naturelles comme étant les dommages matériels directs ayant eu pour cause déterminante l'intensité anormale d'un agent naturel lorsque les mesures habituelles à prendre pour prévenir ces dommages n'ont pu empêcher leur survenance ou n'ont pu être prises. Ces catastrophes peuvent avoir des origines climatiques (Grêle, Tempête, Sécheresse, Tornade), des origines géologiques (Tremblement de Terre, Glissement de terrain, Eruptions volcaniques) ou avoir d'autres causes (Incendie, Inondation). Du point de vue du réassureur, le risque catastrophe est un risque particulier. En effet, c'est un risque de fréquence faible mais qui peut impacter sur l'ensemble du portefeuille.

Figure 1: Principaux événements naturels entre 1988 et 2011

Principaux événements naturels depuis 1988 (Cat-Nat + TGN)

(valeurs mai 2013)

Régime d'assurance	Exercice de survenance	Désignation de l'événement	Coût marché national en M€	Coût marché national en M€ constants * (indice FFB)
Cat-Nat	2011	Subsidence	675	693
Cat-Nat	2011	Innondations du Sud/novembre	215	221
Tempête	2011	Tempête Joachim/décembre	180	185
Cat-Nat	2004-2010	Subsidence	790	959
Cat-Nat	2010	Inondations-Tempête Xynthia/février	745	800
Tempête	2010	Tempête Xynthia/février	735	790
Cat-Nat	2010	Inondations du Var/juin	615	661
Tempête	2009	Tempête Klaus/janvier	1 680	1 872
Tempête	2009	Tempête Quinten/février	200	223
Cat-Nat	2008	Inondations du Centre-Est/novembre	160	178
Cat-Nat	2007	Cyclone Dean Guadeloupe/août	190	222
Cat-Nat	2003	Subsidence	1 420	1 996
Cat-Nat	2003	Inondations du Rhône/décembre	670	942
Cat-Nat	1989 à 2002	Subsidence	2 820	5 041
Cat-Nat	2002	Inondations du Gard/septembre	700	1 021
Tempête	1999	Tempête Lothar et Martin/février	6 860	11 010
Cat-Nat	1999	Inondations de l'Aude/novembre	290	465
Cat-Nat	1999	Inondations-Tempête Lothar et Martin/fév.	170	273
Cat-Nat	1995	Inondations du Nord/janvier-février	350	604
Cat-Nat	1993-1994	Inondations du Nord/décembre-février	110	203
Cat-Nat	1993	Crues du Rhône/octobre	190	351
Cat-Nat	1993	Inondations Bouches du Rhône/sept.	130	240
Cat-Nat	1992	Inondations Vaison la Romaine/sept.	240	462
Tempête	1990	Tempête Daria/février	1 315	2 717
Cat-Nat	1990	Inondations de Besançon/février	150	310
Cat-Nat	1988	Inondations Nîmes/octobre	288	637

* Coût en euros constant revalorisés par l'indice FFB à la fin 2012

Source : FFSA - GEMA et CCR

41

Comme on peut le voir, la France est touchée régulièrement par des catastrophes naturelles dévastatrices. En particulier, la sécheresse , qui n'est pas représentée dans ce tableau, est un risque naturel extrêmement craint, notamment dans le secteur agricole. L'été 2011 nous a bien rappelé à quel point ce risque n'était pas à négliger. En effet, près d'une culture céréalière sur deux a fait l'objet d'une déclaration de sinistre sécheresse.

2.Assurance Multi-Périls sur Récoltes

Les **Assurances Multi-périls sur Récoltes ou Assurance Aléa Climatique** sont apparues en 2005. C'est en effet à partir de cette date que les assureurs ont obtenu les statistiques nécessaires pour appréhender ce type de nouveau risque. Avant cette date, seule la grêle était couverte par les assureurs. Ces assurances ont pour but de protéger et d'assurer les rendements des agriculteurs pendant contre des risques climatiques variés. Les cinq principaux aléas climatiques pris en compte par ce contrat sont la grêle, la tempête, la sécheresse, le gel et les inondations. Malgré les aides gouvernementales pour aider les agriculteurs à souscrire ce type de contrat (l'État prenait en charge 35 % des cotisations des agriculteurs), le marché a eu du mal à se lancer. En 2007 le marché a quasiment stagné en nombre de contrats tout en progressant significativement en termes de primes, du fait de l'augmentation des prix des céréales et oléagineux, mais moins vite en termes de surfaces couvertes. En 2009, les surfaces assurées par ce type de contrat ont baissé, puis ont repris une lente progression jusqu'en 2013. Pourtant, et on l'a vu avec les exemples précédents, le marché agricole est extrêmement exposé aux risques climatiques. En effet, les primes de réassurance agricole en France sont évaluées par l'APREF (L'Association des Professionnels de la Réassurance En France) à environ 50 millions d'euros pour 2013, par rapport à des engagements estimés à 800 millions d'euros.

En 2012, à peine 35 % des cultures du territoire étaient assurés contre l'aléa climatique. Le marché est donc encore extrêmement compétitif pour les réassureurs. Il est d'autant plus compétitif qu'aujourd'hui encore, les statistiques fournies par les assureurs ne sont pas optimales. En effet, un système de collectes de données concernant les données des agriculteurs (surface cultivée, production récoltée, …) est à l'étude mais n'est pas encore au point. Il est ainsi encore délicat, pour les assureurs et les réassureurs d'évaluer les engagements maximums potentiels pris dans ce type de contrat.

3.Problématique de l'étude

Ce manque relatif de données incitent les assureurs et les réassureurs à regorger d'inventivité quant à la tarification de ces contrats. L'objectif de cette étude est d'étudier spatialement un contrat multi-périls sur récoltes par commune de type excédent de sinistres d'une priorité de 100 000 € et d'une portée de 1 900 000 €.

On peut déjà remarquer que du fait que les récoltes agricoles soient des biens très facilement préjudiciables, la priorité de 100 000€, qui est donc assez élevé est une forme de protection contre le risque intéressante pour le réassureur.

La méthode utilisée va être de considérer la localisation spatiale des communes et des sinistres. En effet, des zones géographiques homogènes, c'est-à-dire des zones dans lesquelles les paramètres de sols, de climat et de pratiques culturales sont similaires ont plus de chances d'être atteintes par les mêmes types de sinistres climatiques que les autres. Par exemple, les régions cultivant des vignes seront plus atteintes par le gel, dû à la faible résistance des vignes à ce type de condition climatique.

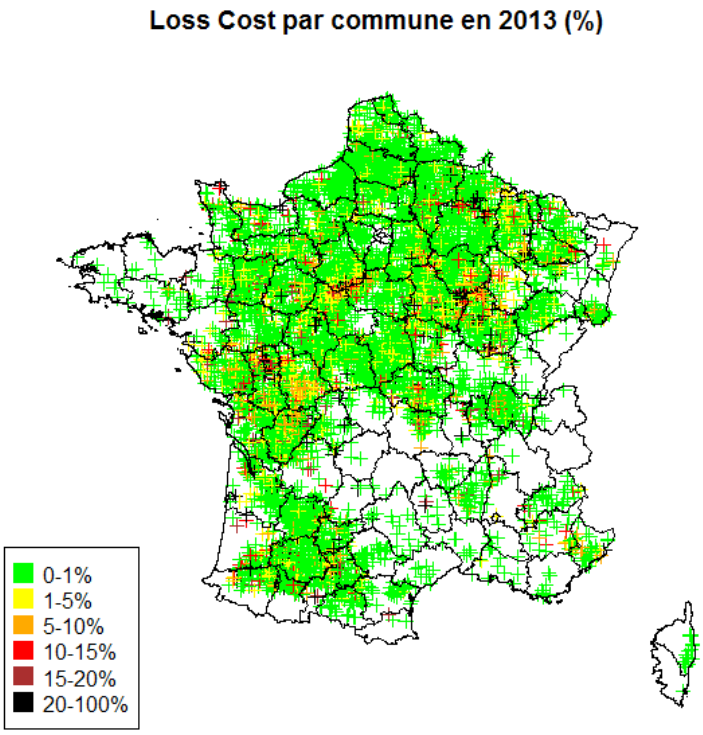
Les méthodes de tarification par expérience classique (modèles fréquence/sévérité) ne prennent pas en compte la concentration des expositions dans des surfaces géographiques limitées. On peut s'attendre à ce que les conditions météorologiques ou le type de sol soient similaires pour des expositions géographiquement proches. Assurer des expositions qui sont similaires et hautement corrélées spatialement avec d'autres expositions localement proches ne résulterait pas forcément en une baisse significative du risque. Le but est de savoir quelle est l'influence de la géographie sur les pertes attendues.

4.Présentation des données utilisées

Liberty Specialty Markets réassure l'Etoile, une société d'assurance mutuelle spécialisée dans l'assurance agricole. Pour mener l'étude, nous utiliserons un historique de sinistres fournies par l'Etoile sur 14 ans (de 2000 à 2013) pour environ 7000 communes. Nous avons les variables suivantes pour chaque sinistre :

- Année du sinistre
- Le numéro de département de la commune sinistrée
- La numéro de commune de la commune sinistrée
- La surface cultivée
- La somme assurée
- La cotisation
- Le montant des indemnités

Figure 2: Carte des Loss Cost en France en 2013



On peut d'emblée s'apercevoir de zones de concentration de sinistres, ce qui corroborerait notre hypothèse : La proximité géographique a une influence sur l'exposition aux risques climatiques. L'idée est de profiter de cette corrélation entre les loss cost des communes proches pour tarifer ce contrat. En effet, s'il y a corrélation entre les Loss Cost des communes alors il est possible d'utiliser l'information contenue dans le Loss Cost d'une commune pour estimer le Loss Cost des communes alentours, en supposant que ces communes sont dans la même zone géographique homogène.

Avant donc de pouvoir utiliser des méthodes permettant de prendre avantage de ces corrélations entre les communes, il faut reconstituer des zones géographiques homogènes. C'est l'objet du prochain chapitre.

Partie II
Classification des régions

Dans l'hypothèse que les loss-cost des communes dépendent des loss cost des communes environnantes ou plus généralement de la zone géographique alentour, l'idée est de modéliser un loss-cost donné en fonction de l'information fournie par les communes environnantes, dans l'hypothèse que les communes environnantes sont dans la même zone géographique homogène. Malheureusement, nous ne possédons pas d'informations concernant la géographie du territoire, ou encore le type de culture pour chaque commune. Il s'agit donc d'utiliser des outils statistiques permettant de créer artificiellement des zones géographiques homogènes. Des méthodes d'analyse spatiale ainsi que de classification permettent de créer et de décrire ce type de régions en utilisant les données cités dans la première partie.

II.1 Généralités sur la classification des données

Le but des méthodes de classification est de construire une partition d'un ensemble d'objets dont on connaît les distances deux à deux. Les classes formées doivent être le plus homogène possible. En classification, toute l'information utile est contenue dans un tableau $n \times n$ donnant les dissemblances ou la distance entre les n individus à classer.

1.Notion de distance

Notons E l'ensemble des n objets à classer. Une distance est une application de $E \times E$ dans R_+ telle que :

$$d(i;j) = d(j;i)$$
$$d(i;j) \quad 0$$
$$d(i;j) = 0 \quad i = j$$
$$d(i;j) \leq d(i;k) + d(k;j)$$

Lorsqu'on a seulement

$$d(i;j) = d(j;i)$$
$$d(i;j) \geq 0$$
$$d(i;i) = 0$$

On parle de dissimilarité. Une similarité est une application s telle que :

$$s(i;j)=s(j;i)$$
$$s(i;j)\geq 0$$
$$s(i;i)\geq s(i;j)$$

Lorsque les données se présentent sous la forme d'un tableau X de p caractères numériques, on utilise souvent la métrique euclidienne classique $M=I$ ou $M=D_{1/s^2}$.

On peut aussi utiliser la métrique de Mahalanobis $M=V^{-1}$, la distance L_1 : $d(i;j)=\sum_k |x_i^k - x_j^k|$ ou encore la distance de Minkowski L_q : $d(i;j)=\left(\sum_k \left(x_i^j - x_i^k\right)^q\right)^{1/q}$

Il existe beaucoup d'autres indices de dissimilarité.

2.Méthode des k-means

La méthode de nuée dynamique, ou k-means, permet de traiter rapidement des ensembles d'effectif assez élevé en optimisant localement un critère d'inertie. On supposera que les individus sont des points de R^p muni de la distance euclidienne.

Étant donné une partition en k groupes d'un nuage de n points, on définit les centres de gravité $g_{1,\dots},g_k$ des k groupes et $I_1,_{\dots},I_k$ les inerties des k groupes. L'inertie est la moyenne des carrés des distances au centre de gravité.

L'inertie totale des n points autour du centre de gravité global g est égale à :

$$I=I_B+I_w$$

où I_w est l'inertie intraclasse et I_B est l'inertie interclasse.

Le but de la méthode des k-means est de chercher la partition telle que I_w soit minimale pour avoir en moyenne des classes bien homogènes. Cela revient à chercher le maximum de I_B .

Cette méthode a plusieurs avantages. Tout d'abord, il existe existe des algorithmes particuliers permettant d'assurer sa convergence. On pourra toujours minimiser I_w . Mais elle a aussi des inconvénients décisifs qui nous dissuaderont de l'utiliser. En effet, il est nécessaire de connaître le nombre de classes à considérer, ou autrement dit, dans notre cas, le nombre de zones géographiques homogènes. Bien entendu, en absence d'informations réelles sur les cultures et la géographie du territoire, nous serons bien incapables de justifier un choix arbitraire du nombre de

classes. De plus, la structure finale des classes dépend fortement du choix initial du nombre de classes. Enfin, cette méthode est très couteuse en calcul. Il est nécessaire de répéter cet algorithme un nombre important fois pour obtenir un regroupement stable ou de dégager les formes fortes.

Cette méthode ne semble pas adaptée à notre problème. Nous allons donc nous pencher sur un autre type de méthode de classification.

3. Classification hiérarchique

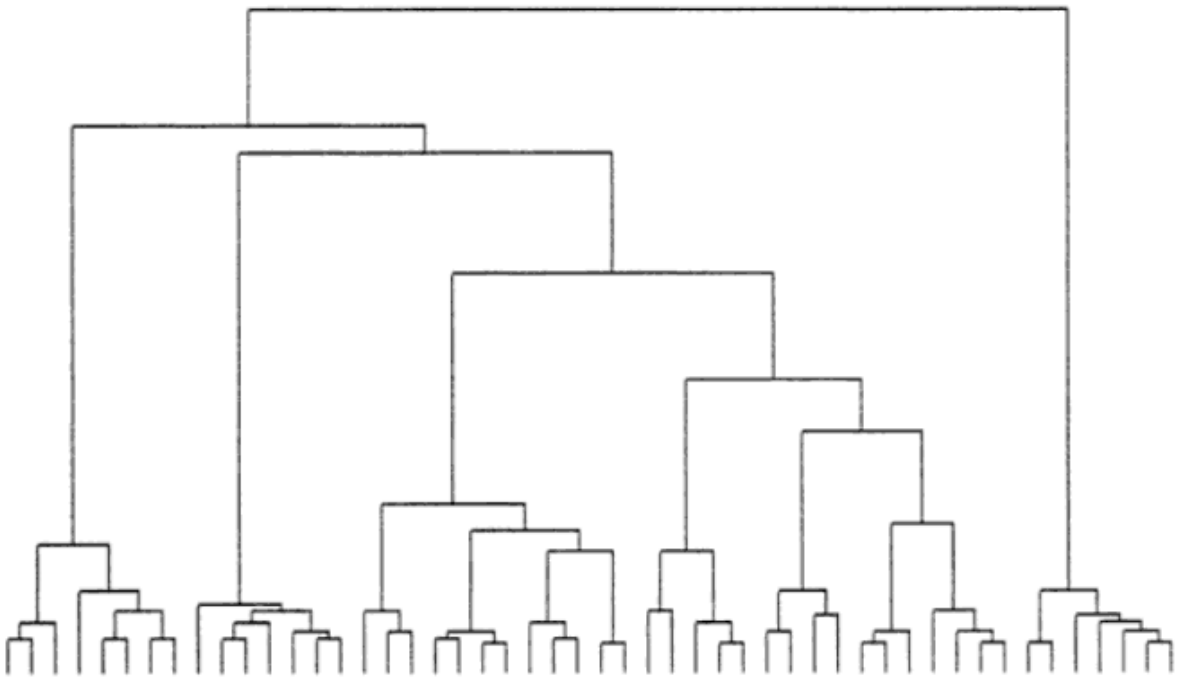
Il existe deux types de classification hiérarchique :

- La **méthode ascendante** qui consiste à appliquer une stratégie de bas en haut pour l'algorithme.

- La **méthode descendante** a une logique inverse et consiste à appliquer une stratégie de haut en bas.

Commençons d'abord par nous intéresser par la classification ascendante. Cette technique fournit un système de classes emboîtées dont l'hétérogénéité augmente avec la taille des classes. Ce système représente une hiérarchie de parties indicée que l'on visualise graphiquement sous la forme d'un arbre hiérarchique indicé appelé aussi dendrogramme.

Figure 3 : Exemple de dendrogramme



D'un point de vue formel, on dit que H_E est une hiérarchie d'un ensemble E si H_E est un sous-ensemble de $P(E)$ (l'ensemble des parties de E) tel que :

- L'ensemble E est un élément de H_E
- $i \in E, \{i\} \in H_E$
- $h \cap h' \in \{\emptyset, h, h'\}$ pour h et h' deux parties quelconques de H_E . Autrement dit, h et h' sont soit disjointes, soit incluses l'une dans l'autre.

L'algorithme utilisé dans le cadre de la classification hiérarchique ascendante est le suivant :

- Munir l'ensemble E des éléments à classer d'une distance ou d'une similarité d . La matrice $M_n^{(d)}$ contient les ressemblances des éléments de E deux à deux.

-Choisir une distance D entre deux parties disjointes de E telle que $D(\{i\},\{i'\}) = d(i;i')$ et utiliser l'algorithme suivant :

- Agréger en un nouvel élément les deux éléments de E les plus proches au sens de D

(donc au sens de d).

- Mettre à jour la matrice des distances en calculant les distances entre l'élément nouvellement formé et les $(n-2)$ éléments restants. Obtention de la matrice $M_D^{(n-1)}$ d'ordre $(n-1)$.

- Rechercher à nouveau les éléments que l'on agrège.

- Ainsi de suite, jusqu'à agréger tous les éléments de E en une seule classe.

L'arbre est ensuite coupé horizontalement pour obtenir un nombre de classes précis. Cependant, il existe plusieurs limites à cet algorithme de base. Premièrement, il nécessite qu'à chaque pas, on passe en revue l'ensemble des distances D entre paires de points pour obtenir la distance la plus petite. De fait, il est pratique pour des ensembles de taille modérée mais est très lourd en calcul pour des ensembles de grandes tailles. De plus, il est difficile de déterminer la coupure définitive de l'arbre. Il n'existe pas de test statistique permettant de guider la décision à prendre pour couper l'arbre. Néanmoins, on peut utiliser des critères externes pour choisir le nombre de classe à retenir. Il s'agit toutefois d'une décision subjective.

4.Critères d'agrégation

Nous utiliserons deux critères d'agrégation des données.

Tout d'abord la **méthode de Ward**, qui est le critère le plus usuel pour la classification ascendante. Le but de cette méthode est d'agréger les individus qui font le moins varier l'inertie l'interclasse (toujours dans le but de minimiser l'inertie intraclasse ou de maximiser l'inertie interclasse comme dans la méthode des k-means).

L'indice de dissimilarité entre deux classes correspond à la perte d'inertie interclasse résultant de leur regroupement.

Prenons g_A et g_B les centres de gravité des classes A et B et g_{AB} le centre de gravité de leur réunion. On a :

$$g_{AB}=\frac{p_A g_A+p_B g_B}{p_A+p_B}$$

où p_A et p_B sont les poids des classes A et B .

L'inertie interclasse étant la moyenne des carrés des distances des centres de classe au centre de gravité total, la variation d'inertie est égale à :

$$p_A d^2(g_A, g) + p_B d^2(g_B, g) - (p_A + p_B) d^2(g_{AB}, g)$$

où g est le centre de gravité de l'ensemble E .

Cette variation est égale à : $\frac{p_A p_B}{p_A + p_B} d^2(g_A, g_B)$

Cette variation est positive, on peut donc poser : $\delta(A, B) = \frac{p_A p_B}{p_A + p_B} d^2(g_A, g_B)$

Cette formule a été généralisée par Lance et Williams :

$$\delta((A, B); C) = \frac{(p_A + p_C)\delta(A, C) + (p_B + p_C)\delta(B, C) - p_C \delta(A, B)}{p_A + p_B + p_C}$$

Ainsi, la somme d'agrégation des différents nœuds de l'arbre est égale à l'inertie totale.

Le critère d'optimisation de cette méthode est assez naturel, ce qui en encourage l'utilisation.

Nous utiliserons ensuite un second critère, celui du **diamètre**. Il est défini par :

$$D_{diam}(h; h') = \sup(d(i, i'); i \in h, i' \in h')$$

Il s'agit de la plus grande distance entre un point h et un point h' . Ce critère crée, en agrégeant en priorité des paires de classes entre lesquelles l'écart D_{diam} est le plus faible, des classes bien compactes entre elles. Cela est particulièrement intéressant dans notre problématique. Nous cherchons en effet créer des zones géographiques homogènes « artificielles ». L'intérêt est donc d'utiliser l'information contenue dans des points proches pour déterminer sur une commune appartient à la même zone géographique. Cependant, cette méthode pourrait être un inconvénient. En effet, ces zones géographiques créées ne prennent pas en compte la vraie situation géographique des communes. Il y a donc des chances que l'on se retrouve avec des communes regroupées dans la même classe en ayant des conditions géographiques totalement opposées. La comparaison des deux critères permettra d'appréhender ce genre de problèmes.

