Université de Montréal

Inférence topologique

par

Noémie Prévost

Département de mathématiques et de statistique
Faculté des arts et des sciences

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en Statistique

février 2014

Université de Montréal

Faculté des études supérieures

Ce mémoire intitulé

Inférence topologique

présenté par

Noémie Prévost

a été évalué par un jury composé des personnes suivantes :

Pierre Duchesne

(président-rapporteur)

Jean-François Angers

(directeur de recherche)

Mylène Bédard

(membre du jury)

Mémoire accepté le:

Le 4 février 2014

SOMMAIRE

Les données provenant de l'échantillonnage fin d'un processus continu (champ aléatoire) peuvent être représentées sous forme d'images. Un test statistique permettant de détecter une différence entre deux images peut être vu comme un ensemble de tests où chaque pixel est comparé au pixel correspondant de l'autre image. On utilise alors une méthode de contrôle de l'erreur de type I au niveau de l'ensemble de tests, comme la correction de Bonferroni ou le contrôle du taux de faux-positifs (FDR). Des méthodes d'analyse de données ont été développées en imagerie médicale, principalement par Keith Worsley, utilisant la géométrie des champs aléatoires afin de construire un test statistique global sur une image entière. Il s'agit d'utiliser l'espérance de la caractéristique d'Euler de l'ensemble d'excursion du champ aléatoire sous-jacent à l'échantillon au-delà d'un seuil donné, pour déterminer la probabilité que le champ aléatoire dépasse ce même seuil sous l'hypothèse nulle (inférence topologique).

Nous exposons quelques notions portant sur les champs aléatoires, en particulier l'isotropie (la fonction de covariance entre deux points du champ dépend seulement de la distance qui les sépare). Nous discutons de deux méthodes pour l'analyse des champs anisotropes. La première consiste à déformer le champ puis à utiliser les volumes intrinsèques et les compacités de la caractéristique d'Euler. La seconde utilise plutôt les courbures de Lipschitz-Killing. Nous faisons ensuite une étude de niveau et de puissance de l'inférence topologique en comparaison avec la correction de Bonferroni. Finalement, nous utilisons l'inférence topologique pour décrire l'évolution du changement climatique sur le territoire du Québec entre 1991 et 2100, en utilisant des données de température simulées et publiées par l'Équipe Simulations climatiques d'Ouranos selon le modèle régional canadien du climat.

MOTS CLÉS : comparaisons multiples, caractéristique d'Euler, champs aléatoires, isotropie, courbures de Lipschitz-Killing, inférence topologique, changement climatique.

SUMMARY

Data coming from a fine sampling of a continuous process (random field) can be represented as images. A statistical test aiming at detecting a difference between two images can be seen as a group of tests in which each pixel is compared to the corresponding pixel in the other image. We then use a method to control the type I error over all the tests, such as the Bonferroni correction or the control of the false discovery rate (FDR). Methods of data analysis have been developped in the field of medical imaging, mainly by Keith Worsley, using the geometry of random fields in order to build a global statistical test over the whole image. The expected Euler characteristic of the excursion set of the random field underlying the sample over a given threshold is used in order to determine the probability that the random field exceeds this same threshold under the null hypothesis (topological inference).

We present some notions relevant to random fields, in particular isotropy (the covariance function between two given points of a field depends only on the distance between them). We discuss two methods for the analysis of non-isotropic random fields. The first one consists in deforming the field and then using the intrinsic volumes and the Euler characteristic densities. The second one uses the Lipschitz-Killing curvatures. We then perform a study of sensitivity and power of the topological inference technique comparing it to the Bonferonni correction. Finally, we use topological inference in order to describe the evolution of climate change over Quebec territory between 1991 and 2100 using temperature data simulated and published by the Climate Simulation Team at Ouranos, with the Canadian Regional Climate Model CRCM4.2.

KEY WORDS : multiple comparisons, Euler characteristic, random fields, isotropy, Lipschitz-Killing curvatures, climate change.

TABLE DES MATIÈRES

Sommaire v

Summary vii

Liste des figures xi

Liste des tableaux xiii

Remerciements 1

Introduction 3

Chapitre 1. Champs aléatoires et tests d'hypothèses multiples 7

1.1. Construction des tests 8

1.2. Contrôle du niveau 9

1.3. Contrôle du taux de fausses découvertes 10

1.4. Champs aléatoires 11

1.4.1. Stationnarité et isotropie 13

1.4.2. Champs aléatoires gaussiens 14

1.5. Modélisation 15

1.6. Inférence topologique 16

Chapitre 2. Notions d'inférence topologique 19

2.1. Complexes simpliciaux 19

2.2. Caractéristique d'Euler 20

2.3. Volumes intrinsèques 21

2.4. Compacité de la caractéristique d'Euler 23

2.5. Exemple 24

2.6. Déformer vers l'isotropie 26
2.7. Déformer vers l'isotropie locale 27
2.8. Remplacer les volumes intrinsèques par les courbures de Lipschitz-Killing 30
2.9. Les courbures de Lipschitz-Killing 31
2.10. Grille de quatre points 33
2.10.1. Indépendance 34
2.10.2. Corrélations isotropes 35
2.10.3. Corrélation horizontale 37
2.11. Comparaison de $\mu_2(S^*)$ et $\mathcal{L}_2(S)$ 37
Chapitre 3. Niveau, puissance et application au changement climatique 47
3.1. Niveau et Puissance 47
3.2. Présentation des données 51
Conclusion 57
Bibliographie 59

LISTE DES FIGURES

2.1 Compacité de dimension zéro à trois d'un champ aléatoire unitaire gaussien en fonction du seuil. 25

2.2 Ensemble d'excursion d'un champ aléatoire gaussien au-delà de quatre seuils différents 26

2.3 Complexe simpliciel (a) avant et (b) après la déformation 28

2.4 Complexes simpliciaux de (a) quatre et (b) neuf points 32

2.5 Coordonnées et composantes du complexe simpliciel de quatre points 34

2.6 Résultat de la déformation pour une simulation de taille (a) 30 et (b) 1 000. 35

2.7 Déformation d'un complexe simpliciel de quatre points pour trois structures de corrélation. 36

2.8 Déformation d'un complexe simpliciel de quatre points pour la covariance horizontale. 38

2.9 Écart relatif entre $\mu_2(S^*)$ et $\mathcal{L}_2(S)$ pour les structures de covariance indépendance, faible, forte et horizontale, pour 4, 9, 16 et 25 points. 39

2.10 Nombre d'itérations de la déformation pour (a) 4 points (b) 9 points (c) 16 points et (d) 25 points. 42

2.11 Écart relatif entre $\mu_2(S^*)$ et $\mathcal{L}_2(S)$ après la déformation pour (a) 4 points (b) 9 points (c) 16 points et (d) 25 points. 44

3.1 Une réalisation d'un champ aléatoire avec le paramètre θ égal à (a) 0,01 (b) 0,1 (c) 0,5 et (d) 0,8. 48

3.2 Centre de la distribution du groupe b sous l'hypothèse alternative . 49

3.3 Courbes de puissance avec le paramètre θ égal à (a) 0,01 (b) 0,1 (c) 0,5 et (d) 0,8. 50

3.4 Température moyenne sur le territoire du Québec. 52

3.5 Valeurs prédites 53
3.6 Distribution géographique du changement climatique en fonction du temps 54
3.7 Points auxquels nous détectons une différence de température significative pour les années (a) 1991-2020 et (b) 2001-2030 (c) 2011-2040 (d) 2021-2050 (e) 2031-2060 et (f) 2041-2070. 55

LISTE DES TABLEAUX

1.1 Configuration des hypothèses et des résultats des tests 11

2.1 Volumes intrinsèques de quelques figures géométriques 22
2.2 Compacités de dimension zéro à trois d'un champ unitaire gaussien au-delà de quelques seuils choisis 25
2.3 Distance entre les points 33
2.4 Longueur des arêtes après la transformation pour les simulations de taille (a) 30 et (b) 1 000 35
2.5 Longueur des arêtes après la transformation pour $\rho_1 = 0{,}2$ et $\rho_1 = 0{,}8$ 37
2.6 Moyenne et écart-type du nombre d'itérations avant l'arrêt de la déformation 41
2.7 Moyenne et écart-type de l'écart relatif entre $\mu_2(S^*)$ et $\mathcal{L}_2(S)$ à l'arrêt de la déformation 43

3.1 Niveaux empiriques 48

REMERCIEMENTS

J’aimerais d’abord remercier mon directeur de recherche Jean-François Angers de m’avoir fait confiance et de m’avoir accordé une grande liberté durant mon travail de recherche. J’aimerais aussi remercier les professeurs et membres du département et en particulier Miguel Chagnon pour sa disponibilité et son dévouement. Je remercie aussi ma famille pour son soutien, et spécialement mon père pour son aide lors de la révision du mémoire. Ensuite, merci à mes ami(e)s, entre autres Ariane, Isabella, Mireille, Fabian et Adriano pour leurs encouragements et pour m’avoir écouté parler de mon travail de recherche en dissimulant parfois leur absence d’intérêt pour le sujet. Finalement, merci à Mayela pour tous les repas qu’elle a préparés et que j’ai mangés en rédigeant mon mémoire, et pour tout le reste.

INTRODUCTION

Dans le cadre de ce mémoire, nous étudions l'évolution du changement climatique sur le territoire du Québec. Le territoire est divisé en 1476 régions pour lesquelles la température moyenne au mois de juillet pour un horizon temporel de 30 ans sera comparée à une période de référence allant de 1961 à 1990 et ce, pour 81 horizons de temps pris entre 1991 et 2100. Il s'agit de données simulées produites par l'Équipe Simulations climatiques d'Ouranos (Consortium sur la climatologie régionale et l'adaptation aux changements climatiques) avec le modèle régional canadien du climat. Plus de détails sont donnés à la section 3.2. Le problème consiste alors à construire une série de tests d'hypothèses dans le but d'identifier les points pour lesquels nous observons une différence significative entre deux horizons de temps donnés tout en contrôlant le niveau global, soit la probabilité d'avoir au moins un résultat positif faux parmi les 1476 tests effectués simultanément. Il faut noter que les données que nous utilisons ont été produites par des simulations d'un modèle régional du climat, conséquemment les résultats obtenus ne représentent pas l'évolution observée du climat.