II.2 Analyse exploratoire spatiale des données

Intérêt de l'analyse spatiale des données est de fournir un cadre pertinent d'analyse de données géo-référencées ou localisées. L'analyse spatiale se prête bien au cadre de notre étude. En effet, elle se distingue de l'analyse statistique classique parce que les observations analysées sont en général dépendantes, et cette dépendance est due à la localisation relative des points de mesure et conduit à un débordement d'information d'un lieu à un autre. Ce débordement crée donc de la redondance de l'information disponible sous la forme de valeurs localisées des indicateurs statistiques étudiés.

1.Aléa spatial

Avant de pouvoir étudier les corrélations spatiales, il convient de définir ce qui s'oppose aux structures de dépendance spatiale. Il s'agit de l'aléa spatial.

Dans ce contexte, soient n points répartis de manière aléatoire dans une région R. On suppose que les positions de ces points n'ont aucune influence les unes entre elles. C'est le concept-clé de l'aléa spatial. De fait, si pour chaque $i=1,..,n$ la variable de Bernouilli $X_i(C)$ décrit l'événement que ce point i se trouve dans la région C. Alors sous l'hypothèse d'aléa spatial, les variables aléatoires $\{X_i(C): i=1,\ldots,n\}$ sont statistiquement indépendantes pour chaque région C. Il s'agit de l'hypothèse fondamentale de l'aléa spatial complet. Cette hypothèse d'aléa spatial complet s'oppose aux structures de dépendance spatiale que nous voulons étudier.

2.Autocorrélation spatiale : Indices de Moran et Geary

L'analyse des structures de dépendance s'effectue grâce à l'étude de autocorrélation spatiale. Elle concerne la corrélation d'une variable X avec sa mesure décalée ou déplacée, et son intensité est une fonction d'importance de ce décalage ou déplacement. Dans un contexte spatial, on mesure la corrélation entre les valeurs prises par X et dans les sites voisins de façon analogue. C'est une méthode similaire à l'autocorrélation des séries temporelles. Les statistiques les plus simples pour mesurer les autocorrélations spatiales sont l'**Indice de Moran** et l'**Indice de Geary**. Ils permettent de valider l'existence d'une autocorrélation spatiale pour une variable donnée, et ce, pour l'ensemble des points.

Commençons par définir la notion de moyenne locale.

Pour tout caractère statistique X observé en n lieux, on appelle moyenne locale de X dans le voisinage d'un point s_j la quantité $\widetilde{X}=\sum_{k=1,k\neq j}^{n} w_{jk}X_k$ où w_{jk} est un poids appliqué à la comparaison entre deux localisations. Il s'agit généralement d'un poids exprimant la distance entre deux observations : $w_{jk}=\frac{1}{d_{jk}}$ avec d_{jk} la distance entre deux observations.

On appelle **Indice d'autorégression de Moran** la quantité :

$$I_M=\frac{\sum_{i=1}^{n} p_i\left(X_i-\bar{X}\right)\left(\widetilde{X}_i-\bar{\widetilde{X}}\right)}{\frac{\sum_{k=1}^{n}\left(X_k-\bar{X}\right)^2}{n}} \quad \text{avec} \quad p_i \text{ tels que } \sum_{k=1}^{n} p_i=1$$

I_M est d'autant plus grand que les valeurs semblables ont tendance à se regrouper en lieu proches. Inversement, il est d'autant plus petit que les valeurs dissemblables ont tendance à se regrouper en lieux proches. De plus, I_M se rapproche de son espérance mathématique sous hypothèse d'indépendance spatiale, et plus les valeurs semblables de X semblent se répartir de manière uniforme sur le domaine étudié. Cette espérance vaut $E\left(I_M\right)=\frac{-1}{n-1}$.

En pratique, on se sert de ce critère pour évaluer la corrélation spatiale. Si l'indice calculé est supérieur à l'espérance mathématique alors les données présentent une autocorrélation spatiale positive. Les valeurs semblables ont alors tendance à se regrouper. Et inversement pour un indice inférieur à l'espérance mathématique. Les valeurs dissemblables ont alors tendance à se regrouper.

On définit de la même manière **l'indice d'autorégression de Geary** la quantité :

$$C=\frac{(n-1)\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} w_{ij}\left(X_i-X_j\right)^2}{2\left(\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} w_{ij}\right)\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2}$$

La valeur de l'indice de Geary varie généralement entre 0 et 2. Les valeurs inférieures à 1 indiquent une autocorrélation spatiale positive alors que celles supérieures à 1 indiquent une autocorrélation négative entre observations.

3.Graphique de dispersion de Moran

Cet outil d'analyse exploratoire présente l'intérêt méthodologique de fournir les outils de visualisation et de compréhension des raisons pour lesquelles la statistique de Moran prend la valeur calculée. Sa rationalité repose sur une interprétation de cet indice comme un coefficient de pente de régression artificielle.

On commence par calculer l'image spatiale de la variable centrée $\dot{X} = P_1 X$ associée à l'opérateur de centrage $P_1 = I_n - \frac{1}{n}\mathbf{1}^t\mathbf{1}$. Son image spatiale est donc $\widetilde{\dot{X}} = G\dot{X} = GP_1 X$. On construit ainsi le graphique de dispersion bivariée $\left\{(\widetilde{\dot{X}}_i, \dot{X}_i); i \in \{1,...n\}\right\}$ alors appelé graphique de Moran.

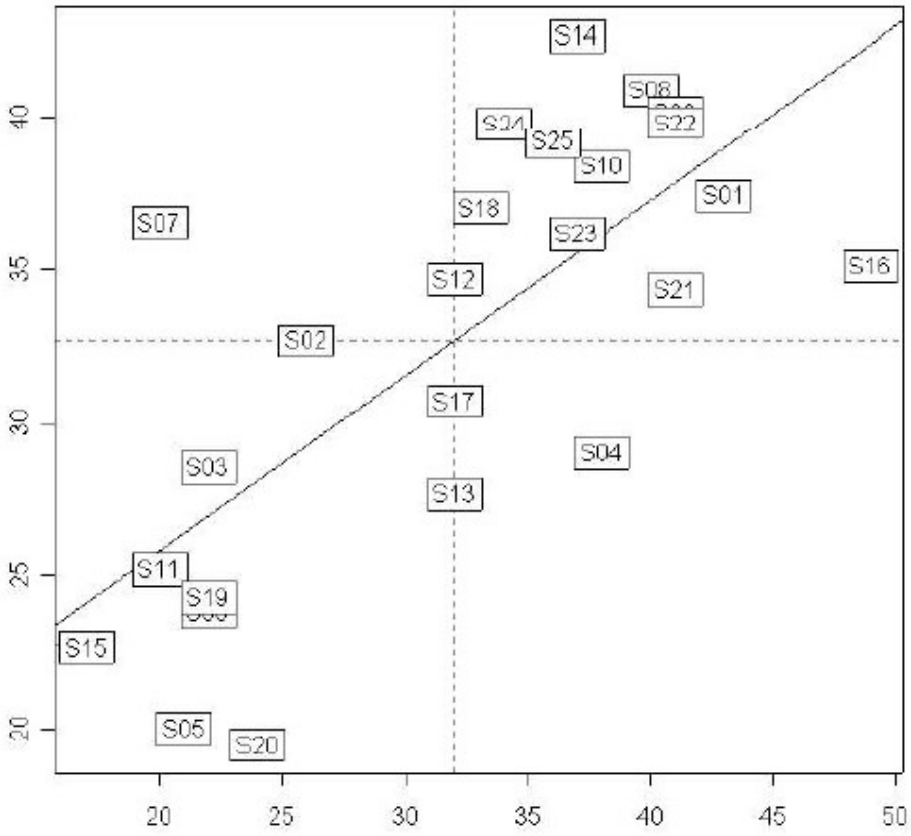


Figure 4 : Exemple d'un graphique de dispersion de Moran

Nous avons en abscisse les valeurs d'une variable et en ordonnée la moyenne des valeurs des voisins (variable « retard »). La droite est l'estimation du modèle de régression linéaire. Les droites en pointillés passent par les moyennes. On interprète alors l'indice de Moran comme le coefficient de régression de la droite. La figure peut être découpée en 4 quadrants : Bas-Gauche, Bas-Droite, Haut-Gauche et Haut-Droit. Ainsi, les cadrants en Bas-Gauche et Haut-Droit correspondent à un regroupement spatial de valeurs similaires alors que les cadrants Bas-Droite et Haut-Gauche correspondent à un regroupement spatial de valeurs non-similaires.

4.Autocorrélation spatiale locale

Les indicateurs locaux d'association spatiale (LISA) sont discutés de façon notable par Anselin (1995). Ils ont pour but d'affiner les méthodes de repérage et de description des structures d'autocorrélation spatiale. Celui-ci propose une implantation locale de l'indice d'autorégression de Moran afin d'apprécier le degré de similarité par petits voisinages. En effet, l'indice de Moran ne décrit une région qu'en une seule valeur et ne donne aucune information sur les variations locales de l'indice et sur sa dispersion. Il se peut que certaines parties de la région présentent un arrangement spatial dispersé alors que d'autres appartiennent au typé groupé. L'indice global ne représente donc une information exploitable que si la région d'étude est suffisamment homogène. Mais généralement, elle ne l'est pas. Cette méthode requiert aussi que les données soient agrégées par zone, d'où l'utilité de la classification hiérarchique. La procédure calcule l'indice d'autocorrélation de Moran pour chaque observation, dans ce cas-ci chaque secteur de dénombrement,en fonction des observations localisées à proximité (voisinage) ;ceci permet d'obtenir une mesure des similarités et différences locales.

L'indice de Moran appliqué localement a pour expression :

$$I_i = \frac{\left(Z_i - \bar{Z}\right)}{S_Z^2} \sum_{j=1}^{n} w_{ij}\left(Z_j - \bar{Z}\right)$$

II.3 Mise en œuvre de la méthode

Dans un premier temps, il va s'agir de tester et d'étudier la structure de dépendance spatiale des Loss Cost des communes du portefeuille 2013 dans sa totalité, en utilisant les indices de Moran et Geary. Il s'agit d'une approche préliminaire nous permettant de savoir si, au niveau global de la France, les valeurs semblables des Loss Cost ont tendance ou non à se regrouper. Cela nous

permettrait de justifier l'hypothèse de présence de zones géographiques homogènes. Dans un deuxième temps, nous effectuerons une classification hiérarchique ascendante sur les données spatiales dans l'espoir d'obtenir des zones géographiques homogènes précises. Ensuite, nous étudierons la structure de dépendance spatiale dans chaque zone créée, d'abord de manière globale, puis en utilisant des autocorrélations locales spatiales.

1.Autocorrélation spatiale du portefeuille dans son ensemble

Grâce à la librairie spdep du logiciel R, nous pouvons calculer l'autocorrélation spatiale des communes de la France entière. L'idée est d'analyser la distribution spatiale des Loss Cost et d'en déduire s'il y a corrélation spatiale entre les différentes communes.

L'hypothèse H_0 sera : H_0 : « La répartition spatiale des Loss Cost des communes dépend d'un processus spatial aléatoire. » Bien évidemment, notre but est de rejeter cette hypothèse. En effet, nous supposons que la répartition spatiale des Loss Cost présente une agrégation statistiquement significative.

La première étape est de constituer le graphe de voisinage. La statistique spatiale s'appuie sur la quantification du voisinage et la structure spatiale s'exprime comme une relation quantitative, et mesure sur chaque couple de deux points l'importance du premier élément pour le second. Il faut néanmoins se demander ce qu'est un voisin . Jusqu'à quelle distance seuil peut-on considérer que le Loss Cost d'une commune est dépendant du Loss Cost d'une commune environnante ? L'idée est d'utiliser les indices de Moran et Geary pour estimer cette distance. La distance idéale à considérer pour les voisins correspondra à celle qui maximisera l'indice de Moran et qui minimisera l'indice de Geary. On ne pondérera pas les voisins, c'est-à-dire que chaque voisin d'une commune aura la même influence sur la commune qu'un autre voisin. Notre intuition nous dirait que dans un modèle ayant trait à des aléas climatiques, l'information qu'une commune apporte à une autre est inversement proportionnelle à la distance entre ces deux communes. Néanmoins, pour des raisons de temps de calcul, nous nous en abstiendrons. On peut néanmoins noter que des modèles existent pour essayer d'étudier la qualité de l'information d'une commune à une autre, tels que les modèles de crédibilité spatiale. Ils constitueraient cependant un axe d'amélioration substantiel à notre modèle.

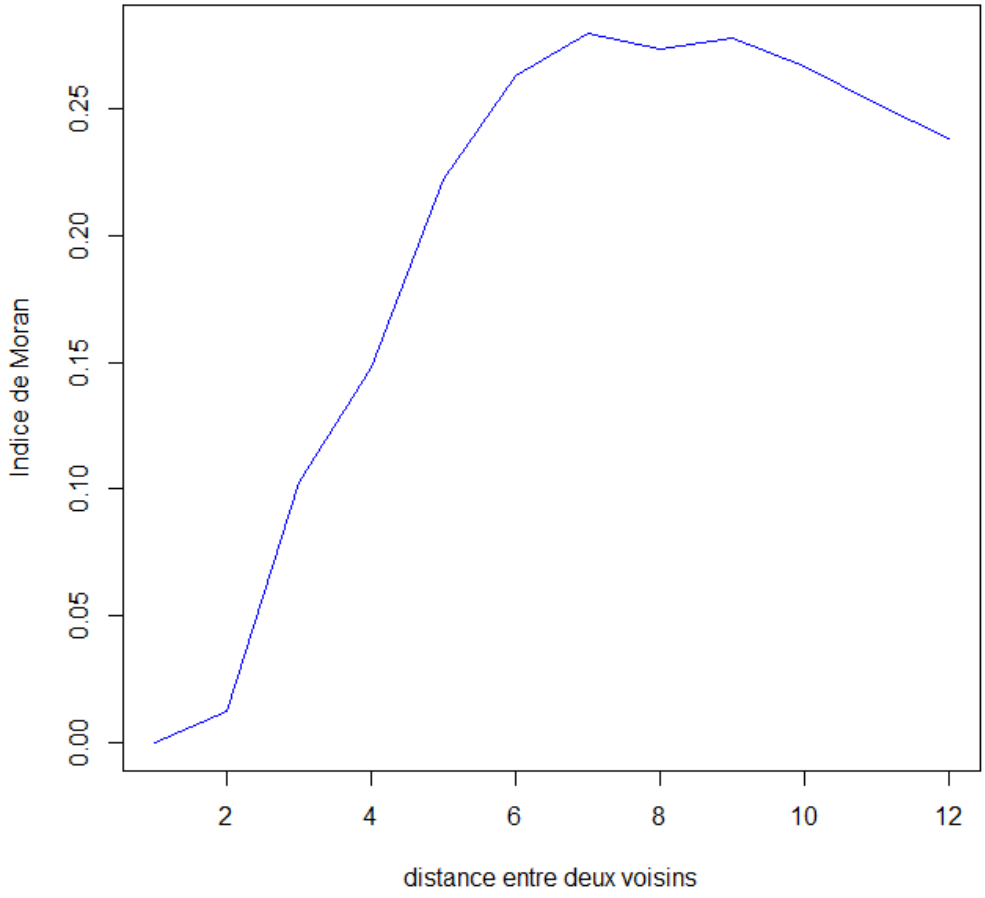


Figure 5 : Courbe décrivant l'indice de Moran en fonction de la distance seuil

On peut voir que la courbe atteint son maximum à 7km de distance. C'est la distance seuil que nous utiliserons donc pour l'indice de Moran. Pour cette distance, nous obtenons un indice

$I_M = 0{,}2796927573$ d'espérance $E(I_M) = -0.0003362475$ et une p-value $p < 2{,}2 \quad 10^{-16}$. On peut donc en tirer deux conclusions. Premièrement, l'hypothèse H_0 est rejetée : Il existe bien une structure de dépendance spatiale entre les Loss Cost des communes de France. De plus, I_M étant supérieur à son espérance, les données présentent une autocorrélation positive : les valeurs de loss-cost semblables ont tendance à se regrouper spatialement.

Observons maintenant le graphique de dispersion de Moran associé à notre étude.

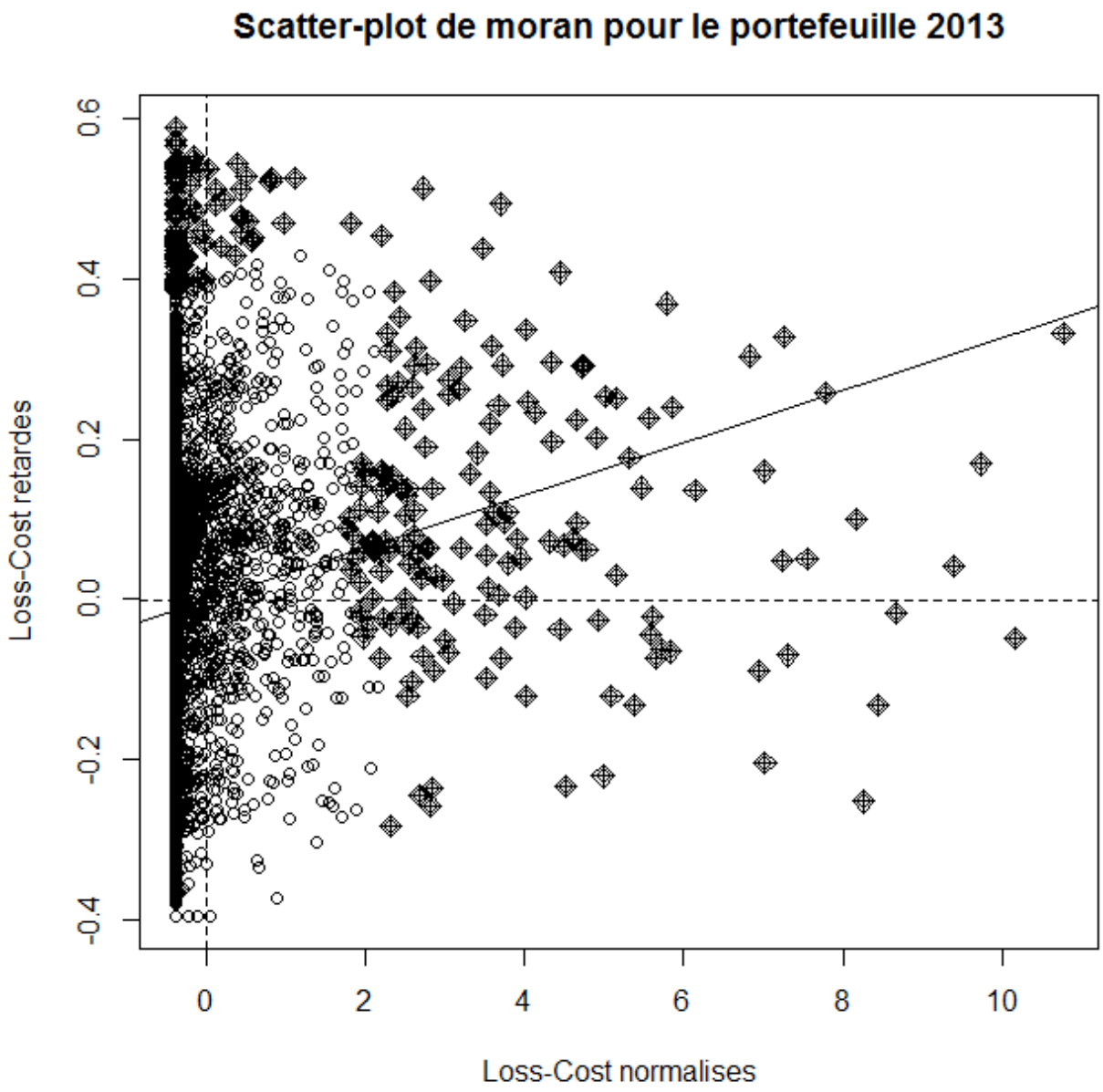


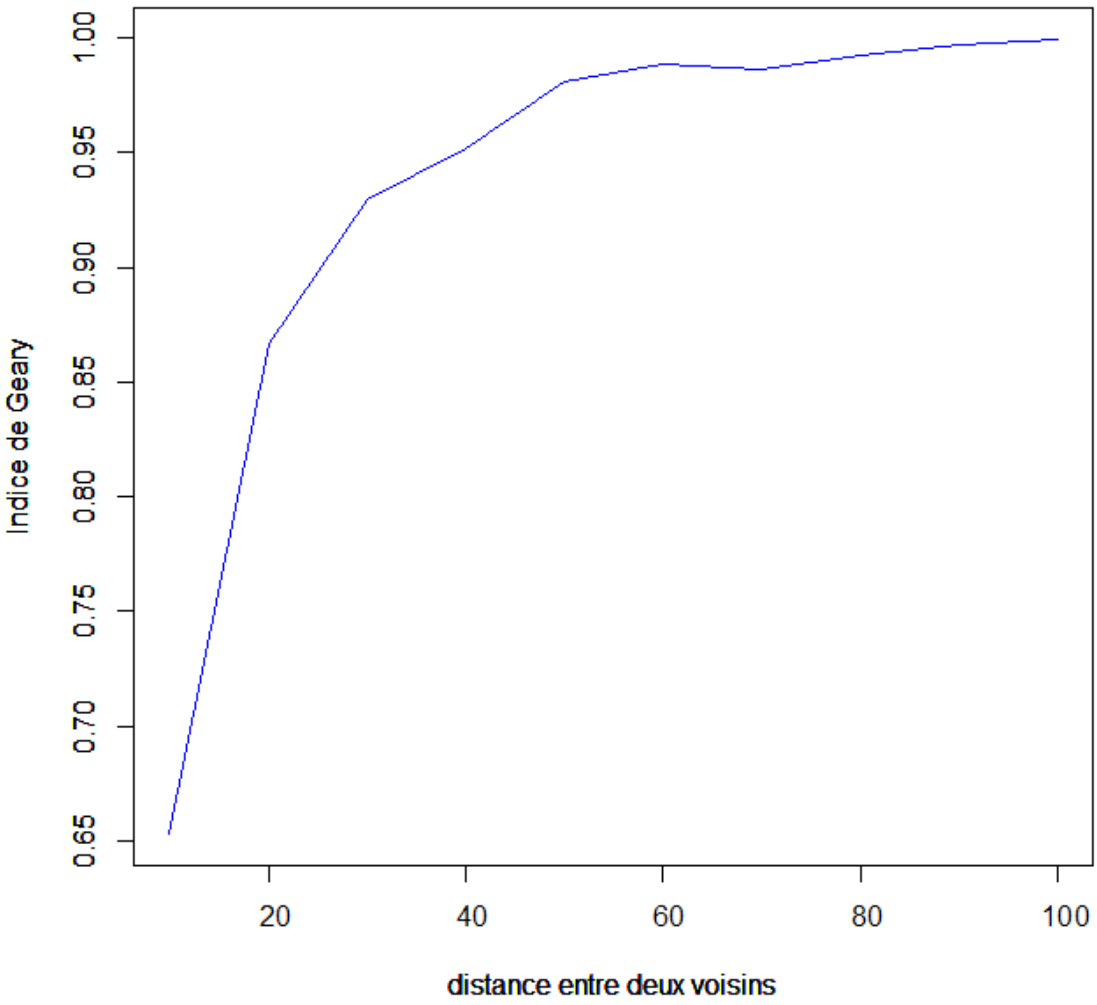
Figure 6: Graphique de dispersion de Moran

L'avantage principal du scatter-plot est que données étant normalisées, on peut voir directement où se situent les points extrêmes. Il s'agit de ceux éloignés de deux fois l'écart-type par rapport à la moyenne. Ils sont représentés par des losanges sur le graphique. L'idée pourrait être d'étudier localement ces points pour essayer de déterminer leur nature. Néanmoins, l'interprétation de la nature de ces points est un champ qui dépasse le cadre de l'analyse exploratoire des données, et nous ne le traiterons pas dans le cadre de ce mémoire.