Il existe de nombreuses méthodes pour contrôler le niveau d'un ensemble de tests. Parmi celles-ci, la correction de Bonferroni est une méthode très largement utilisée. Cependant, vu le nombre élevé de tests à effectuer, nous nous attendons à ce que cette méthode ne permette pas de détecter les différences de température étant donné que la correction sera très sévère.

Une solution, présentée par Benjamini et Hochberg (1995), consiste à contrôler le taux de résultats positifs faux plutôt que la probabilité d'en avoir au moins un. Cette quantité est appelée *false discovery rate*. Bien que l'approche généralement acceptée soit de contrôler le niveau, Benjamini (2010) rapporte qu'il y a eu un changement d'attitude avec l'arrivée de la recherche en génétique, où il fallait parfois procéder à plusieurs milliers de tests simultanément. Contrôler le niveau causant une perte de puissance importante, les chercheurs avaient besoin d'une approche alternative. Il est important de mentionner que

la correction de Bonferroni, tout comme le FDR sous sa forme initiale (Benjamini et Hochberg, 1995), suppose que les données des différents tests simultanés sont indépendantes, et si ce n'est pas le cas, ces méthodes donnent lieu à une correction trop sévère. Les données que nous étudions représentent plutôt un échantillonnage fin d'un processus continu (champ aléatoire) et nous pouvons nous attendre à ce que les points voisins tendent à être corrélés. De plus, étant donné que l'inférence porte sur les caractéristiques topologiques d'un phénomène aléatoire continu, comme la présence de sommets, les corrections de Bonferroni et du FDR sont arbitraires puisqu'elles dépendent du nombre de points d'échantillonnage et non du phénomène étudié (Chumbley et Friston, 2007).

Nous nous intéressons donc à une méthode qui permette d'incorporer à notre analyse la corrélation spatiale entre les différents points d'échantillonnage. Cette méthode s'applique spécifiquement aux situations où les tests multiples sont en fait une discrétisation d'un test global sur un ensemble continu. Il s'agit d'utiliser la caractéristique d'Euler pour déterminer la probabilité qu'un champ aléatoire dépasse un seuil donné. Worsley (1996) présente une introduction sur le sujet et l'histoire du début de son utilisation en astrophysique et en imagerie médicale. Nous nous référerons à cette méthode par le nom d'inférence topologique.

Au départ, l'inférence topologique a été développée pour le cas particulier d'un champ aléatoire isotrope, c'est-à-dire pour lequel la corrélation entre deux points dépend uniquement de la distance qui les sépare. Cependant en pratique, nous rencontrons souvent des champs anisotropes, ce qui a été abordé dans le cadre du problème d'interpolation (kriging) par Sampson et Guttorp (1992). Leur idée consiste à calculer une déformation du champ en modifiant les coordonnées des points d'échantillonnage de façon à obtenir un espace isotrope, pour ensuite modéliser la covariance du champ à l'aide d'un modèle isotrope.

Dans le domaine de l'inférence topologique, Worsley *et al.* (1999) reprennent l'idée de Sampson et Guttorp (1992) en décrivant une méthode simple permettant de déformer le treillis sur lequel sont représentés les points d'échantillonnage, afin d'obtenir l'isotropie au niveau de l'échantillon étudié. Taylor et Worsley (2007) présentent une méthode pour procéder à des tests multiples sur des champs aléatoires isotropes ainsi que sur des champs anisotropes qui sont des fonctions de champs gaussiens sans qu'il soit nécessaire de déformer le champ pour le rendre isotrope. Ensuite, Chamandy *et al.* (2008) proposent

un développement pour des champs non gaussiens mais pour lesquels la statistique de test sera un champ approximativement gaussien, en faisant appel au théorème limite central.

Ce mémoire est composé de trois chapitres qui sont divisés comme suit. Dans le premier chapitre, nous décrivons le type de problème que nous allons étudier et nous utilisons une modélisation univariée pour procéder à la construction d'un ensemble de tests d'hypothèses. Ensuite, nous abordons le problème des comparaisons multiples et nous décrivons les corrections de Bonferroni et du FDR. Puis, nous présentons certaines notions liées à l'étude des champs aléatoires, notamment la stationnarité et l'isotropie, ainsi que le modèle de champ aléatoire gaussien unitaire. Finalement, nous présentons une seconde modélisation des données, en les représentant comme des points d'un champ aléatoire et nous introduisons la méthode d'inférence topologique.