On voit que le coefficient de régression de la régression est clairement positif ce qui confirme nos calculs précédents. On observe néanmoins un nombre non-négligeable de points dans des zones présentant des associations spatiales dissimilaires. Cela pourrait indiquer l’existence de différents régimes d'association spatiale.

Appliquons le même raisonnement avec l'indice de Geary.

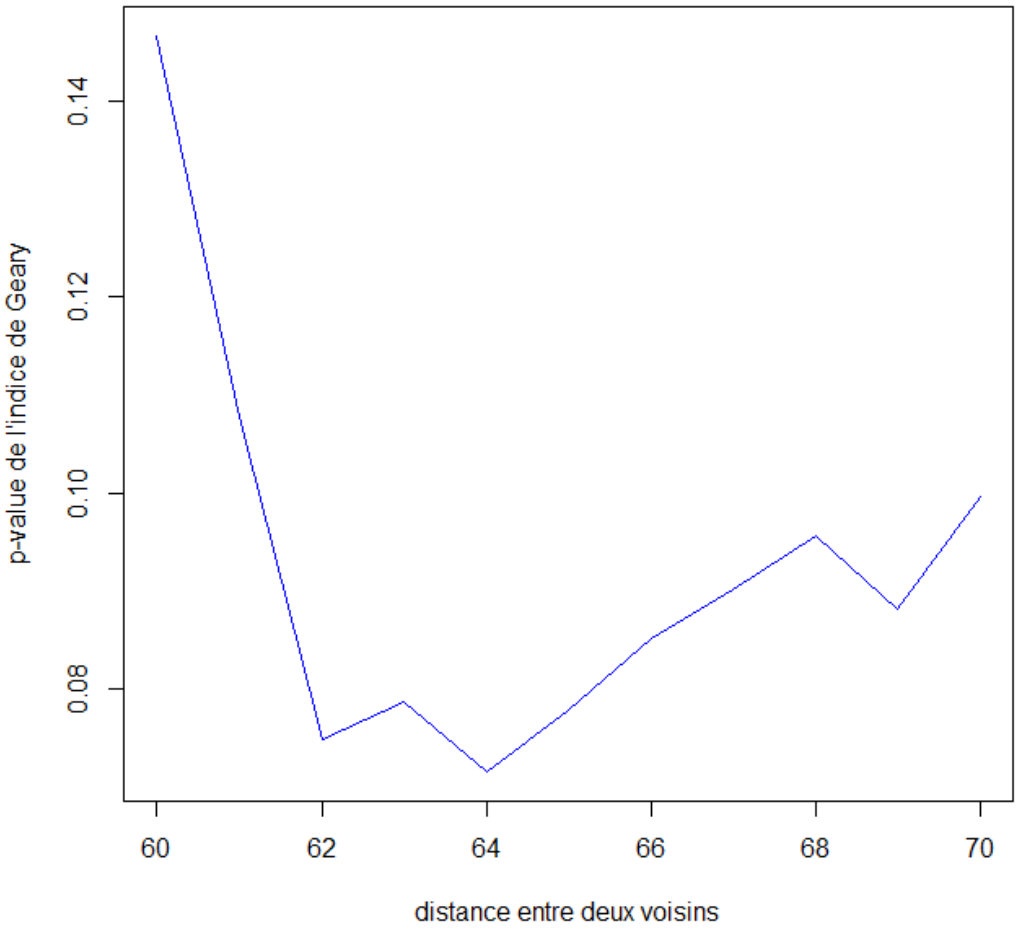
Figure 7 : Courbe décrivant l'indice de Geary en fonction de la distance seuil



On constate que la courbe de l'indice de Geary en fonction du temps semble tendre vers son espérance : 1. On peut trouver la valeur critique de la distance seuil telle que la p-value nous donne un niveau de confiance à 95 %. On obtient $d_{critique}$ = 64km. C'est la distance seuil que nous utiliserons pour déterminer C , l'indice de Geary.

Nous obtenons un indice $C=0{,}8426541295$ d'espérance $E(C)=1$ et une p-value $p=0{,}05$. L'hypothèse H_0 est rejetée, et comme C est inférieur à 1, on peut conclure qu'il existe une autocorrélation positive entre les communes.

Figure 8 : Courbe décrivant les p-values en fonction de la distance seuil



La différence de valeur seuil entre I_M et C provient sûrement du fait que C est plus sensible aux variations locales. De fait, on peut déjà anticiper le fait qu'il risque d'avoir de grandes similarités au niveau local, et ce, jusqu'à un rayon de 70km. Une autre explication provient du fait que l'indice de Moran se base sur la moyenne (ou plutôt sur les écarts à la moyenne) et est donc moins sensible aux valeurs abérrantes que l'indicateur de Geary qui lui va se baser sur les écarts entre voisins.

2.Test de significativité de la composante spatiale des indices

Nous avons utilisé les indices de Moran et de Geary pour décrire la structure d'association globale des communes du portefeuille. Il faut maintenant vérifier si la composante spatiale des-dits indices est significative ou non. Pour cela, nous allons effectuer des simulations de Monte-Carlo. Les simulations de Monte-Carlo permettent de tester l'hypothèse que la valeur d'un indice est le résultat d'un processus aléatoire. Si cette hypothèse est rejetée, alors la valeur de cet indicateur est jugée significative. Voici le principe de base : Les valeurs d'une variable aléatoire (ici les loss-cost) sont distribuées de manière aléatoire (suivant un critère de répartition choisi) entre les unités spatiales, afin de pouvoir calculer une valeur simulée de l'indice. Les simulations sont répétées en un nombre suffisamment élevé de fois pour que les valeurs simulées puissent être approximées par une loi de distribution. Cette distribution de valeurs simulées sera utilisée comme base de comparaison pour la valeur observée de l'indice. Si la valeur observée se trouve à l'intérieur des valeurs critiques données par les seuils de significativité, alors on peut conclure qu'elle est non-significative. Au contraire, si la valeur observée est en dehors des valeurs critiques, elle n'apparaît pas comme le résultat d'un processus aléatoire et peut donc être considérée comme significative.

Dans le cas des indices de Moran et de Geary, le test le plus utilisé est le test des permutations. Il s'agit de garder la même structure de données, mais pour chaque simulation, le données sont permutées de manière aléatoire entre les unités spatiales. Une fois les simulations avec les données permutées réalisées, l'indicateur observé peut être comparé avec la distribution des valeurs simulées. Son rang dans cette distribution permet de calculer une pseudo p-value et de conclure quant à la significativité de la composante spatiale de l'indicateur.

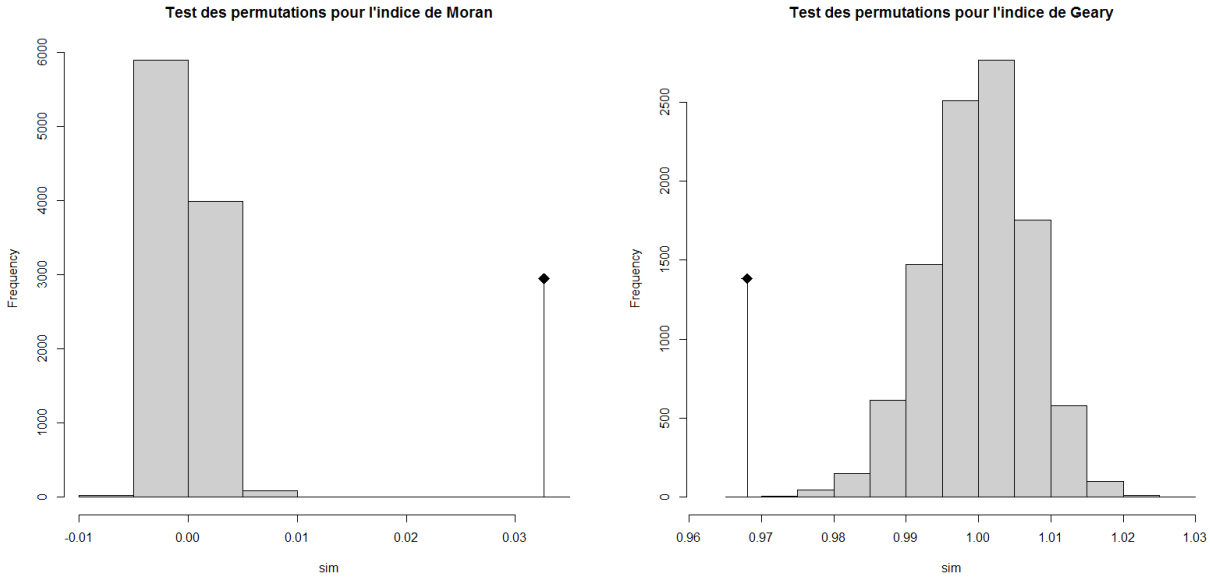


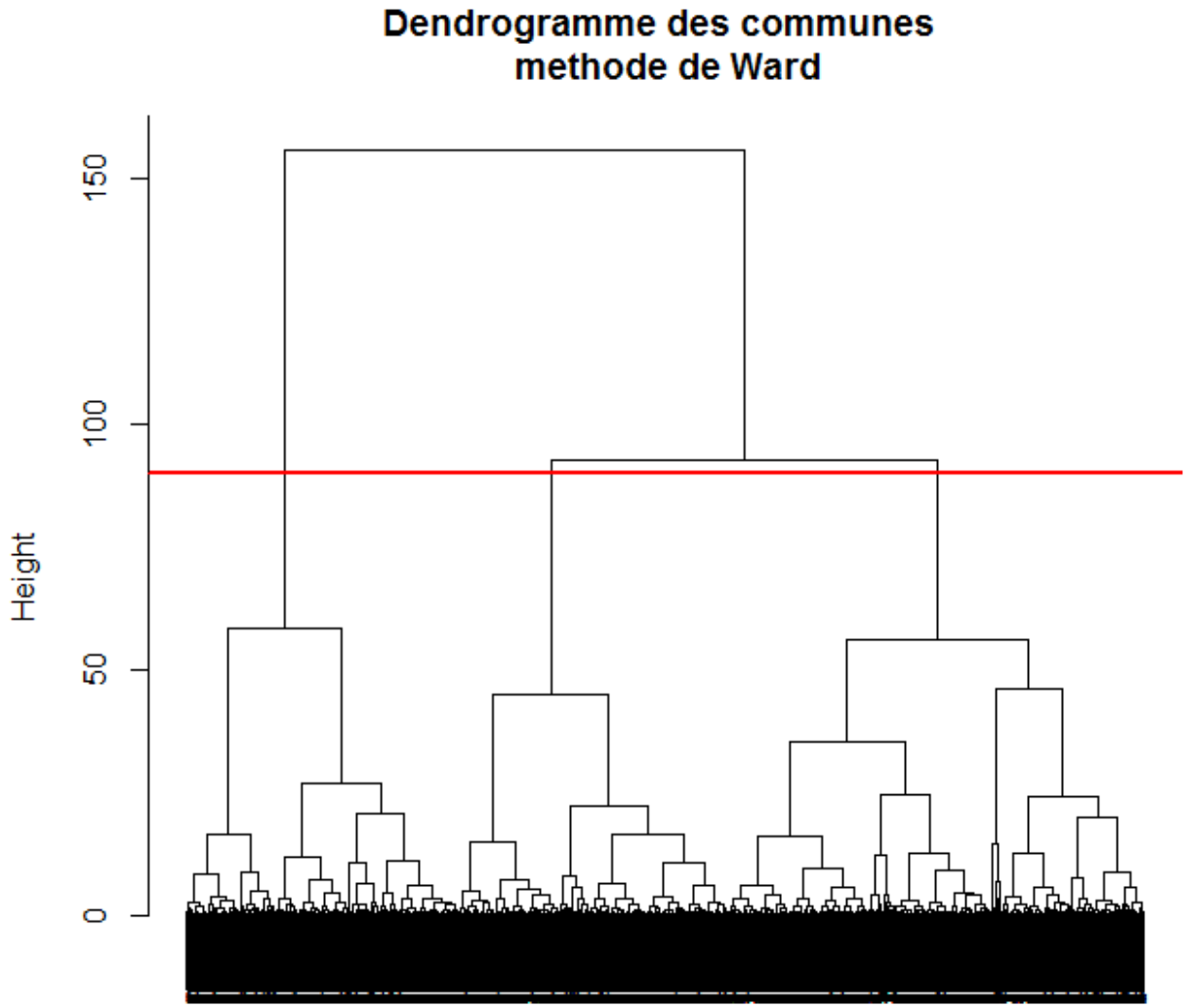
Figure 9: Tests de permutations des indices d'autocorrélation globaux

Nous avons effectué 10 000 simulations de Monte-Carlo pour chaque indice. On voit que les valeurs observées sont bien en dehors des valeurs critiques. On obtient des pseudo p-values de 1×10^{-4} pour l'indice de Moran et 2×10^{-4} pour l'indice de Geary, nous permet de rejeter l'hypothèse de processus aléatoire et donc de s'assurer de la significativité des indices.

3.Classification ascendante hiérarchique du portefeuille

La troisième étape est donc maintenant de réaliser une classification ascendante hiérarchique sur les données. Nous savons en effet à présent qu'il existe une autocorrélation spatiale entre les communes. Le but est de les séparer en zones géographiques homogènes de manière à étudier la structure de dépendance spatiale de manière beaucoup plus fine.

Figure 10 : Dendrogramme associé à la méthode de Ward sur les coordonnées des communes du portefeuille 2013

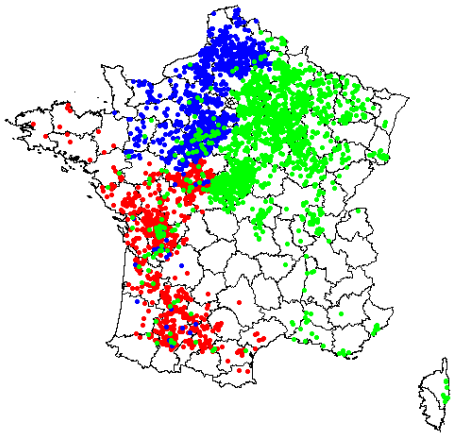


Nous avons utilisé la métrique euclidienne comme mesure géométrique et le critère de Ward comme critère d'agrégation des données. Ces deux méthodes se complètent très bien et apportent généralement de bons résultats.

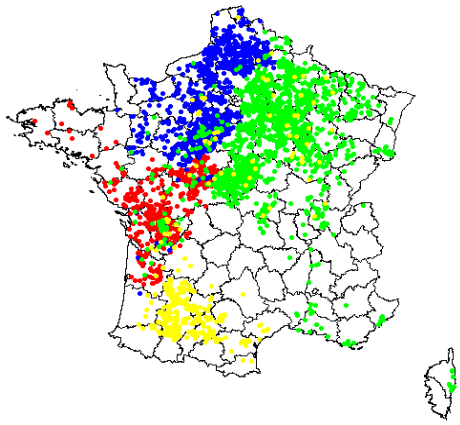
Suivant le niveau où le dendrogramme est coupé, différentes classes seront obtenues. Il sera utile de comparer les coupures et choisir celle qui aura les classes les plus homogènes.

Nous coupons le dendrogramme en 3 à 5 classes et nous regardons l'évolution des variances intraclasses pour les classes considérées.

3 classes



4 classes



5 classes

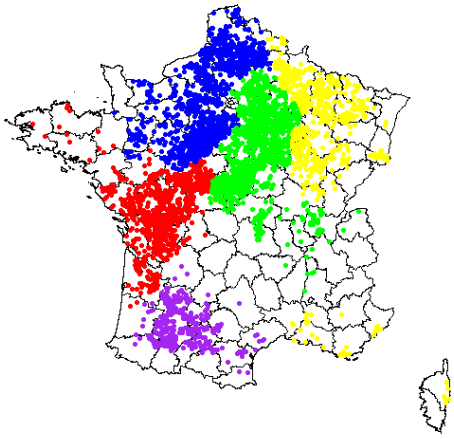


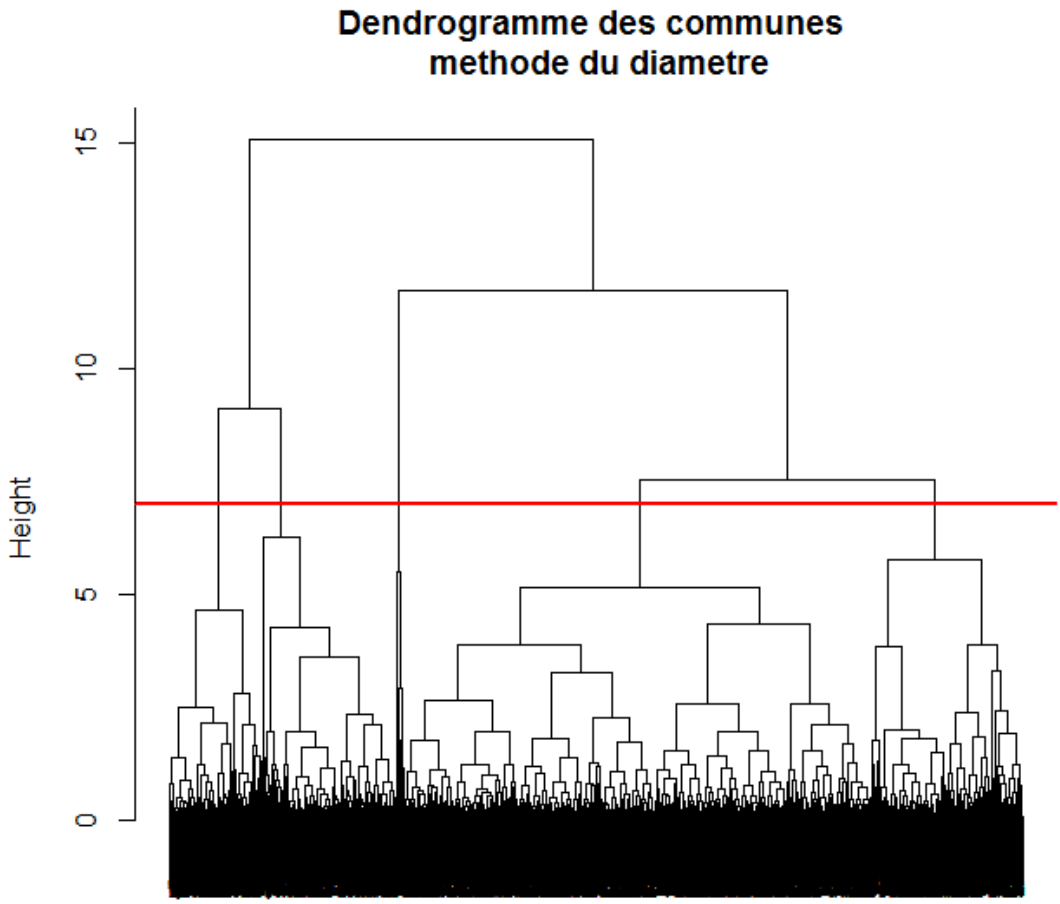
Figure 11 : Classes associées à la classification hiérarchique ascendante utilisant la méthode de Ward

On voit que lorsque l'on coupe le dendrogramme en 4 classes, celles-ci sont encore quelque peu agrégées et mélangées. On retiendra donc 5 classes. Ces classes sont relativement homogènes en terme de localisation, sauf la classe jaune. En effet, cette classe a une grande majorité de ses points dans le nord-est de la France et quelques autres dans le sud-est et la Corse. Le climat entre ces deux régions est pourtant très différent, et les cultures produites aussi. Ces deux régions ne devraient pas être dans la même classe. On peut y voir une limite du modèle de Classification Ascendante Hiérarchique. Cette méthode reste au final une méthode statistique, qui ne prend pas en compte les données agricoles, ou météorologiques. Un mélange avec un modèle physique ou économétrique pourrait très certainement permettre une classification des données beaucoup plus fine.

Dans le graphique des pertes d'inertie relatives donné en annexe, on peut voir une perte d'inertie lorsque l'on passe de 4 à 5 partitions, ce qui nous conforte dans notre choix d'utiliser 5 partitions.

Utilisons maintenant le critère du diamètre dans notre classification ascendante hiérarchique.

Figure 12: Dendrogramme associé à la méthode du diamètre sur les coordonnées des communes du portefeuille 2013



Regardons la distribution spatiale des classes.

On peut voir que cette méthode regroupe effectivement les données dans des clusters bien compacts. D'un point de vue spatial, toutes les régions sont bien séparées les unes des autres. Elles se sont pas mélangées. De plus, contrairement à la méthode de Ward, le sud-est et la Corse ne sont pas dans le même cluster que le nord-est, ce qui est plus cohérent d'un point de vue géographique.

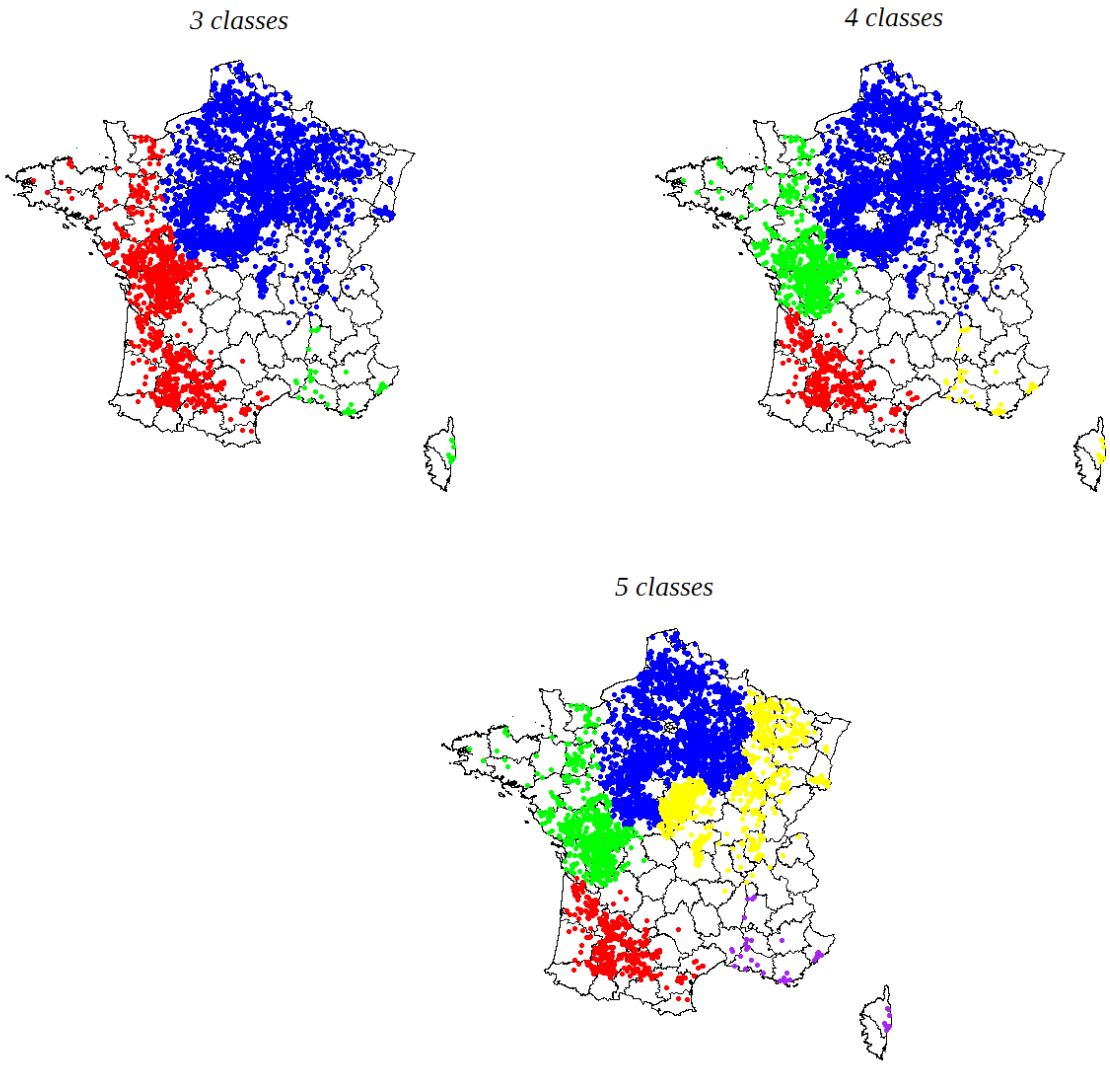


Figure 13: Classes associées à la classification hiérarchique ascendante utilisant la méthode du diamètre

Le graphique des pertes d'inertie donné en annexe ne nous permet pas de voir une perte d'inertie significative pour une partition donnée. On choisira une partition en 5 classes car ce nombre reste en accord avec le nombre de partitions trouvées avec la méthode de Ward.

Dans la suite du mémoire, nous utiliserons donc la partition obtenues par le critère du diamètre. Nous pouvons essayer d'interpréter les régions obtenues grâce à la carte des climats donnée en annexe.

La première classe, en rouge, regroupe le sud-ouest de la France, caractérisée par un climat océanique aquitain plutôt doux. On y retrouvera des villes telles que Bordeaux ou Toulouse.

La deuxième classe en vert regroupe le nord-ouest de la France. Les zones comprises dans

cette classe sont sous un climat océanique nord-ouest. On y retrouvera des villes telles que Poitiers, Rennes ou Nantes.

La troisième classe en bleu, regroupe le centre et le nord de la France, soumis à un climat océanique dégradé. On y retrouvera des villes telles que Paris, Lille ou Tours.

La quatrième classe en jaune regroupe une partie du centre ainsi que l'est de la France. C'est la région la plus hétérogène, climatiquement parlant car elles regroupe à la fois des zones sous climat océanique dégradé mais aussi des zones sous climat semi-continental et des zones sous climat montagnard. D'une manière générale, le climat est plus rude dans la quatrième classe que dans les autres classes. On y retrouvera des villes telles que Strasbourg, Clermont-Ferrand ou encore Dijon.

La cinquième classe en violet regroupe le sud-est de la France ainsi que la Corse. Les communes de cette classe sont soumises à un climat méditerranéen chaud. On y retrouvera des villes telles que Marseille, Nice ou encore Avignon.

Notre analyse nous montre clairement que nous arrivons à des régions semblables à la carte des climats, ce qui nous conforte dans notre choix de segmentation des régions.

Nous appellerons par la suite les classes par leurs attributs :

- Région 1 : Région sud-ouest (en Rouge)

- Région 2 : Région ouest (en Vert)

- Région 3 : Région nord (en Bleu)

- Région 4 : Région est (en Jaune)

- Région 5 : Région sud-est (en Violet)

4.Autocorrélations spatiales locales des régions

Maintenant que nous avons réussi à séparer les communes du portefeuille en zones géographiques homogènes, nous allons pouvoir étudier plus en détail les autocorrélations spatiales locales de chaque zone. L'idée est de représenter sur une carte la structure de dépendance locale de chaque commune par rapport aux communes de son voisinage. On utilisera les valeurs de l'indice de Moran Local Normalisé : $Z(I_i)=\frac{I_i-E(I_i)}{\sqrt{(V(I_i))}}$ De même qu'avec les indices de autocorrélation spatiale globaux, l'indice de Moran Local Normalisé, de forts indices supérieurs ou inférieurs à zéro

révèlent respectivement d'importantes similarités ou différences, au niveau du voisinage d'une unité spatiale donnée. La figure 14 nous montre la carte des LISA normalisés pour la région 4. Les cartes des autres régions sont données en annexe.

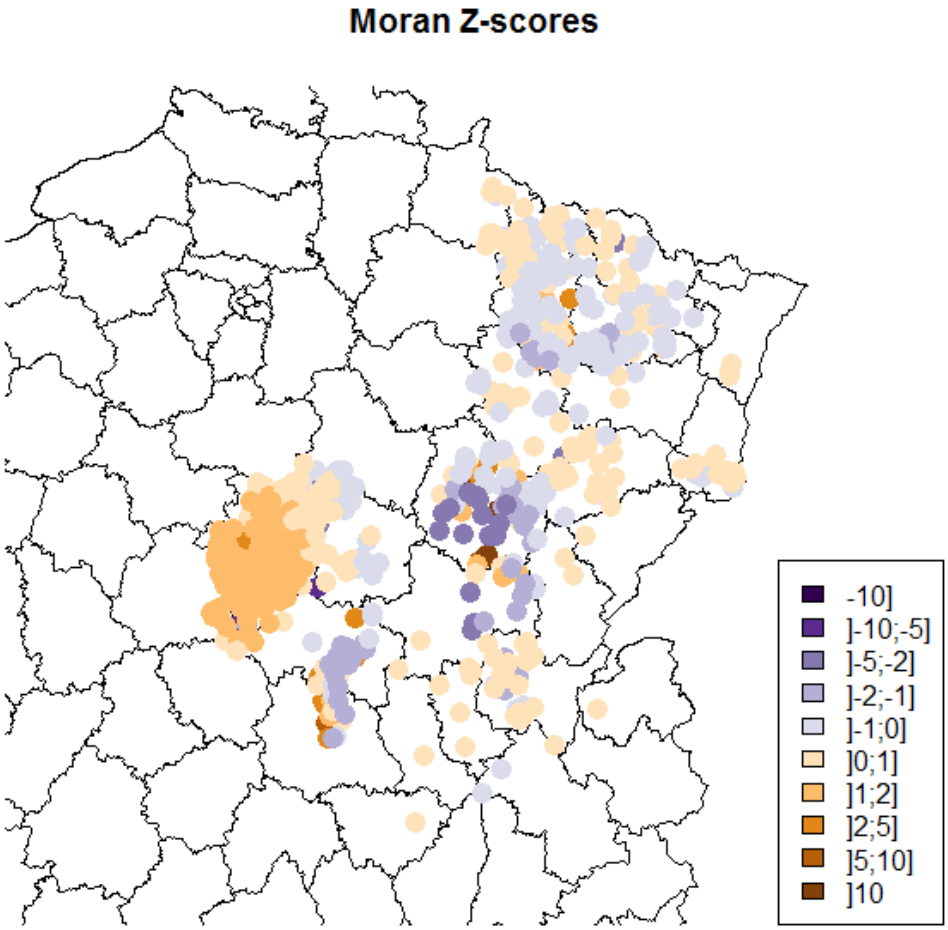


Figure 14: Carte des Indices de Moran Locaux pour la région 4

Le degré d'autocorrélation spatiale est exprimé par le biais d'un dégradé de couleurs révélant, en bleu, les zones à autocorrélation négative, et en marron, les zones à autocorrélation positive. Plus la valeur est élevée, plus le degré d'autocorrélation spatiale est élevé. Cette carte permet, combinée à la carte de la distribution spatiale des Loss-Cost, de déterminer les ensembles de clusters similaires, mais aussi ceux présentant des différences avec leur voisinage. Un cluster important d'indices positifs au centre de la France est visible. Cela pourrait être dû à la présence de

cultures similaires, et donc sensibles au même type de risque. On voit aussi plusieurs petits clusters d'indices négatifs.

Bien entendu, les LISA reposent sur l'hypothèse statistique de normalité. Il est donc nécessaire de vérifier la signification statistique des structures en clusters de la région. La figure 15 nous donne une carte de signification statistique des LISA de la région.

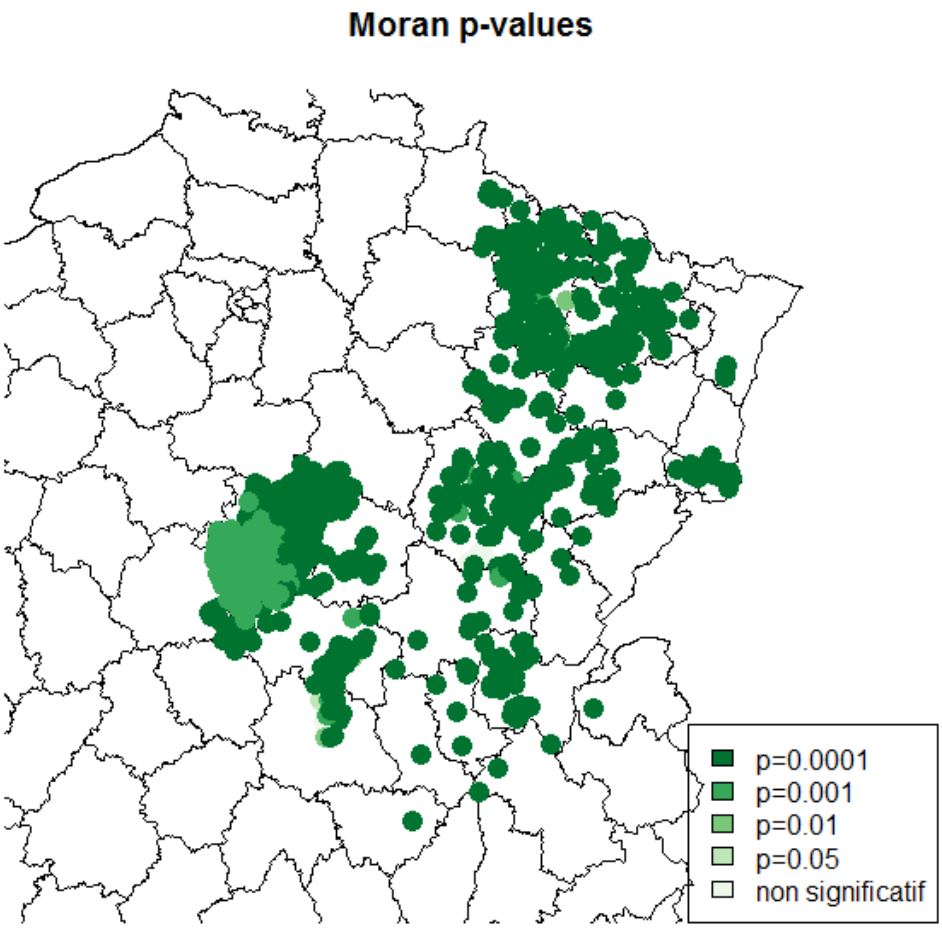


Figure 15: Carte de signification statistique des Indices de Moran Locaux de la région 5

La p-value est calculée pour chaque location de la région. Plus elle est faible, plus l'éventuel cluster qu'elle forme avec son voisinage est statistiquement significatif. Dans notre exemple, plus

une commune est foncée, plus elle est corrélée (positivement ou négativement) avec son voisinage. On remarque immédiatement que les LISA sont globalement assez significatifs dans la région ce qui conforte l'hypothèse de clusters spatiaux au niveau local. Les autres régions présentent aussi des LISA statistiquement significatifs qui mettent en évidence des clusters spatiaux.

On pourrait néanmoins évoquer une limite du LISA et même plus généralement de l'indice de Moran. L'indice de Moran calcule l'écart de la valeur d'un point et celle de ses voisins à la moyenne. Les valeurs éloignées de la moyenne sont donc bien plus mis en avant que les valeurs proches de la moyenne. Les valeurs proches de la moyenne se retrouvent alors avec des niveaux d'autocorrélation spatiale non-significatifs et n'influencent pas l'autocorrélation globale. De fait, les communes où les LISA ne sont pas significatifs généralement des zones où les valeurs sont toutes proches de la moyenne.

Partie III
Régression locale

Nous avons dans un premier temps étudié la structure de dépendance spatiale globale des communes du portefeuille. Puis, après avoir découpé le territoire en zones géographiques homogènes grâce à des méthodes de classification. Enfin, nous avons étudié la structure de corrélation locale dans chaque région.

L'idée est donc maintenant de lisser cette tendance spatiale. En effet, il semble raisonnable de penser que le risque climatique d'une commune dépend non seulement de sa location, mais aussi des lieux environnants. Le lissage spatial permettra de lutter contre les variations locales de Loss-Cost brusques. Il s'agit de capter l'information des communes voisines en supposant que les communes proches les unes des autres ont plus de chances d'avoir un risque spatial similaire que celles qui sont éloignées.

III.1 Présentation du modèle de régression locale

La régression locale (LOESS ou LOWESS) est une méthode de régression non paramétrique reposant sur la méthode des k-plus proches voisins. L'un des avantages de ce type de régression est qu'elle est simple d'utilisation. Elle permet en effet aux données de « parler pour elles-mêmes ». Elle a été créée par Cleveland en 1979 sous le nom de LOWESS (**LO**cally **W**eighed **S**catterplot **S**moothing) puis généralisée en 1988 par Cleveland et Devlin sous le nom LOESS (**Lo**cally Weighted Regr**ESS**ion). Dans la suite de l'étude, la régression locale fera référence à la procédure LOESS.

1.Généralités théoriques

Le but de la régression LOESS est d'estimer la relation $y_i = f(x_i) + \epsilon_i$ dans laquelle la fonction $f(x_i)$ est inconnue et ϵ_i est le terme d'erreur, représentant les erreurs aléatoires dans les observations. On suppose que les erreurs ϵ_i sont identiques et indépendamment distribuées de moyenne 0 et de variance finie $var(\epsilon_i) = \sigma^2$. L'idée de la régression locale consiste à utiliser un modèle de régression défini uniquement dans le voisinage d'un point x_o d'intérêt. On notera $N(x_o)$ ce voisinage. Ce modèle de régression est basé sur la formule de Taylor-Young :

Soit f une fonction de classe C^n sur un intervalle I et $a \in I$. Alors pour tout réel $x \in I$,

on a : $f(x)=f(a)+\sum_{k=1}^{n-1}\frac{f^{(k)}(a)}{(k-1)!}(x-a)^k+o(|x-a|^n)$. Cette formule montre que l'on peut approcher n'importe quelle fonction de classe C^n par un polynôme. L'inégalité de Jackson nous permet de raffiner cette approximation :

Soit f une fonction de classe C^n sur un segment $[a,b]$. Alors :

$\min_{g\in P_k}\max_{x\in[a,b]}|g(x)-f(x)|\leq C(\frac{b-a}{2k})^n$ avec P_k l'espace linéaire des polynômes de degré n .

Regardons maintenant comment obtenir un lissage LOESS à partir d'un point x_o .

La première étape est de définir une fonction de poids. Nous allons définir cette fonction de poids de telle sorte que seules les valeurs contenues dans la fenêtre de lissage $[x_o+h(x_o);x_o-h(x_o)]$ seront considérées comme une estimation de $f(x_o)$. Des fonctions telles que la fonction tricube de Tukey permettent de répondre à ces exigences :

$$W(u)=\begin{cases}(1-|u|^3)^3, & |u|\leq 1\\ 0, & |u|>1\end{cases}$$

La séquence des poids est alors définie par $w_i(x_o)=W(\frac{x_i-x_o}{h(x)})$. On définit la fenêtre par une procédure similaire aux k plus proches voisins. Dans la fenêtre de lissage, f(x) est approximée par un polynôme. Par exemple, une approximation quadratique :

$f(x)\approx\alpha_0+\alpha_1(x-x_o)+\frac{1}{2}\alpha_2(x-x_o)^2$ pour $x\in[x_o-h(x_o),x_o+h(x_o)]$

Pour une fonction continue, le théorème de Taylor nous permet d'estimer la qualité de cette approximation.

Pour obtenir les estimateurs de la régression locale, on cherche le $\alpha=(\alpha_0,\alpha_1,\alpha_2)'$ minimisant

$\hat{\alpha}=\underset{\alpha\in R^3}{arg\,min}\sum_{i=1}^{n}w_i(x_o)[\boldsymbol{Y}_i-\{\alpha_0+\alpha_1(x-x_o)+\frac{1}{2}\alpha_2(x-x_o)^2\}]^2$ et on définit $\hat{f}(x_o)=\hat{\alpha}_0$

2.Régression locale multivariée

Dans le cas de notre étude, les données sont géospatiales, elles dépendent donc deux

variables indépendantes, à savoir la latitude et la longitude. Pour pouvoir effectuer un lissage spatial de ce genre de données, il est nécessaire de prolonger les caractéristiques de la fonction LOESS aux données multivariées.

Soit un modèle de régression à deux variables :

$Y=f(x_{i1},x_{i2})+\epsilon_i$ avec $f(x,y)$ inconnue.

Autour d'un point $x_o=(x_{o1},x_{o2})$ une approximation locale quadratique sera sous la forme :

$$f(x_1;x_2)=\alpha_0+\alpha_1(x_1-x_{o1})+\alpha_2(x_2-x_{o2})+\alpha_3(x_1-x_{o1})(x_2-x_{o2})+\frac{1}{2}\alpha_4(x_1-x_{o1})^2+\frac{1}{2}\alpha_5(x_2-x_{o2})^2$$

Une fois que l'on définit la distance entre un point x et x_o, et un pas h on peut définir la séquence de poids : $w_i(x_o)=W(\frac{\|x_i,x_o\|}{h})$.

La distance usuelle utilisée pour un espace R^d est $\|x\|^2=\sum_{j=1}^{d}(x_j/v_j)^2$. Un choix naturel pour les v_j est l'écart-type des variables.

III.2 Application du modèle de régression LOESS

1.Préparation des données

Les données utilisées pour la régression sont les loss-cost de 2000 à 2013 des communes présentes dans le portefeuille 2013. Nous prendrons le loss-cost moyen entre 2000 et 2013 pour chaque commune pour deux raisons. Premièrement parce que cela permet de supprimer toute corrélation intertemporelle de l'étude. Les loss-cost pourraient effectivement être corrélés d'une année à une autre sur des communes voisines, mais cela n'est pas considéré essentiel dans l'estimation des valeurs. Les méthodes de lissage spatial ne s'intéressent qu'à la corrélation spatiale. La deuxième raison est que les méthodes non-paramétriques ont été conçues en supposant que l'échantillon que l'on désire étudier ne contient pas d'égalités pour les variables explicatives x_1 et x_2 . Pour remédier à ce problème, on prend donc la moyenne de la variable réponse et on accorde à la nouvelle coordonnée un poids w_i^* égal au nombre de valeurs pour le couple $(x_{1i};x_{2i})$, c'est-à-dire le nombre de fois que la commune a été présente dans le portefeuille depuis 2000.

Une deuxième transformation est effectuée au niveau des loss-cost. Il s'agit d'une

transformation logit du type $z=\log\left(\frac{y}{1-y}\right)$. L'idée est d'augmenter la qualité de la régression. En effet, les Loss-Cost sont compris entre 0 et 1 et sont généralement plus proches de 0. Cette transformation permet de rendre la distribution des loss-cost moins dissymétrique vers la gauche et permet d'assurer que les loss-cost estimés ne seront pas négatifs.

2.Paramètres de la fonction LOESS

La procédure LOESS est une procédure non-paramétrique dans le sens où le praticien ne spécifie pas la forme final de la courbe (ou en l'occurrence de la surface) de lissage. Cependant, certains paramètres doivent être implémentés pour assurer une bonne qualité de lissage.