Dans le second chapitre, nous présentons les notions théoriques relatives à l'inférence topologique, telles que la caractéristique d'Euler, les volumes intrinsèques, les courbures de Lipschitz-Killing et les compacités du champ aléatoire au-delà d'un seuil donné. Ensuite, nous donnons des exemples de déformation des points d'échantillonnage avec la méthode présentée par Worsley *et al.* (1999) et nous appliquons aussi la méthode décrite par Taylor et Worsley (2007), qui permet de travailler directement avec les champs anisotropes.

Au troisième chapitre, nous présentons une étude de niveau et de puissance par simulation comparant l'inférence topologique avec la méthode de Bonferroni. Nous constatons que plus les données possèdent une forte corrélation de courte portée, plus le gain en puissance par rapport à Bonferroni augmente. Finalement, nous appliquons la méthode de l'inférence topologique au jeu de données de températures simulées mentionné précédemment. Nous décrivons l'évolution du changement climatique entre 1991 et 2100 à l'aide de 81 comparaisons comportant 1476 tests individuels pour lesquels nous contrôlons le niveau global.

Chapitre 1

CHAMPS ALÉATOIRES ET TESTS D'HYPOTHÈSES MULTIPLES

Dans ce chapitre, nous présentons différentes méthodes permettant de construire une série de tests d'hypothèses tout en contrôlant l'erreur de type I de façon globale. Nous introduisons aussi quelques notions permettant d'aborder l'analyse des champs aléatoires.

Considérons un ensemble fini de points $P = \{s_1, \ldots, s_N\}$, $N \in \mathbb{N}$, pour lesquels nous disposons d'une série d'observations $X_1, \ldots, X_J$ d'un même vecteur aléatoire $X_j = (X_{j1} \ldots X_{jN})$. De plus, un vecteur de réalisations X_j représente un échantillonnage fin d'un processus continu (ou champ aléatoire) défini sur un espace paramétré bidimensionnel $S \subset \mathbb{R}^2$. Par échantillonnage fin nous entendons que l'ensemble de points échantillonné est fini en pratique, mais conceptuellement, est vu comme un continuum. Par exemple, une image numérique est formée d'un nombre fini de pixels de couleurs mais nous la percevons comme continue si les pixels sont suffisamment petits. Nous souhaitons comparer deux sous-groupes d'observations au moyen d'un ensemble de statistiques de test, noté $t_1, \ldots, t_N$. Le problème consiste alors à déterminer une région critique qui permet de contrôler l'erreur de type I de l'ensemble des tests, soit la probabilité de commettre au moins une erreur de type I (niveau), ou encore le taux d'erreur de type I.

Une façon simple de contrôler l'erreur de type I consiste à voir l'ensemble de tests que nous voulons construire comme une série d'inférences séparées et d'appliquer une méthode pour tenir compte de la multiplicité. Parmi les nombreuses approches connues, nous décrivons dans les prochaines sections les méthodes de Bonferroni, de Tukey, et le contrôle du taux de fausses découvertes (FDR). Lorsque les observations aux différents points sont corrélées, ces méthodes sont trop conservatrices car la corrélation spatiale est ignorée. En effet, plus un échantillon est auto-corrélé, moins il contient d'observations

effectivement indépendantes, et moins grande est la probabilité d'y observer des valeurs extrêmes. De plus, il faut noter que le nombre de tests pour lesquels nous effectuons la correction est artificiel, puisqu'il dépend de l'échantillonnage qui a été fait et non du processus continu duquel les observations proviennent. Ceci n'a pas d'effet si nous contrôlons le taux de fausses découvertes, mais aura un impact lorsque nous contrôlons la probabilité d'avoir au moins une erreur de type I, comme c'est le cas avec les méthodes de Bonferroni et de Tukey. Nous présentons ensuite une méthode qui utilise les propriétés de l'ensemble d'excursion du champ aléatoire sous-jacent au-delà du seuil des tests, afin de contrôler le niveau de façon globale.

1.1. Construction des tests

Nous avons une série de vecteurs d'observations $X_1, \ldots, X_J$, $X \in \mathbb{R}^N$, où les N composantes d'une observation représentent les valeurs observées aux N points d'échantillonnage $P = \{s_1, \ldots, s_N\}$. La série de vecteurs d'observations $X_1, \ldots, X_J$ est séparée en deux groupes a et b et nous étudions le contraste entre ces deux groupes, pour chacune des N composantes, c'est-à-dire pour chacun des points d'échantillonnage. Nous modélisons la distribution des observations en chaque point en ignorant la distribution conjointe pour nous contenter des distributions marginales. Nous appliquons le modèle

$$X_{k,j} = \mu_k + e_{k,i}, \tag{1.1.1}$$

où $X_{k,1}, \ldots, X_{k,n_k}$ sont les valeurs observées dans l'échantillon du groupe k, $k = a, b$. Les erreurs e_i, $i = 1, \ldots, n_k$ sont indépendantes et identiquement distribuées selon la loi normale centrée-réduite. Afin de confronter les deux hypothèses

$$\left\{\begin{array}{ll} H_0 : & \mu_a = \mu_b \\ H_1 : & \mu_a \neq \mu_b \end{array}\right\}, \tag{1.1.2}$$

nous pouvons construire N tests pour les hypothèses suivantes :

$$\left\{\begin{array}{ll} H_{0i} : & \mu_{a,i} = \mu_{b,i} \\ H_{1i} : & \mu_{a,i} \neq \mu_{b,i} \end{array}\right\}, \tag{1.1.3}$$

$i = 1, \ldots, N$, où μ_{ki} est l'espérance des observations au point i pour le groupe k. En chaque point, la statistique t pour la comparaison de deux échantillons

indépendants à variances égales est donnée par

$$t = \frac{\bar{x}_a - \bar{x}_b}{\sqrt{s_c^2\left(\frac{1}{n_a} + \frac{1}{n_b}\right)}}, \text{ où } \bar{x}_k = \frac{1}{n_k}\sum_{j=1}^{n_k} x_{k,j} \tag{1.1.4}$$

et s_c^2 est l'estimateur de variance combinée. Pour chacun de ces tests prit individuellement, nous pouvons rejeter l'hypothèse nulle au niveau de confiance α lorsque $|t| \geq t_{n_a+n_b-2,1-\alpha/2}$. Maintenant, si nous utilisons un seuil commun u pour l'ensemble des tests, ceux-ci prennent la forme

$$\left\{\begin{array}{ll} \text{si } |t_i| \geq u & \text{Rejeter l'hypothèse } H_{0i} \\ \text{si } |t_i| < u & \text{Ne pas rejeter l'hypothèse } H_{0i} \end{array}\right\}, \tag{1.1.5}$$

$i = 1, \ldots, N$. Il reste à choisir le seuil de façon à contrôler l'erreur de type I au niveau global.

1.2. Contrôle du niveau

Dans cette section, nous décrivons deux méthodes permettant de déterminer un seuil global u tel que le niveau, soit la probabilité d'avoir au moins une fausse découverte parmi les N tests effectués, soit inférieur ou égal à une constante α.

La méthode de Bonferroni consiste à choisir les niveaux des tests individuels de façon à ce que leur somme soit égale au niveau global souhaité. En effet, puisque

$$\begin{aligned} &1 - \mathbb{P}\left(\bigcup_{i=1}^{N} |T_i| \geq u | H_i = H_{0i}\right) \\ &\leq 1 - \sum_{i=1}^{N} \mathbb{P}\left(|T_i| \geq u | H_i = H_{0i}\right) = 1 - \sum_{i=1}^{N} \alpha_i, \end{aligned} \tag{1.2.1}$$

nous pouvons choisir le niveau des tests comme $\{\alpha_i = \alpha/N\}_{i=1,\ldots,N}$ de façon à obtenir un niveau global inférieur ou égal à α. S'agissant d'une borne supérieure, il y a une perte de puissance lorsque la correction est trop sévère. Ceci dit, cette méthode a pour avantage d'être très simple à appliquer et de ne nécessiter aucun présupposé sur les échantillons comparés.

L'approche de Tukey, quant à elle, permet de contrôler avec exactitude le niveau de l'ensemble de tests lorsque les tailles échantillonnales n sont égales et que les échantillons sont indépendants et proviennent de distributions ayant la même variance ($\Sigma = \sigma I_N$). Il s'agit d'associer une valeur-p à l'écart maximum

constaté entre toutes les paires de moyennes en observant que

$$\mathbb{P}\left(\bigcup_{i=1}^{N}|T_i| \geq u\right) = \mathbb{P}\left(|T|_{max} \geq u\right). \tag{1.2.2}$$

Ensuite, $\mathbb{P}(|T|_{max} > u)$ est un quantile de la distribution exacte de $Q_{N,N(n-1)} = \max_{i=1\ldots N(2\times N-1)} \frac{|\bar{X}_{a_i} - \bar{X}_{b_i}|}{s_c/\sqrt{n}}$. La méthode de Tukey n'est pas efficace dans le cas qui nous intéresse, car nous ne comparons pas toutes les moyennes. En effet, nous avons N points pour lesquels nous comparons deux moyennes. Il y a donc $2\times N$ moyennes et seulement N comparaisons, parmi $N(2 \times N - 1)$ comparaisons possibles.

Notez que nous avons présenté les tests en supposant que les hypothèses sont bilatérales, mais lors de l'application, nous allons plutôt procéder à des tests unilatéraux à droite. Dans ce cas, le quantile utilisé est celui de $1 - \alpha$ plutôt que $1 - \alpha/2$ et le niveau à contrôler est donné par $\mathbb{P}(T_{max} > u)$ plutôt que $\mathbb{P}(|T|_{max} > u)$.