D'abord, le paramètre de lissage α . Ce paramètre définit la largeur de la fenêtre de lissage définie précédemment. Il est défini entre 0 et 1. Un $\alpha=1$ indique que toutes les valeurs du modèle sont utilisées dans les régressions locales. Plus%alpha sera élevé, plus la courbe sera lissée. En effet, plus la fenêtre de lissage est grande, plus il y a d'observations prises en compte. De fait, chaque observation aura donc proportionnellement moins d'influence sur la régression. De plus, une plus grande fenêtre de lissage implique qu'il y aura moins d'observations qui changeront en passant d'une fenêtre à une autre. Ces deux facteurs tendent à stabiliser les courbes de régression locales. Malheureusement, il n'existe pas de méthode absolue permettant de déduire un α optimal pour la méthode LOESS. L'idée va être de comparer des courbes avec des α différents et en déduire le plus adapté.

Le deuxième paramètre est le degré du polynôme de régression, λ . Si $\lambda=1$, les polynômes de régression sont des droites. Si $\lambda=2$, des équations quadratiques sont utilisées. Celles-ci compliquent le processus de lissage mais permettent généralement d'obtenir des courbes de lissage ayant une adéquation aux données satisfaisante. Normalement, il est possible de donner le degré que l'on veut au polynôme de régression. Néanmoins, cela est extrêmement gourmand en tant de calcul. De plus, le logiciel R ne nous permet pas de dépasser le degré 2. C'est ce degré que nous utiliserons dans la suite de notre étude.

3.Mise en œuvre de la procédure

On va appliquer la régression LOESS à chaque zone géographique créée précédemment afin de bien appréhender les effets d'autocorrélation spatiale. On prendra pour exemple la région 4.

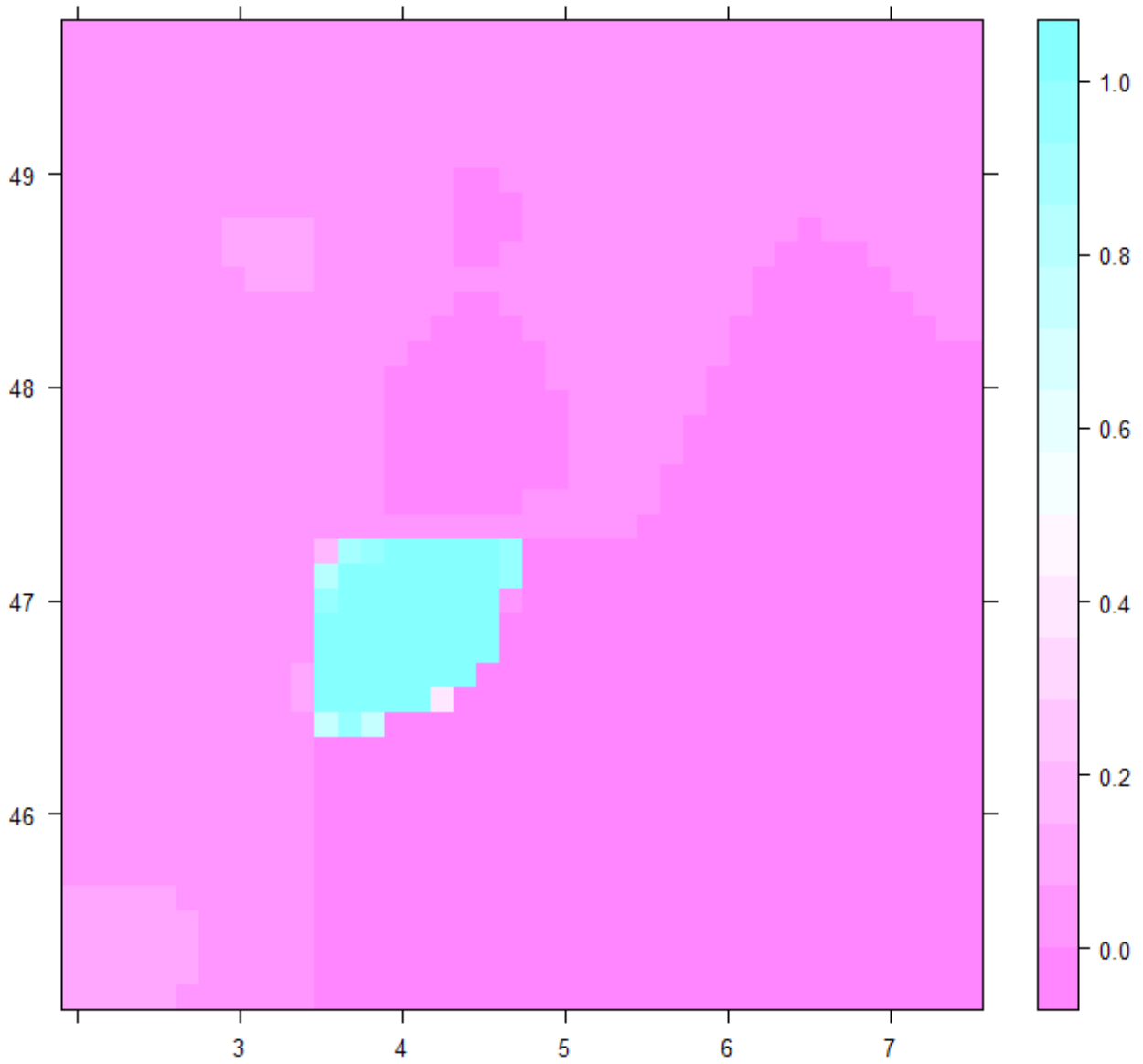


Figure 16 : Lissage spatial local pour la région 4

Nous avons testé différentes valeurs de α et nous avons choisi $\alpha = 0{,}75$. Cette valeur permet d'avoir un lissage spatial fin, tout en conservant les grandes structures spatiales. La figure 16 nous montre directement une des limites de la méthode LOESS. Il est assez difficile d'interpréter les résultats en 3 dimensions. Néanmoins on peut remarquer un fort cluster de Loss-Cost élevés autour du point de coordonnées (4,47). La carte des indices de Moran Locaux montre qu'il y avait de fortes corrélations spatiales positives dans cette zone. Cependant la Figure 17 montre que s'il y a

effectivement un pic de Loss-Cost, il n'est pas aussi élevé que dans le lissage spatial.

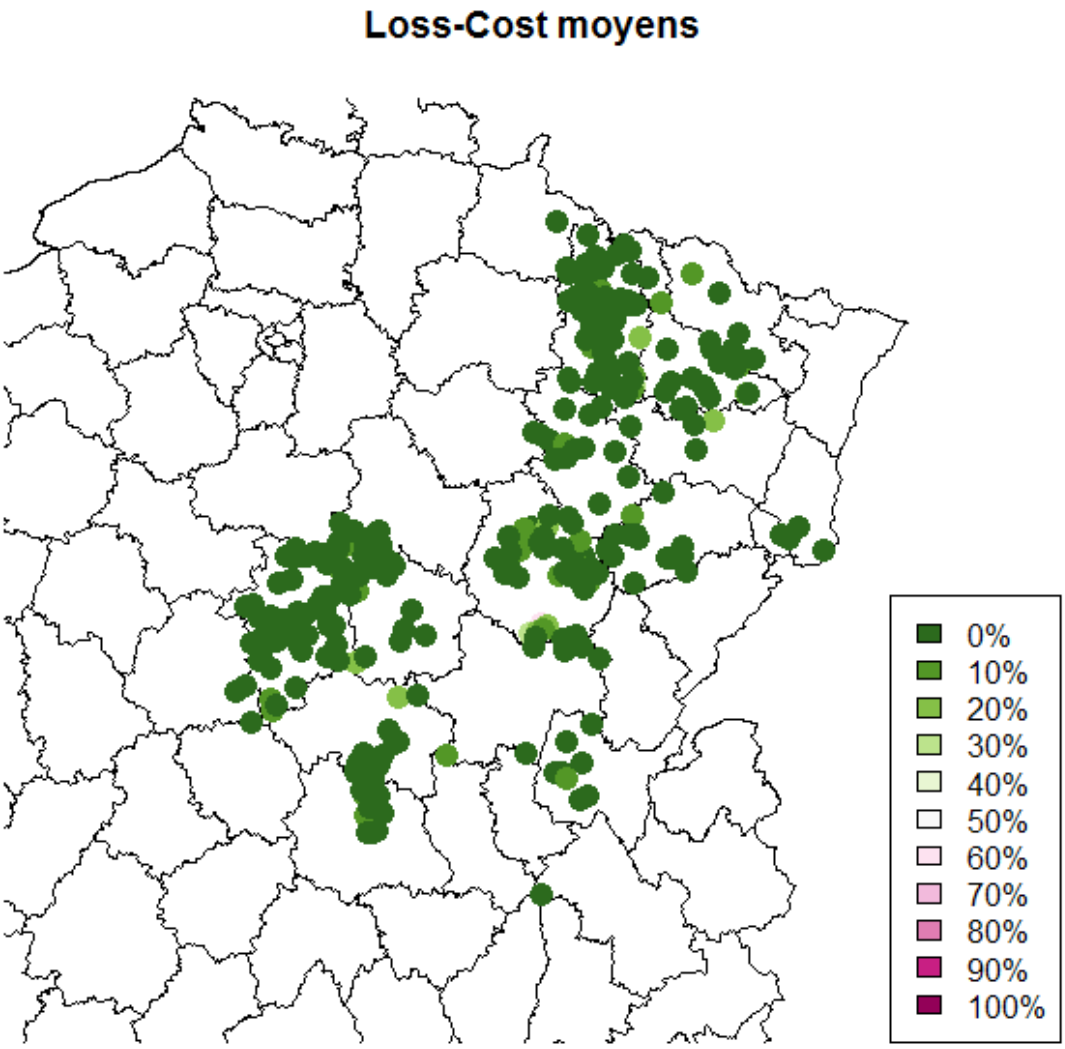


Figure 17: Loss Cost moyens entre 2000 et 2013 pour la région 4

Ce décalage entre le lissage et la réalité pourrait provenir du fait que les données sont très éparpillées dans l'espace. De fait, la fenêtre de lissage doit être assez large pour estimer un nombre suffisant de valeurs, ce qui est contradictoire avec le caractère local de la régression.

4.Validation des hypothèses

La méthodologie de la procédure LOESS se base sur des hypothèses importantes. L'une d'entre-elles est que les erreurs ϵ_i sont indépendantes et identiquement distribuées. Il convient de vérifier cette hypothèse. La procédure LOESS n'est pas une méthode paramétrique. Cependant, on peut appliquer les diagnostics utilisés en régression linéaire. Pour cela, le moyen le plus simple est de tracer un QQ-plot des résidus.

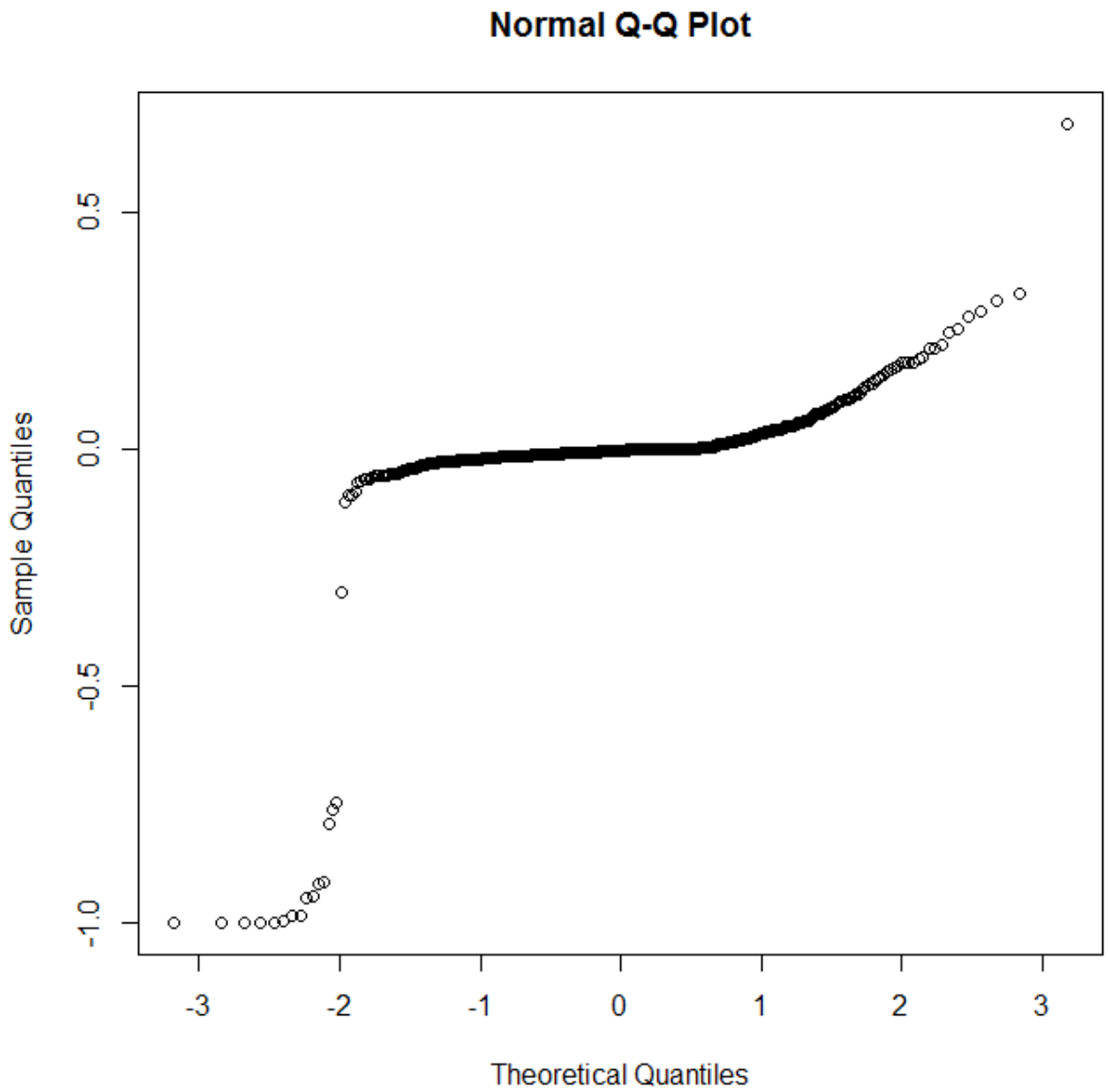


Figure 18 : QQ-plot des résidus pour la région 4

Le QQ-plot nous montre que l'hypothèse de normalité des résidus n'est pas respectée. En effet, dans un cas de normalité, les points devraient se positionner suivant la première diagonale.

Cela nous permet de comprendre un peu mieux ces pics de Loss-Cost dans la figure 16. Le modèle doit être trop sensible aux valeurs hautes et pas assez aux valeurs moyennes. L'idée pour corriger cela serait de retravailler sur la valeur de α ou de tester si un polynôme de degré 3 aurait ferait office de meilleur ajustement.

De plus, on peut remarquer que même dans le cas d'un modèle fiable, on peut quand même constater que la surface de la fonction LOESS n'a pas d'expression définie. De fait, il serait compliqué de pouvoir exporter le modèle, ce qui est une limite significative pour ce type de régression. Des modèles de lissage tels que les splines de régression pourraient pallier à ce genre de difficultés.

Conclusion

L'objectif de ce mémoire était l'étude spatiale des répercussions de phénomènes climatiques sur un portefeuille de territoires agricoles. On partait en effet de l'hypothèse qu'il y avait une dépendance spatiale entre les communes. C'est-à-dire que l'information d'un sinistre dans une commune se diffusait dans les communes environnantes. Il s'agissait donc dans un premier temps de quantifier cette information, cette corrélation.

Pour se saisir de cette information, nous avons utilisé des techniques d'analyse exploratoire spatiale. Les indices de Moran, globaux et locaux, ainsi que ceux de Geary nous ont permis d'appréhender cette diffusion d'information d'une commune à son voisinage, d'abord au niveau global puis localement. L'utilisation de la classification ascendante hiérarchique pour créer des zones géographiques homogènes nous a permis d'affiner l'étude des corrélations entre les communes du portefeuille. On peut néanmoins citer deux limites à cette analyse. Premièrement, lorsque nous avons séparé le territoire en plusieurs zones, nous n'avons pas pris en compte les dépendances spatiales entre ces zones, notamment les effets de bords. En effet, certaines de ces zones avaient des frontières communes, ou, du moins, des communes d'une zone très proches des communes d'une autre zone. Les corrélations spatiales de ces communes n'ont pas été calculées. La deuxième limite de cette analyse est que, par définition de l'analyse exploratoire, elle ne permet pas d'appréhender le risque de manière concrète.

Pour remédier à ce problème, nous avons essayé d'effectuer un lissage spatial des loss-cost. Ce lissage devait nous permettre de appréhender les zones de risque. La méthode employée a été la régression LOESS, qui a été choisie pour sa simplicité à mettre en œuvre. Les résultats ont malheureusement été décevants. En effet, nous n'avons pas réussi à obtenir un lissage fiable. Cela est sûrement dû à un mauvais choix des paramètres de lissage.

Cela étant dit, ce modèle pourrait tout à fait être la base d'une nouvelle tarification du traité. En supposant que nous réussissions à effectuer un lissage fiable des Loss-Cost, nous pourrions utiliser des méthodes de simulation de sinistres pour simuler l'exposition de Liberty Specialty Markets par rapport à ce lissage spatial.

Bibliographie

Anselin L., 1995, « Local indicators of spatial association – LISA ». Geographical Analysis, 27, 93-115.

Andriantavy C., 2008, « Tempêtes : Prise en compte de la dépendance spatiale avec la théorie des copules », Mémoire d'actuaire

APREF, 2014, « Développement du Marché Assurance et Réassurance Récoltes en France »

Roger S. Bivand, Pebesma E., Gómez-Rubio V., 2013, « Applied Spatial Data Analysis with R »

Bouzidi H., 2014, « Mesure du risque pour un portefeuille d'assurance entreprise soumis aux catastrophes naturelles », Mémoire d'actuaire

Caloz R., Collet C., 2011, « Analyse spatiale de l'information géographique »

D. Chessel, S. Ollier & S. Dray, 2003, « Ordination sous contrainte spatiale »

William S. Cleveland; Susan J. Devlin, 1988, « Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting », *Journal of the American Statistical Association*, Vol. 83, No. 403. (Sep., 1988), pp. 596-610.

Cohen-Salmon L., Ginghaye Fongang D., 2014, « Modélisation du risque gel en France », Mémoire d'actuaire

Descheemaekere D., Perron B., 1996, « Les événements naturels en France : étude du risque tempête pour une compagnie d'assurances », Mémoire d'actuaire

Droesbeke J.-J., Lejeune M., Saporta G., 2006, « Analyse statistique des données spatiales »

Manfred M. Fisher, Hank J. Schollten, David Urwin, 2005, « Spatial Analytical Perspectives on GIS»

Gezici F., 2003, « Spatial analysis of regional inequalities in Turkey »

Darrell L. Hueth,William H. Furtan, 1994, « Economics of Agricultural Crop Insurance: Theory and Evidence »

Hurlin C., 2007, « Econométrie et Statistique Non Paramétrique (Partie II). Régressions Non Paramétriques Univariées »

Rafael A. Irizarry, 2003, « Applied Nonparametric and Modern Statistics »

William G. Jacoby, 1999, « Loess: a nonparametric, graphical tool for depicting relationships

between variables »

Lacoste P., 2013, « Réassurance »

Mathis J., 2009, « Elaboration d'un zonier en assurance de véhicules par des méthodes de lissage spatial basées sur des simulations MCMC », Mémoire d'actuaire

Morency C., 2006, « Étude de méthodes d'analyse spatiale et illustration à l'aide de microdonnées urbaines de la Grande Région de Montréal », les Cahiers Scientifiques du Transport N° 49/2006 - Pages 77-102

Nakache J.-P., Confais J., 2004, « Approche pragmatique de la classification : Arbres hiérarchiques, Partitionnements »

Oliveau S., 2005, « Les indicateurs locaux d'association spatiale (LISA) comme méthode de régionalisation »

Poon J., 2011, « Spatial cross-sectional credibility models with general dependance structure among risks », Mémoire

Pothon A., 2013, « Le risque inondation en France Métropolitaine », Mémoire d'actuaire

Sanseverino-Godfrin V., 2009, « L'evolution du cadre juridique des risques naturels et industriels »

Sapporta G., 2011, « Probabilités, analyse des données et statistique »

Frank F. Schnapp, James L. Driscoll, Gary R. Josephson, « Ratemaking Considerations for Multiple Peril Crop Insurance »

Smith, T.E., 2014, « Notebook on Spatial Data Analysis »

M. Tivadar, Y. Schaeffer, A. Torre, F. Bray, « OASIS - un Outil d'Analyse de la Ségregation et des Inégalites Spatiales »