1.3. Contrôle du taux de fausses découvertes

La méthode de Bonferroni fonctionne bien lorsque le nombre de tests n'est pas très élevé. Par contre, lorsque le nombre d'hypothèses à tester est de plusieurs milliers, comme c'est le cas dans certains problèmes de génétique, le niveau des tests individuels peut devenir trop petit pour permettre aux tests d'avoir une puissance suffisante en pratique. Une autre approche au problème des comparaisons multiples, proposée par Benjamini et Hochberg (1995), consiste à contrôler le taux de fausses découvertes. Le tableau 1.1 représente la distribution des résultats des tests selon le rejet ou non-rejet de l'hypothèse nulle et selon que l'hypothèse nulle est vraie ou fausse. Selon la notation du tableau, U représente le nombre de tests pour lesquels l'hypothèse nulle n'est pas rejetée alors qu'elle est vraie (vrais négatifs), V, le nombre de tests où l'hypothèse nulle est rejetée alors qu'elle est vraie (faux positifs), W, le nombre de tests pour lesquels l'hypothèse nulle n'est pas rejetée alors qu'elle est fausse (faux négatifs) et S, le nombre de tests où l'hypothèse nulle est rejetée alors qu'elle est fausse (vrais positifs). Le FDR peut s'écrire, selon la notation du tableau 1.1, comme $\mathbb{E}(Q)$, où $Q = V/R$ lorsque $R \geq 0$ et 0 sinon, alors que le niveau s'écrit $\mathbb{P}(V \geq 1)$.

TABLE 1.1. Configuration des hypothèses et des résultats des tests

	Non rejet de H_0 $\lvert T_i\rvert < u$	Rejet de H_0 $\lvert T_i\rvert \geq u$	Total
H_0 Vraie	U	V	N_0
H_0 Fausse	W	S	$N - N_0$
Total	$N - R$	R	N

Dans l'article, les auteurs soulignent deux propriétés du taux de fausses découvertes. Premièrement, contrôler le FDR est équivalent à contrôler le niveau au sens faible, c'est-à-dire sous l'hypothèse nulle globale ($\bigcap_{i=1}^{N} H_i = H_{0i}$). En effet, sous l'hypothèse nulle globale, $N_0 = N$, d'où $V = R$ et nous pouvons montrer que $\mathbb{E}(Q) = \sum_{v=1}^{N} V/R \times \mathbb{P}(V = v) = \mathbb{P}(V \geq 1)$. La méthode de Bonferroni, quant à elle, contrôle le niveau au sens fort c'est-à-dire pour toute configuration des hypothèses. Ensuite, plus le nombre d'hypothèses nulles fausses augmente, plus le gain en puissance de la méthode du FDR augmente par rapport au contrôle du niveau.

Il est indiqué, selon les auteurs, de contrôler le niveau dans les cas où une conclusion basée sur un ensemble d'inférences serait vraisemblablement erronée si au moins une des inférences l'était. Contrôler le taux de fausses découvertes serait plus approprié lorsqu'une faible proportion d'inférences erronées ne change pas la validité de la conclusion globale, ou encore si plusieurs décisions séparées sont prises sans qu'il n'y ait une décision globale, comme pour l'analyse exploratoire d'une grande quantité de variables.

1.4. CHAMPS ALÉATOIRES

Dans cette section, nous introduisons des outils permettant de modéliser les processus stochastiques qui sont définis de façon continue sur des espaces à plusieurs dimensions.

Définition 1.4.1. *De façon générale, nous appelons champ aléatoire ou processus stochastique, la collection de variables aléatoires* $\{\mathcal{X}(s) : s \in S\}$*, où* S *est un espace paramétré. Un espace paramétré de dimension* d *est un ensemble non orienté indexé par un ensemble de* d *paramètres.*

Il peut être plus pratique de diviser l'espace paramétré en ses composantes géométrique et temporelle, menant à la notation $\{\mathcal{X}(s,t) : s \in S, t \in \mathbb{R}^+\}$. Au chapitre 3, le processus $\mathcal{X}(s,t)$ représente la température de l'air au sol au point s et au temps t. Cela dit, lorsque nous travaillons avec le champ d'un seul contraste temporel, la dimension temps est éliminée et nous retrouvons la

notation $\{\mathcal{X}(s) : s \in S\}$. L'espace S représente donc une région géographique et nous le traitons comme un sous-ensemble du plan $\mathbb{R}^2$. Nous pourrions aussi modéliser le sol comme une surface bidimensionnelle comprise dans $\mathbb{R}^3$, si nous voulions tenir compte de la topographie. En effet, la théorie nous permet de définir un champ aléatoire sur une variété, soit un espace topologique abstrait pouvant être vu localement comme approximativement euclidien, mais nous nous contenterons d'étudier le cas $S \subset \mathbb{R}^2$, ou S est un ensemble connexe à deux dimensions.

Maintenant que nous avons défini le champ aléatoire, nous pouvons étudier ses caractéristiques et y associer un modèle statistique. Puisque le champ aléatoire représente une collection infinie de variables aléatoires, il n'est pas possible de définir une distribution de probabilité en fonction d'un vecteur d'espérance et d'une matrice de variance-covariance. Nous définissons donc les fonctions d'espérance et de covariance, ainsi que les distributions de dimension finie. La fonction d'espérance associe l'espérance de la variable aléatoire à ses coordonnées dans l'espace paramétré, alors que la fonction de covariance décrit la covariance entre deux points donnés en fonction de leurs coordonnées.

Définition 1.4.2. *La fonction d'espérance d'un champ aléatoire* $\mathcal{X}(s)$ *est définie comme* $\mu(s) = \mathbb{E}\{\mathcal{X}(s)\}$ *et sa fonction de covariance est donnée par*

$$c(s_1, s_2) = \mathbb{E}\{(\mathcal{X}(s_1) - \mu(s_1))(\mathcal{X}(s_2) - \mu(s_2))\},$$

où s_1 *et* s_2 *sont des éléments de* P, $c(s_1, s_2)$ *est une fonction non négative et* $c(s, s)$ *correspond à la variance de* $\mathcal{X}(s)$.

Une distribution de dimension finie d'un champ aléatoire est la distribution conjointe d'un sous-ensemble fini de variables aléatoires lui appartenant.

Définition 1.4.3. *La distribution de dimension finie d'un champ aléatoire* $\mathcal{X}(s)$ *est la collection* $\{F_P\}$ *où* $P = (s_1, \ldots, s_N)$, $1 \leq N < \infty$ *est un ensemble fini de points de* S, *et*

$$\begin{cases} F_P : \mathbb{R}^N \mapsto [0, 1] \\ F_P(P, x_1, \ldots, x_n) = \mathbb{P}\{\mathcal{X}(s_1) \leq x_1, \ldots, \mathcal{X}(s_N) \leq x_N\}. \end{cases}$$

Pour récapituler la notation, nous avons ici $\mathcal{X}(P) = X = X_1 \ldots X_N$, de sorte que $\mathcal{X}(s_1)$ est équivalent à X_1. Maintenant que nous avons quelques outils nous permettant de modéliser les champs aléatoires, voyons comment nous pouvons les classifier en spécifiant certaines caractéristiques de leurs fonctions d'espérance et de covariance.

1.4.1. Stationnarité et isotropie

Les processus qui remplissent les conditions de stationnarité et d'isotropie forment une classe importante car leurs caractéristiques d'invariance permettent d'importantes simplifications. Nous définissons d'abord la stationnarité stricte et la stationnarité faible.

Définition 1.4.4. *Un processus $\mathcal{X}(s)$ est dit strictement stationnaire si ses distributions de dimension finie sont invariantes au décalage des coordonnées, c'est-à-dire que pour tout vecteur d de même dimension que S, les distributions de dimension finie de $\{\mathcal{X}(s_1)\dots\mathcal{X}(s_N)\}$ et $\{\mathcal{X}(s_1+d)\dots\mathcal{X}(s_N+d)\}$ sont identiques.*

Une condition moins forte, et plus facile à utiliser est la stationarité faible (de second ordre, en covariance), définie comme suit.

Définition 1.4.5. *Un processus $\mathcal{X}(s)$ est dit faiblement stationnaire si sa fonction moyenne est constante ($\mu(s)=\mu$) et sa fonction de covariance spatiale dépend seulement du vecteur d décrivant l'écart entre les deux points, c'est-à-dire qu'elle est invariante au décalage des coordonnées, ce que nous notons $c(s,s+d)$, ou encore, avec un abus de notation, $c(d)$.*

Si nous ajoutons à la stationnarité faible l'invariance de la fonction de covariance par rapport à l'orientation, nous obtenons la condition d'isotropie.

Définition 1.4.6. *Un processus $\mathcal{X}(s)$ est dit isotrope si sa fonction moyenne est constante ($\mu(s)=\mu$) et sa fonction de covariance spatiale dépend seulement de la norme du vecteur d décrivant l'écart entre les deux points, c'est-à-dire qu'elle est invariante au décalage des coordonnées et à l'orientation, ce que nous notons $c(s,s+d)=c(|d|)$, en faisant un abus de notation.*

Avant d'aller plus loin, nous devons définir la notion de dérivée directionnelle (ou spatiale). Il s'agit en fait de la dérivée partielle L^2 dans la direction t. La définition est tirée du livre de Adler *et al.* (2009), section 2.4.3, en prenant le cas particulier où les dérivées partielles sont de second ordre et les directions sont prises dans $\mathbb{R}^2$.