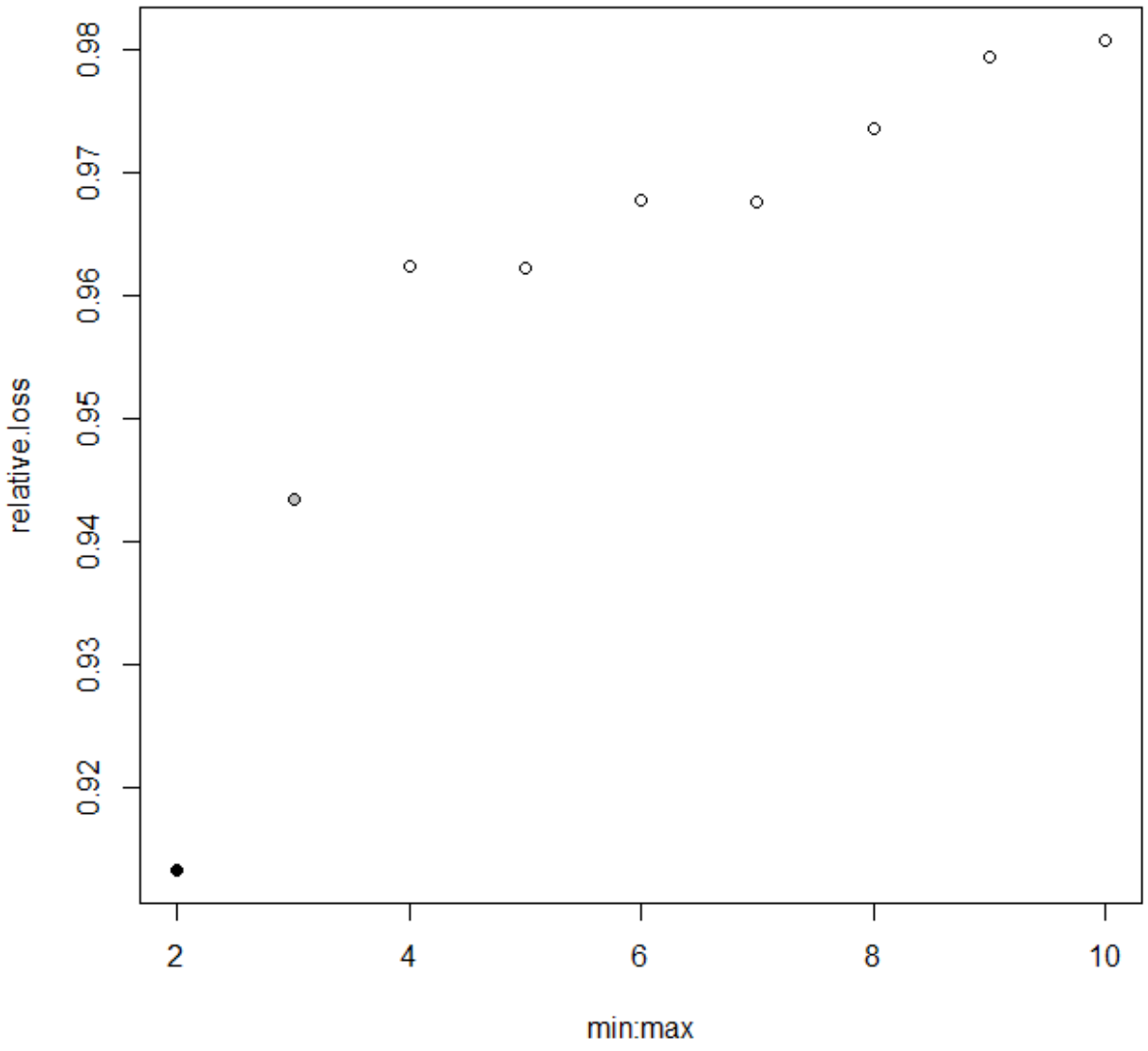
Valat C., 2008, « Tarification des traités catastrophes au Canada », Mémoire d'actuaire

Vandal N., 2005, « La régression non paramétrique multidimensionnelle Théorie et application à une étude portant sur la densité mammaire », mémoire

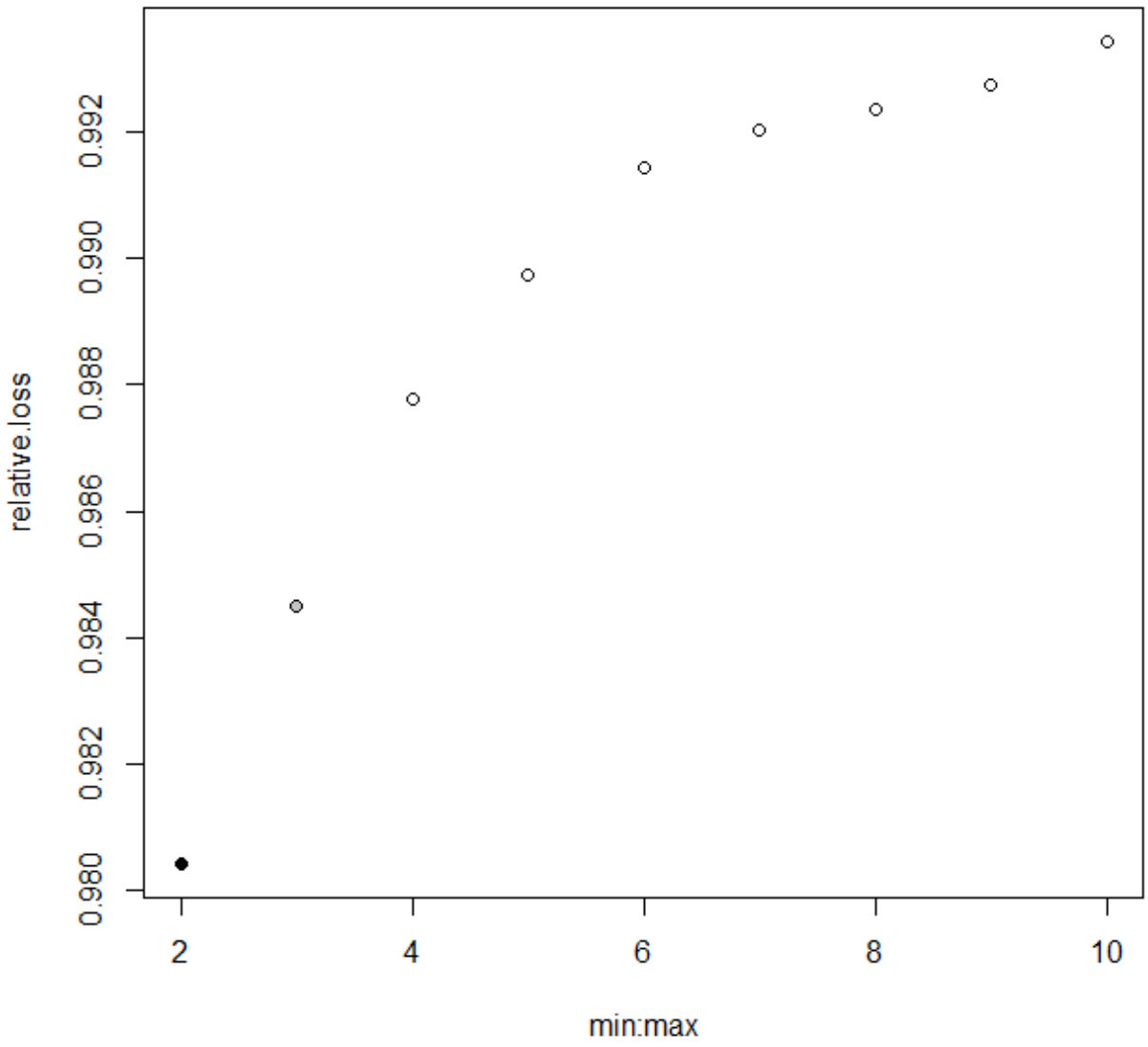
Yeatman J., 1998, « Manuel international de la réassurance », Economica, Paris, 1998, 380 p

Annexes

A.Inerties relatives pour la classification ascendante hiérarchique avec les méthodes de Ward et du diamètre

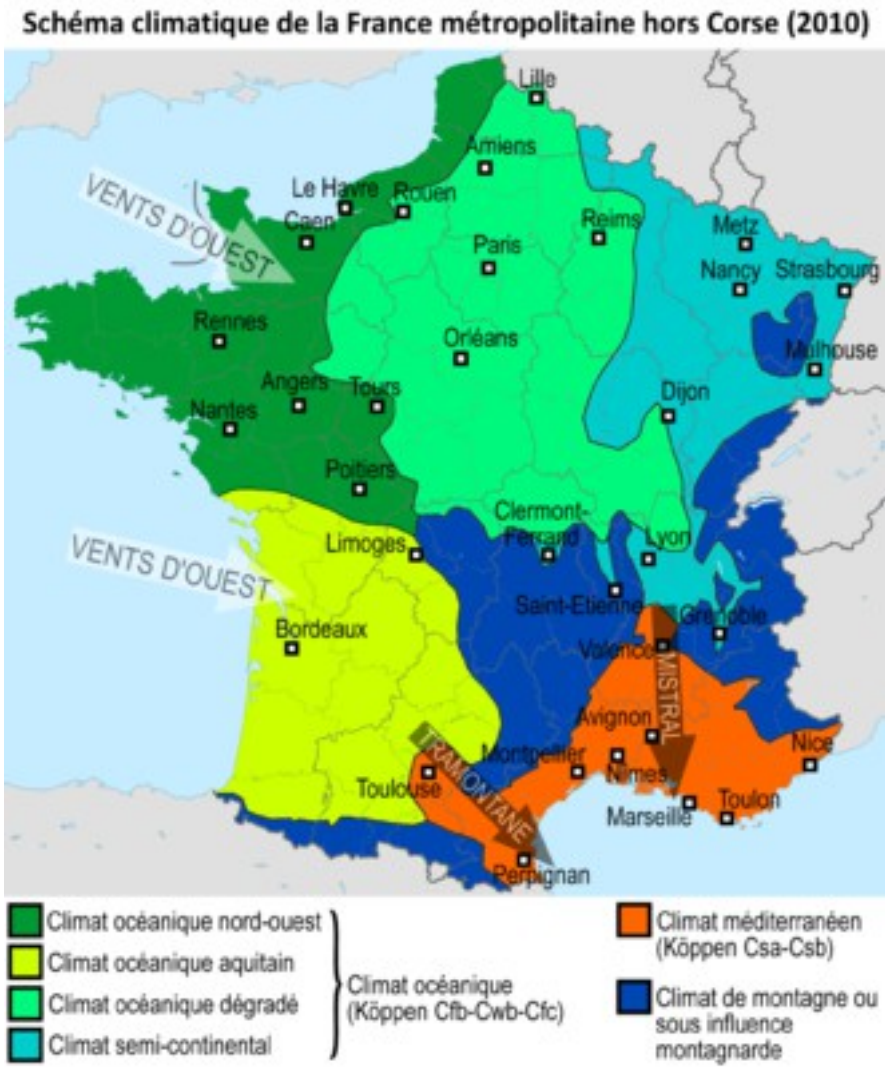


Méthode de Ward



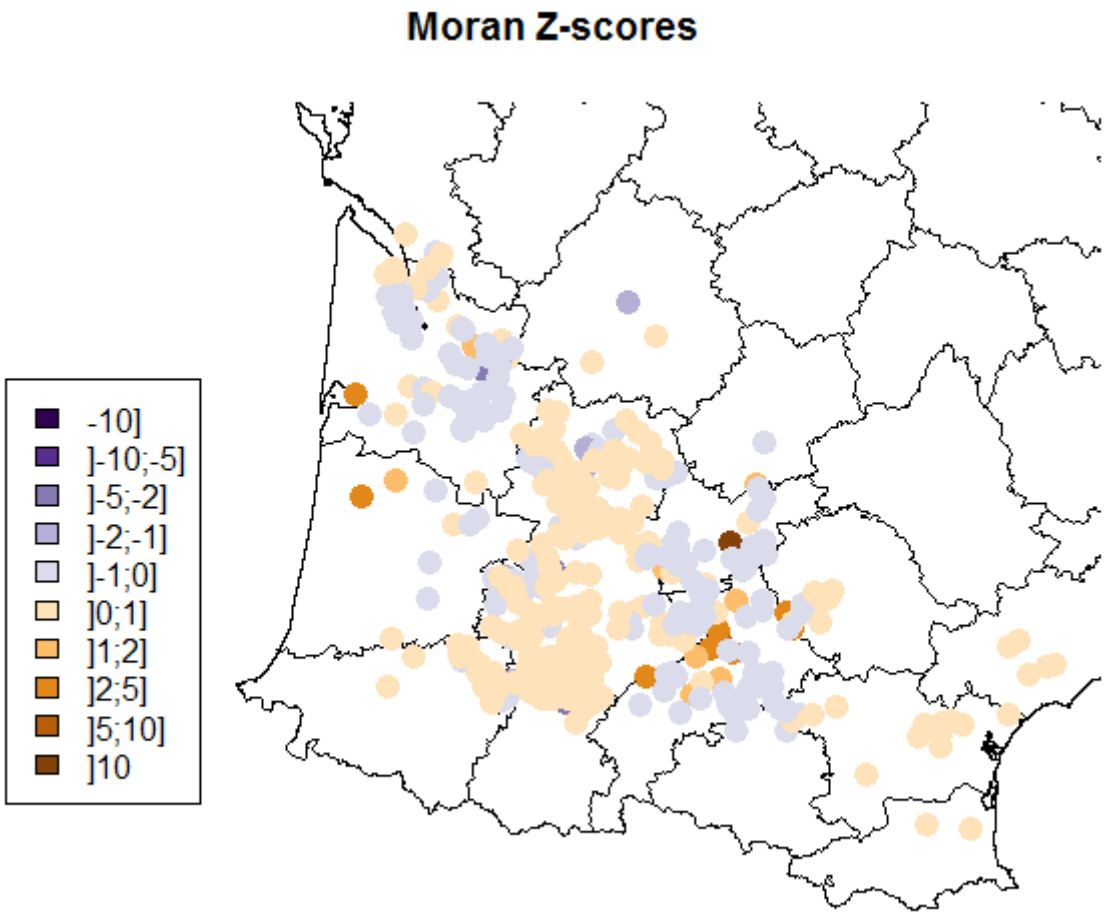
Méthode du diamètre

B.Carte des climats de la France

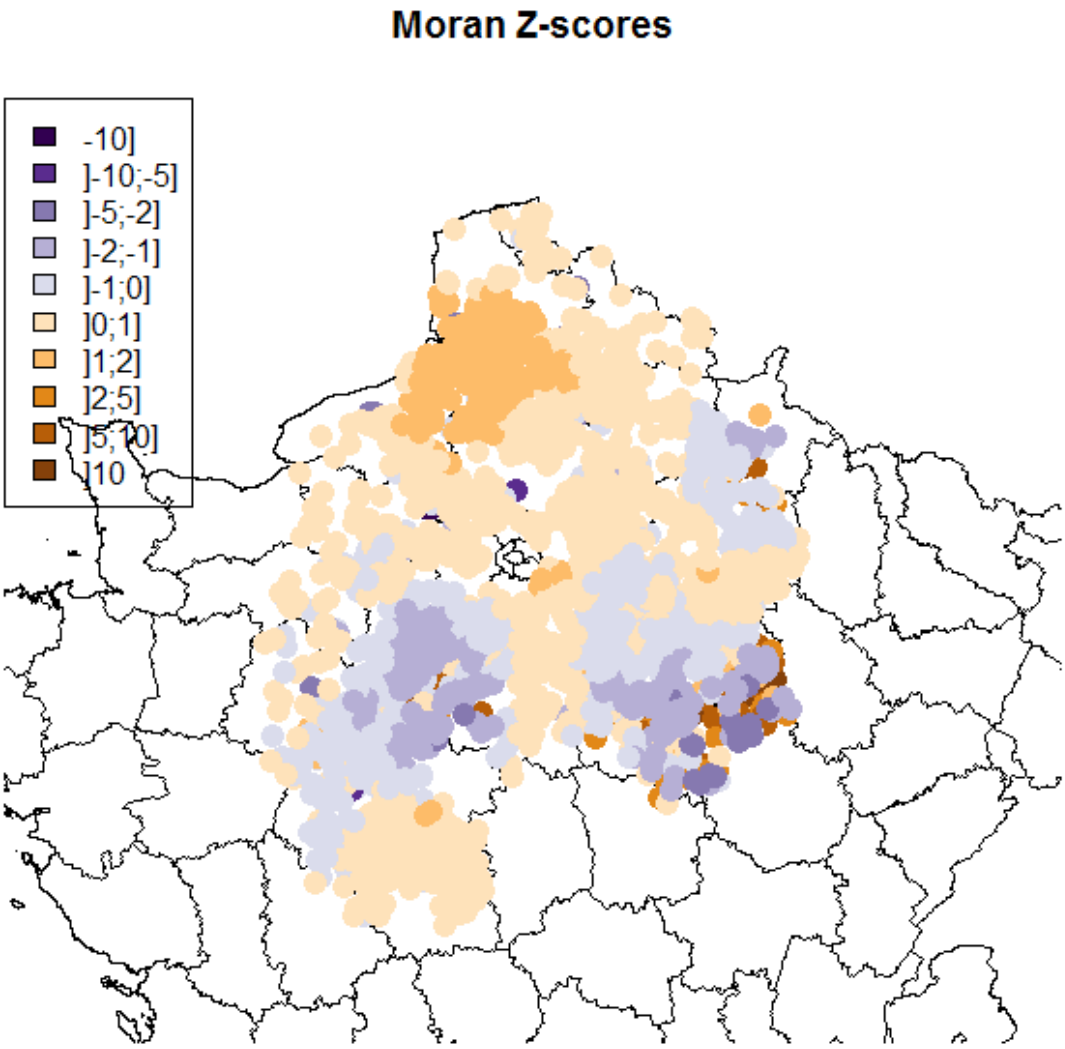


Source Wikipedia

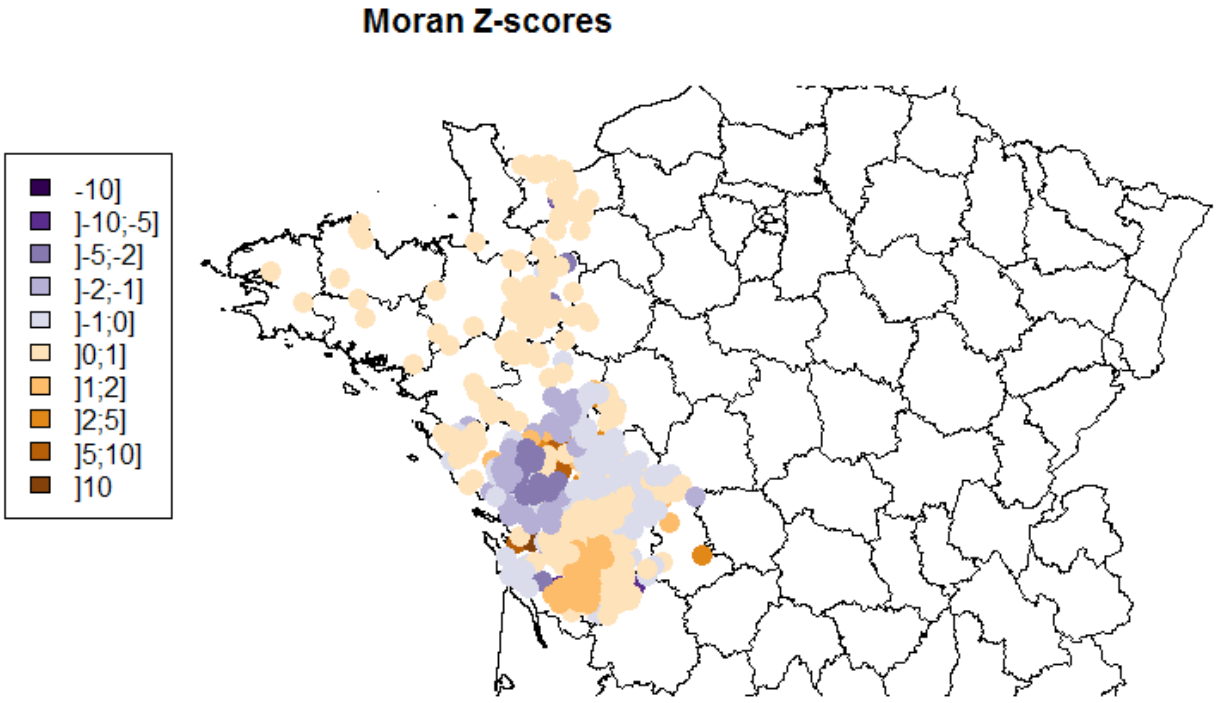
C.Cartes des LISA pour les régions 1,2,3 et 5



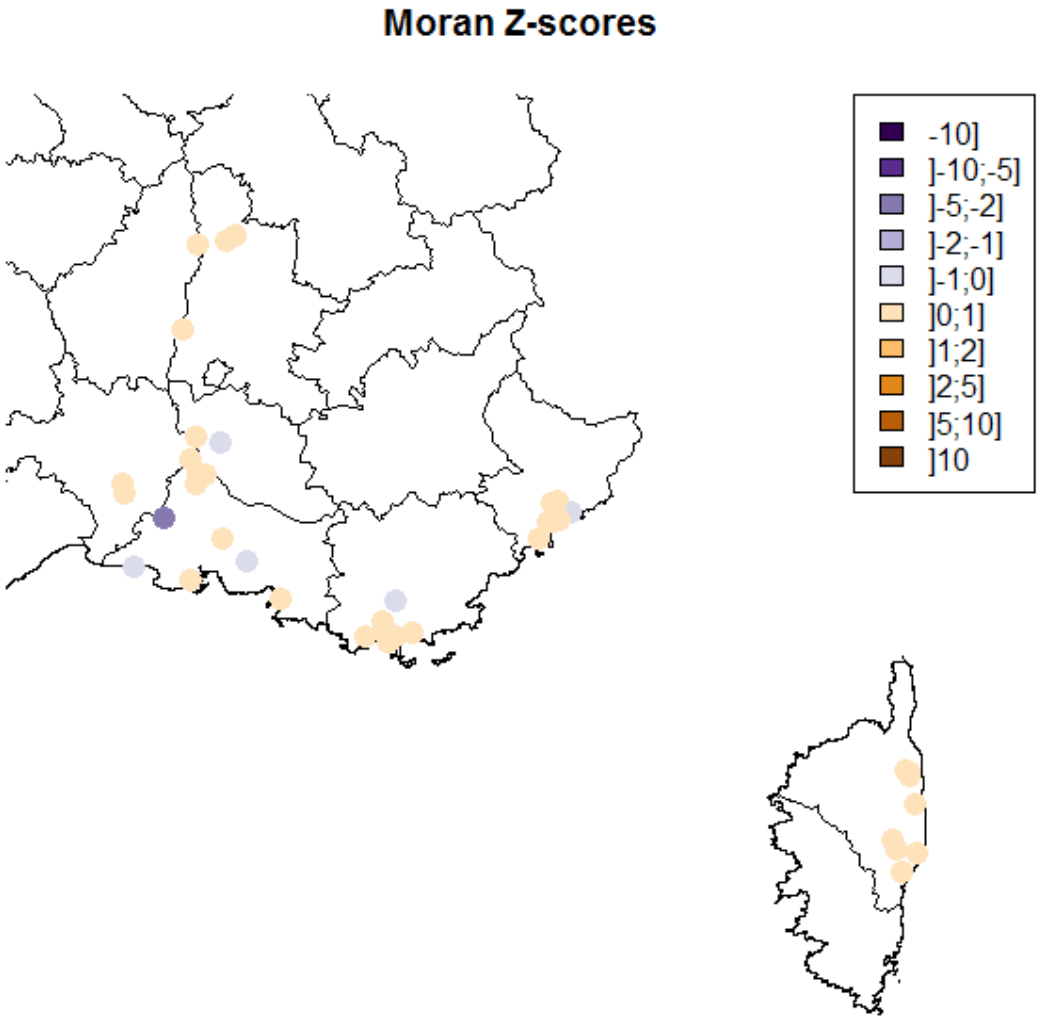
Région 1



Région 2

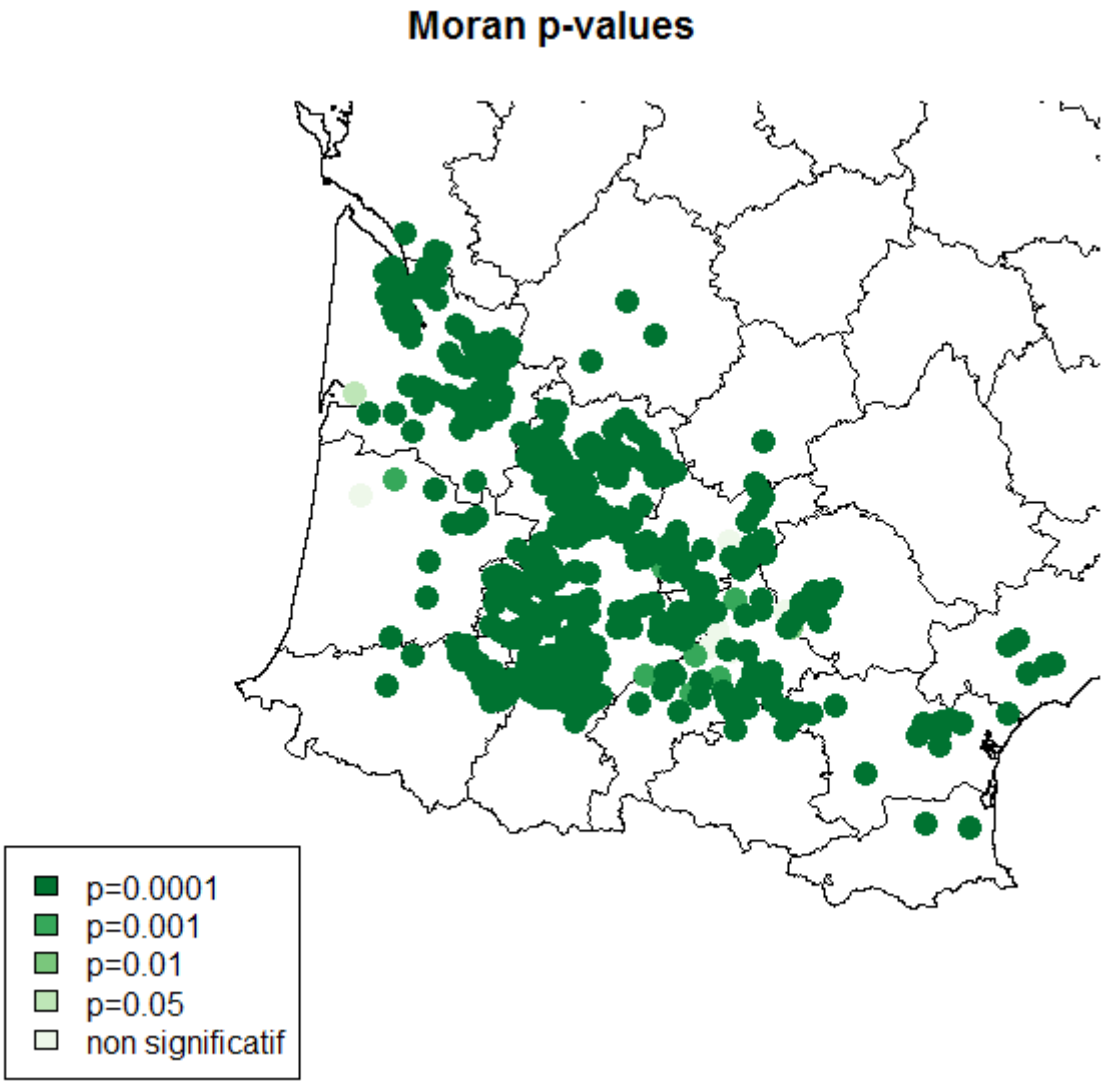


Région 3

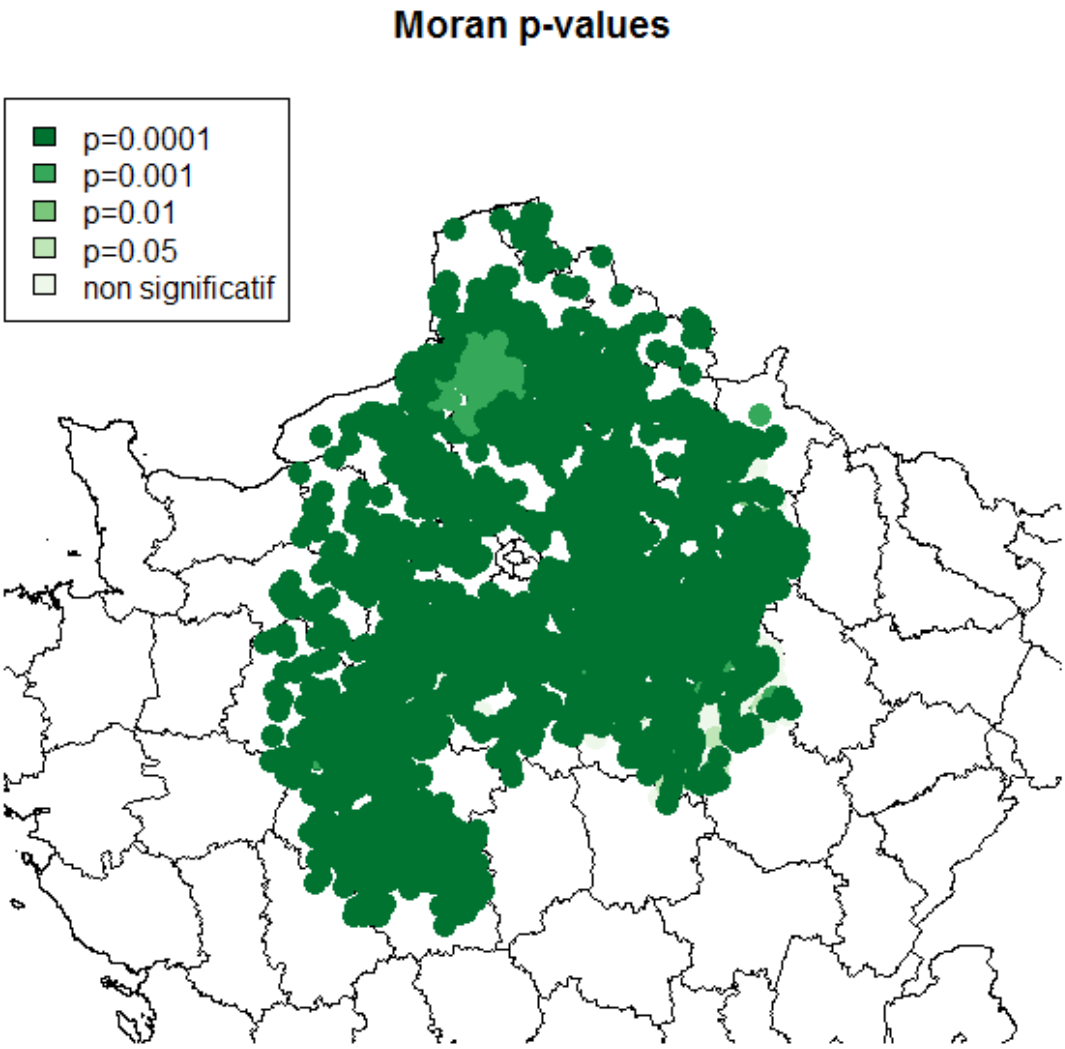


Région 5

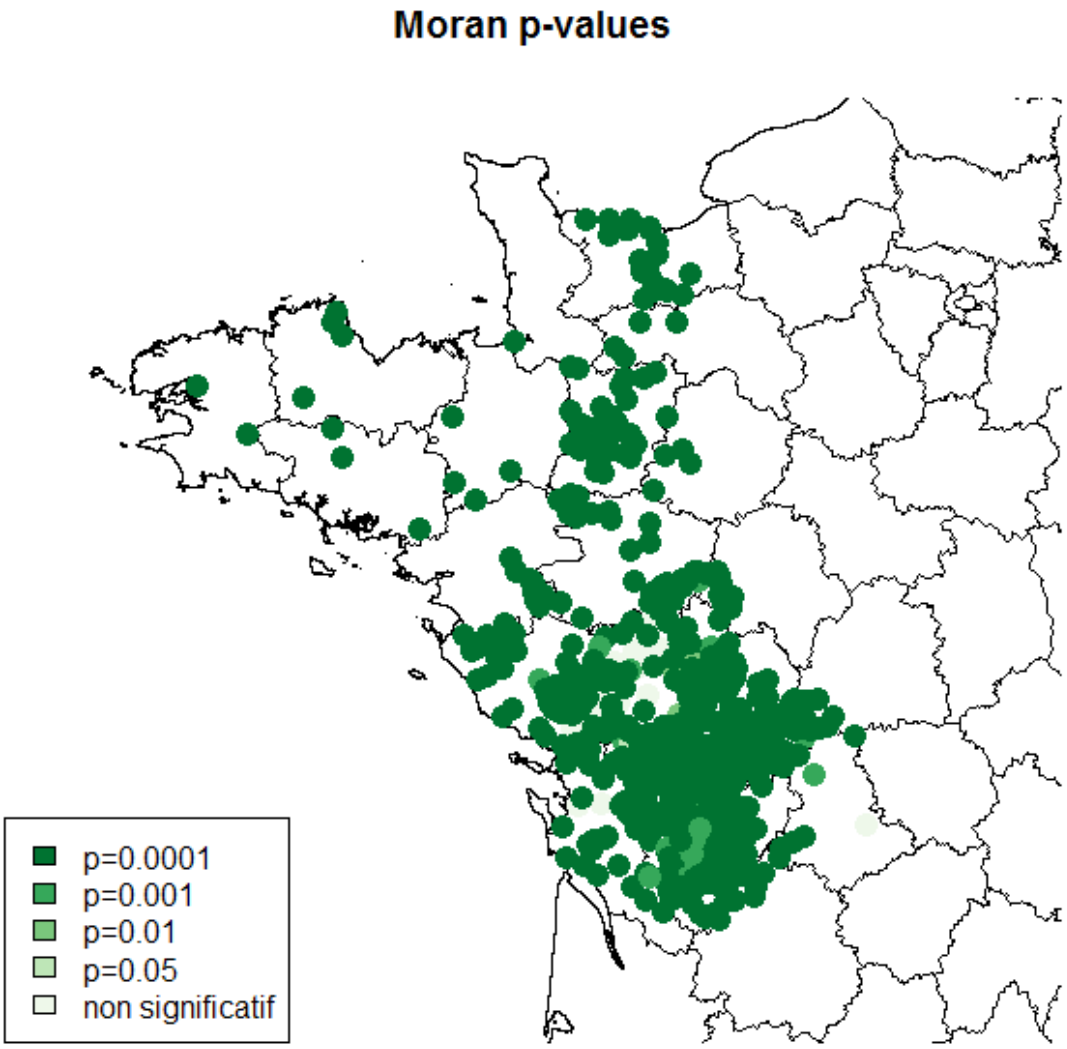
D.Carte de significativité des LISA pour les régions 1,2,3 et 5



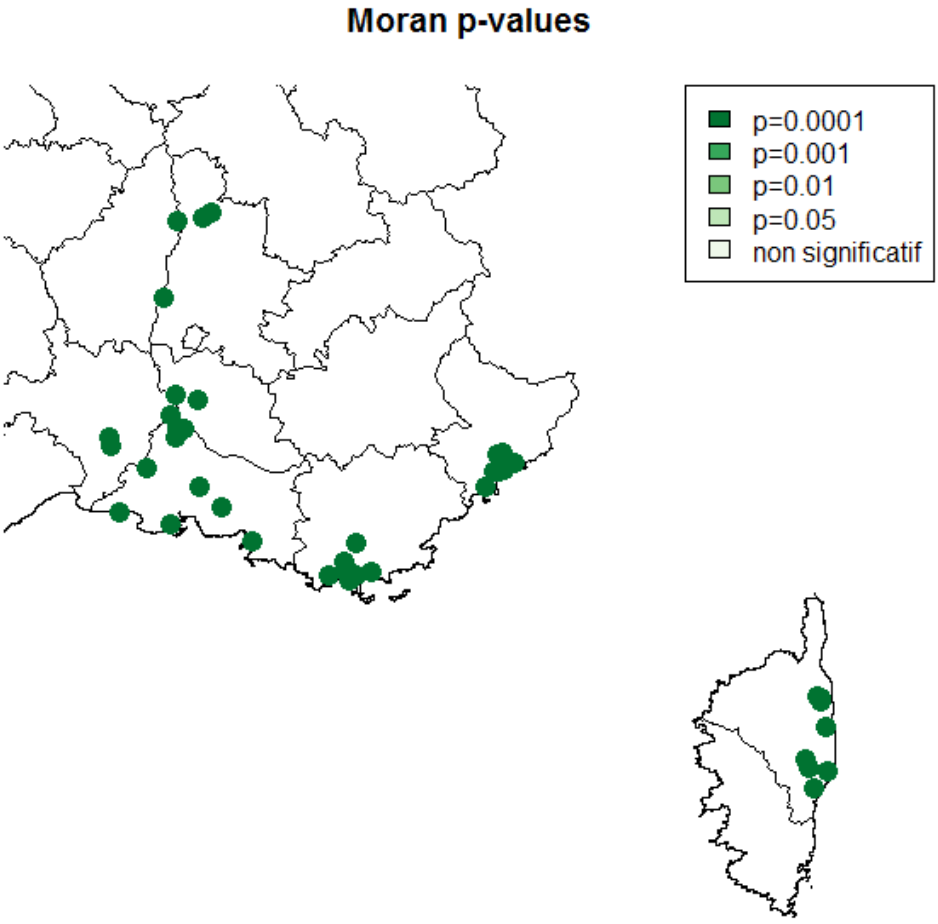
Région 1



Région 2

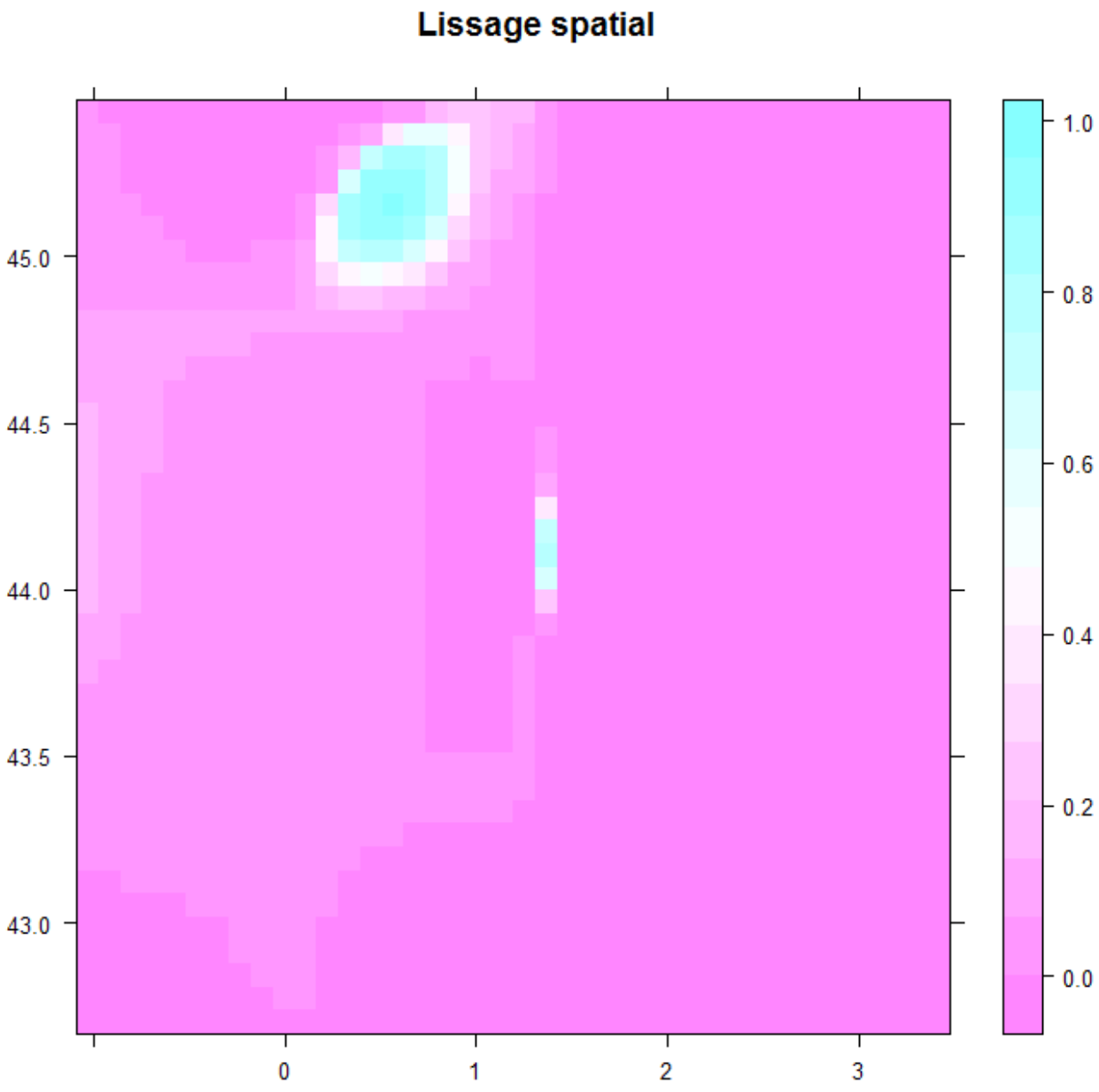


Région 3

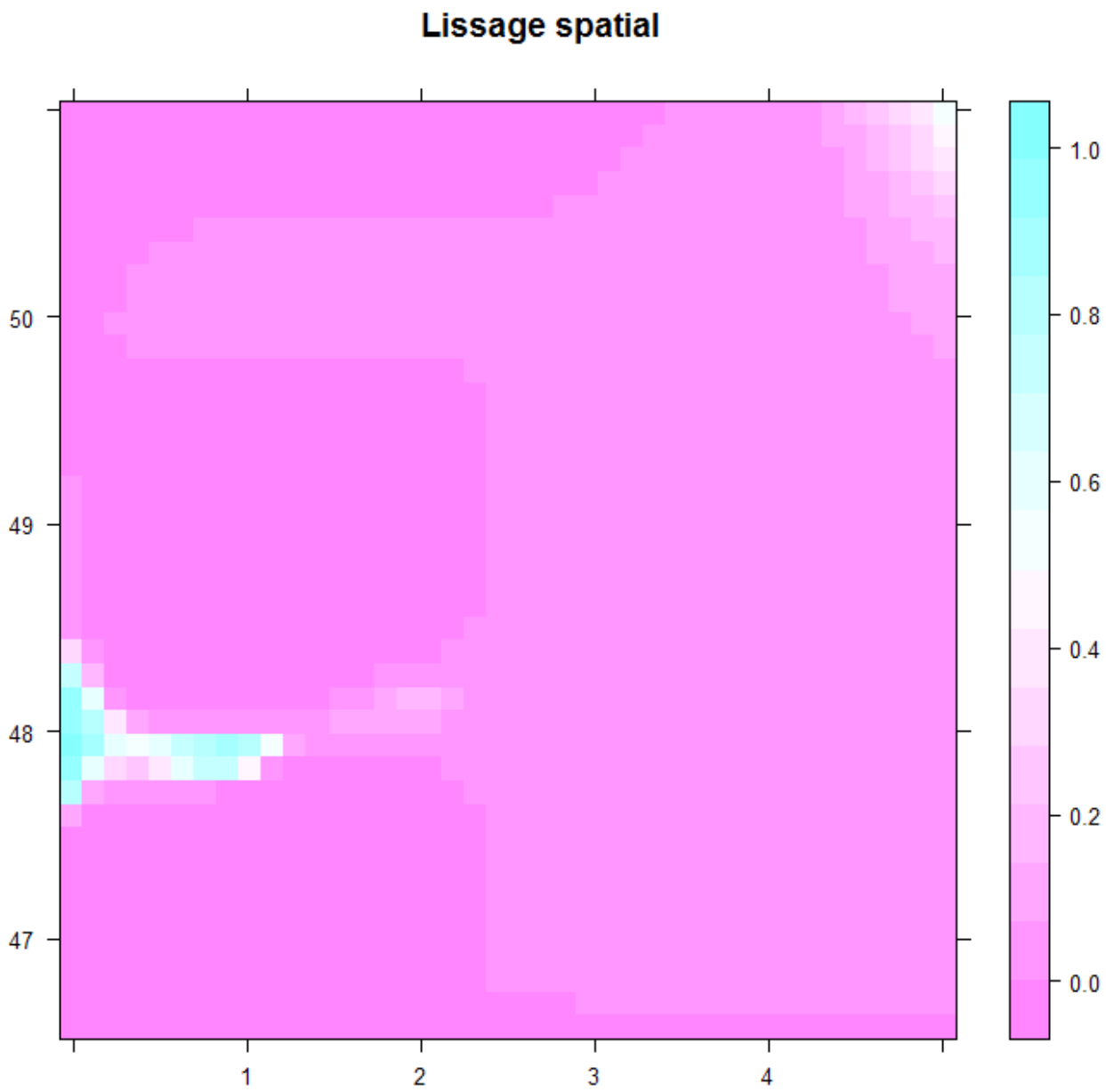


Région 5

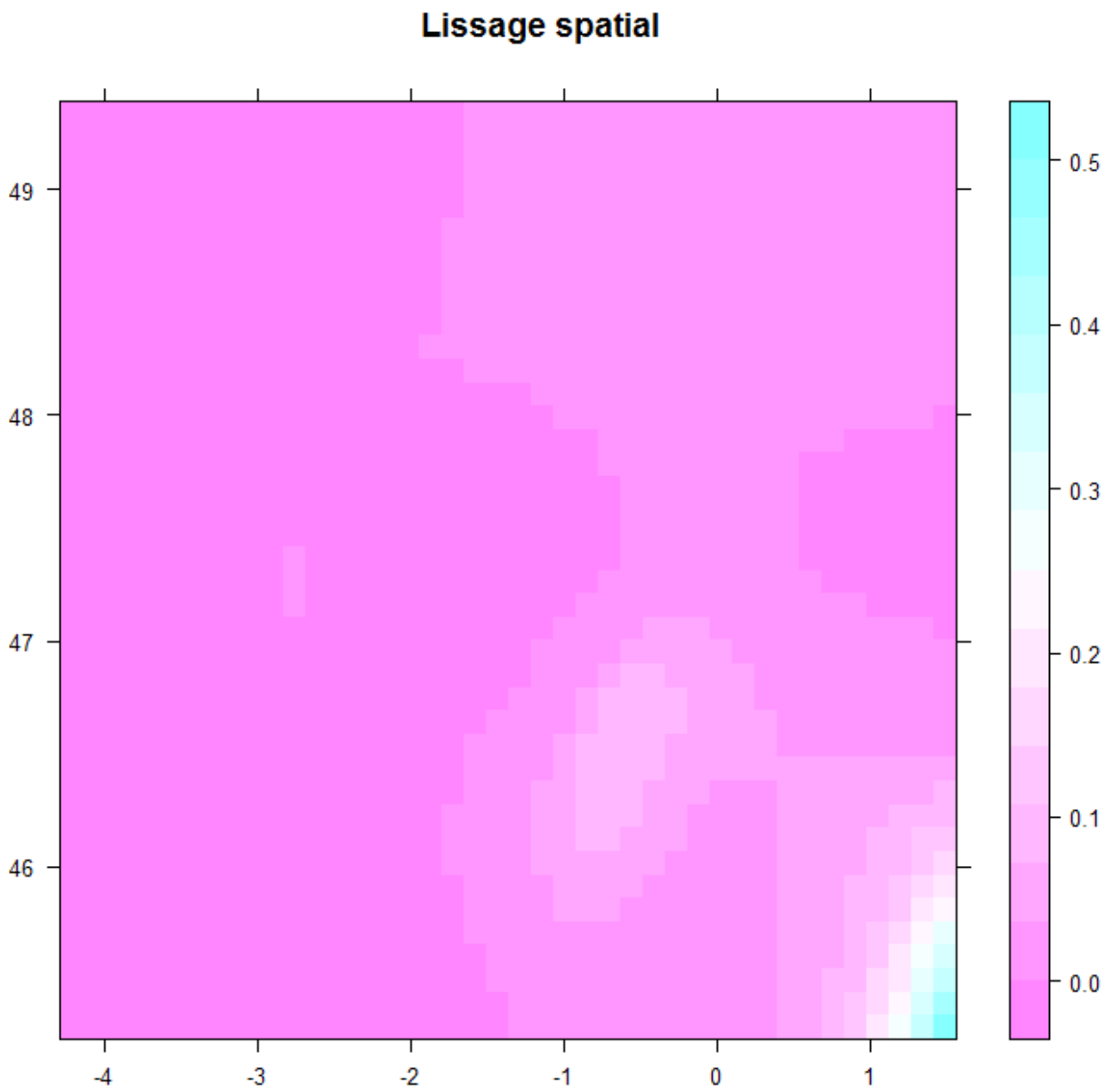
E.Lissages spatiaux des régions 1,2,3 et 5



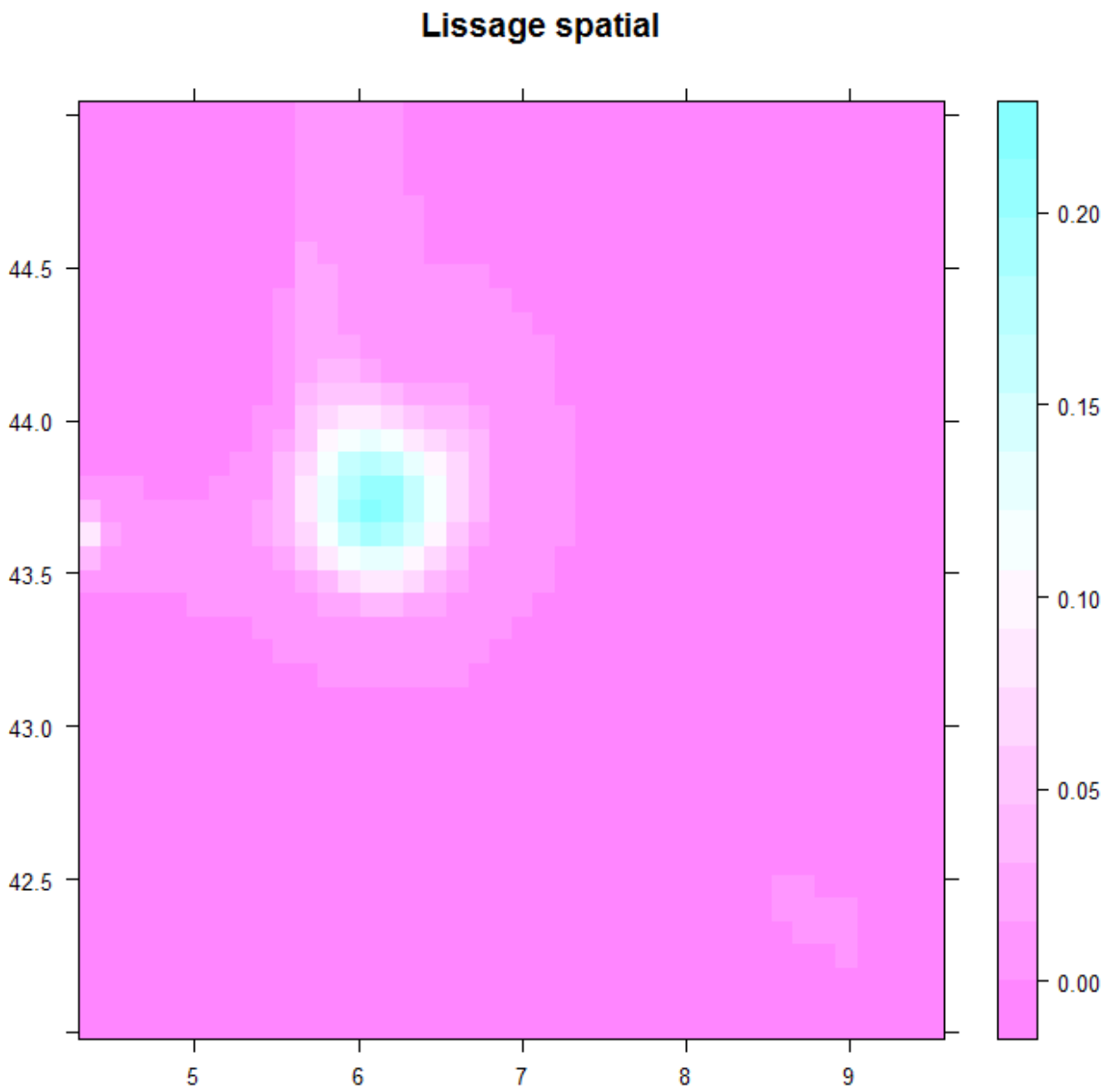
Région 1



Région 2

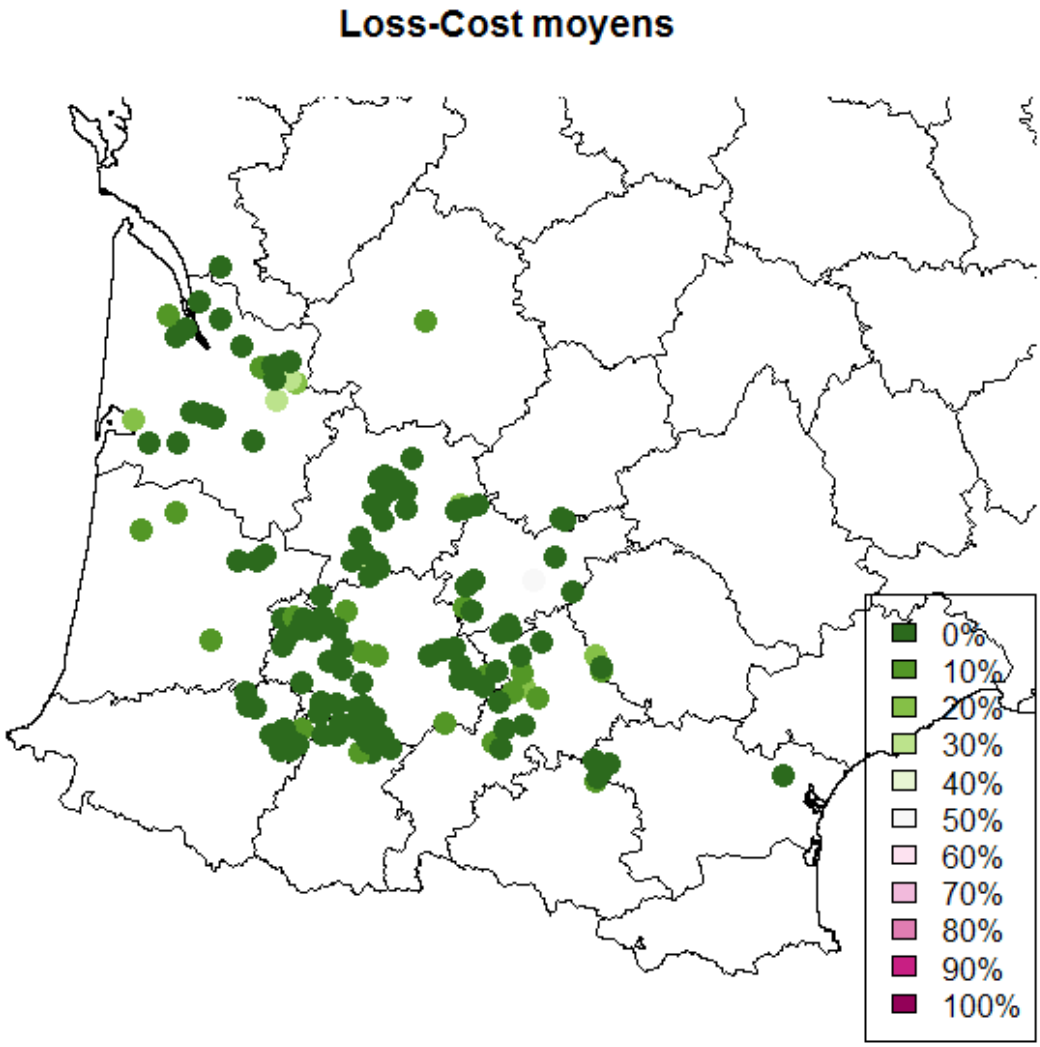


Région 3

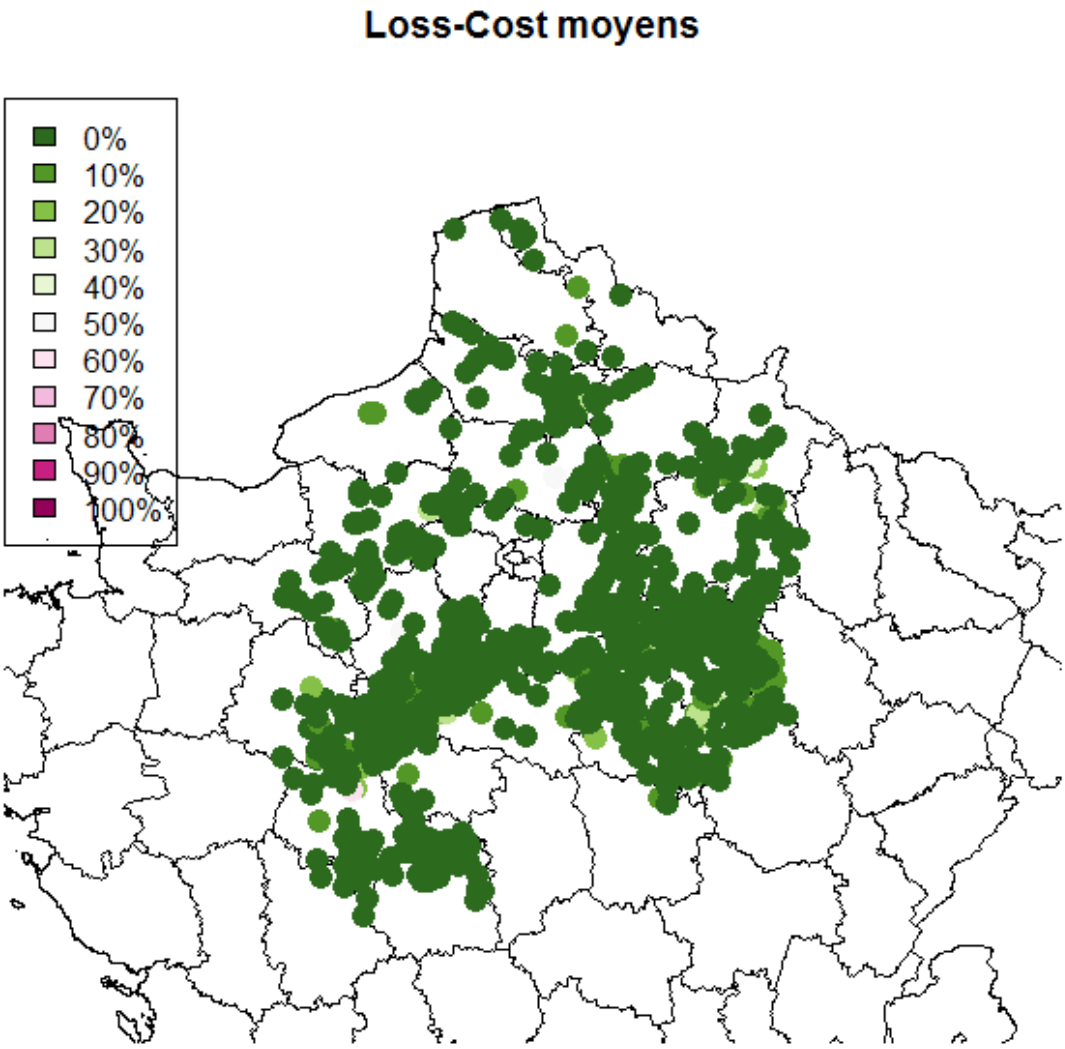


Région 5

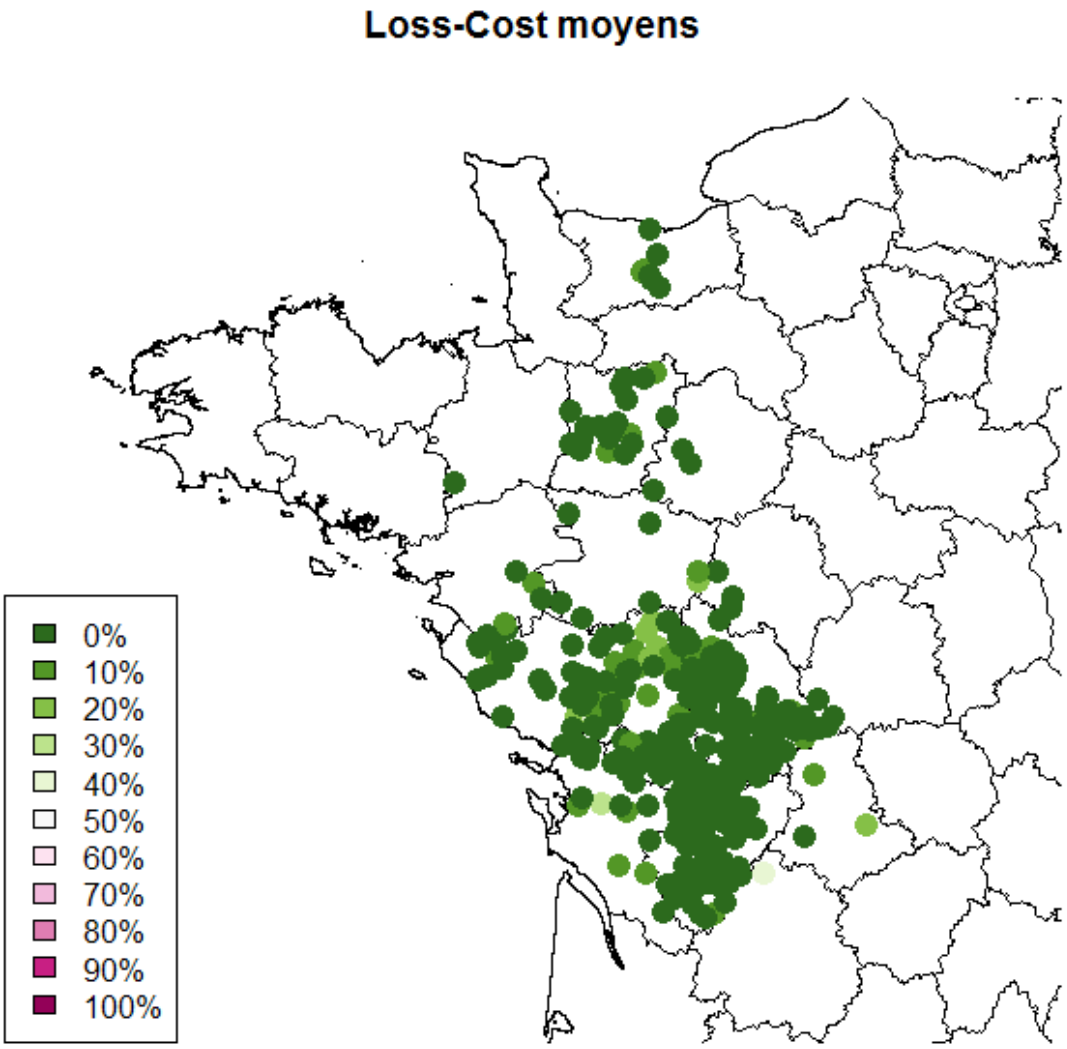
F.Loss-costs moyens pour les régions 1,2,3 et 5



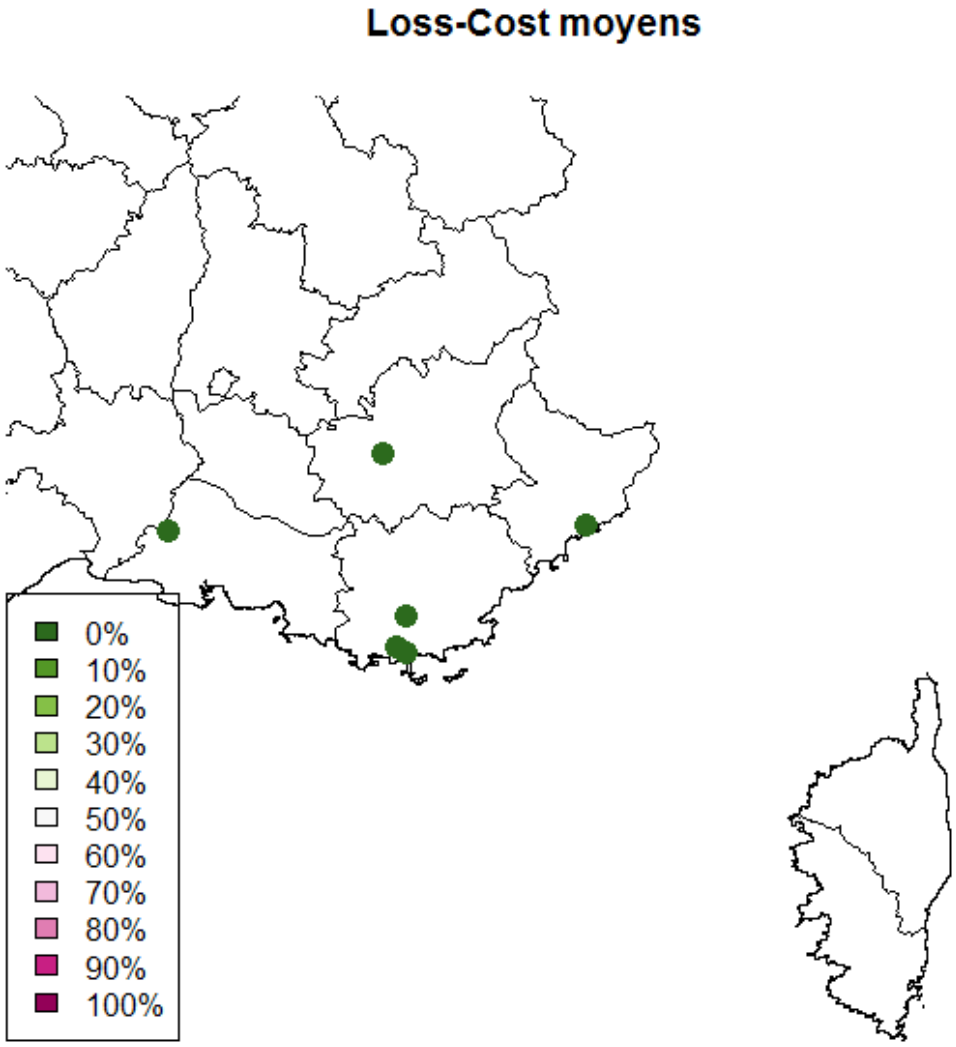
Région 1



Région 2



Région 3



Région 5