Définition 1.4.7. *Pour définir les dérivées directionnelles, choisissez un point $t\in\mathbb{R}^2$ et une séquence de deux directions t'_1,t'_2, que nous dénotons par $t'=t'_1,t'_2$. Nous pouvons par exemple choisir $t'=(e_1,e_2)=((0,1);(1,0))$. Nous disons que $\mathcal{X}(t)$ a une dérivée L^2 partielle dans la direction t' si la limite*

$$D^2_{L^2}(\mathcal{X}(t,t')) \equiv \lim_{h_1,h_2\to 0} \Delta^2\mathcal{X}(t,t',h_1,h_2) \tag{1.4.1}$$

existe en carré moyen, où $h = (h_1, h_2)$. *Ici,* $\Delta^2\mathcal{X}(t, t', h_1, h_2)$ *est la différence symétrique*

$$\Delta^2\mathcal{X}(t, t', h_1, h_2) = \sum_{s\in\{0,1\}^2} (-1)^{2-(s_1+s_2)}\mathcal{X}(t + s_1h_1t'_1 + s_2h_2t'_2) \tag{1.4.2}$$

$$= \mathcal{X}(t + h_1t'_1 + h_2t'_2) - \mathcal{X}(t + h_1t'_1) - \mathcal{X}(t + h_2t'_2) \tag{1.4.3}$$

et la limite dans (1.4.1) *est interprétée séquentiellement, c'est-à-dire qu'il faut faire tendre d'abord* h_1 *vers 0, puis* h_2 *vers 0. Ensuite,* s_1 *et* s_2 *sont les composantes de* s *dans le produit cartésien* $s \in \{0, 1\}^2 = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$.

Les dérivées directionnelles d'un champ aléatoire permettent de caractériser la relation de dépendance de courte portée entre ses points. Si un champ aléatoire est stationnaire et isotrope alors en particulier ses dérivées directionnelles sont non corrélées et leur variance est constante, c'est-à-dire que la matrice du second moment spectral du processus est constante sur la diagonale et nulle ailleurs, ce que nous écrivons $\Lambda(s) = -\ddot{c}(0) = \mathrm{Var}(\dot{\mathcal{X}}(s)) = \lambda I_D$, où le point dénote la dérivée spatiale et $c(d)$, la fonction de covariance pour deux points éloignés d'une distance d. Il s'avère que plusieurs propriétés des champs gaussiens dépendent uniquement de ces propriétés qu'on appelle isotropie locale, et en pratique c'est cette condition que nous utiliserons. Pour plus de détails, voir la section 2.4.3 de Adler *et al.* (2009).

Définition 1.4.8. *Un champ aléatoire* $\mathcal{X}(s)$ *est dit localement isotrope si la matrice de son second moment spectral est de la forme* $\Lambda(s) = \lambda I_D$.

Nous avons vu plusieurs définitions liées au concept d'isotropie, parmi lesquelles nous retenons la définition d'isotropie locale. Bien qu'elle soit la plus faible des conditions, elle est suffisante pour notre application et elle a pour avantage d'être facile à utiliser.

1.4.2. Champs aléatoires gaussiens

Nous avons maintenant introduit les notions nécessaires pour permettre la définition du champ aléatoire gaussien et en particulier du champ gaussien unitaire, que nous utiliserons comme modèle pour l'hypothèse nulle lors de l'analyse d'un champ de statistiques T. Nous rappelons d'abord la définition de la distribution normale multidimensionnelle.

Définition 1.4.9. *Soit* $X = (X_1, \ldots, X_N)^T \in \mathbb{R}^N$ *un vecteur aléatoire, et soit* $a \in \mathbb{R}^N$ *un vecteur de constantes. Nous dirons que* X *suit une distribution normale multidimensionnelle de dimension* N *si et seulement si*

$$\forall a,\ a^TX \sim \mathbb{N}(a^T\mu, a^T\Sigma a).$$

Les paramètres décrivant la distribution sont le vecteur d'espérance $\mu = (\mathbb{E}(X_1), \ldots, \mathbb{E}(X_N))^T$ *et la matrice de variance-covariance* $\Sigma = \mathbb{E}\{(X - \mu)(X - \mu)^T\}$. *En notant par* $|\Sigma|$ *le déterminant de la matrice* Σ, *la distribution de probabilité de* X *est alors donnée par*

$$f(x) = (2\pi)^{-N/2}|\Sigma|^{-1/2} \exp\left\{-\frac{1}{2}(X - \mu)^T \Sigma^{-1}(X - \mu)\right\},$$

et on écrit $X_j \sim \mathbb{N}_N(\mu, \Sigma)$.

Définition 1.4.10. *Un champ aléatoire gaussien défini sur un espace paramétré* S *est un champ aléatoire* $\mathcal{X}(s)$ *pour lequel pour tout ensemble de* N *points* $P = \{s_1, \ldots, s_N\} \in \mathbb{R}^N$ *où* $1 \leq N < \infty$, *la distribution de dimension finie de* $\{\mathcal{X}(s_1), \ldots, \mathcal{X}(s_N)\}$ *est une normale multidimensionnelle à* N *dimensions.*

Nous pouvons maintenant définir le modèle que nous allons utiliser comme hypothèse nulle, le champ aléatoire gaussien unitaire.

Définition 1.4.11. *Un champ aléatoire gaussien unitaire, noté* $Z(s)$ *est un champ gaussien de moyenne nulle, de variance unitaire et qui possède la propriété d'isotropie locale, où* $\lambda = 1$, *c'est-à-dire que la matrice de son second moment spectral est de la forme* $\Lambda(s) = I_D$.

Ce modèle peut s'appliquer pour tout champ gaussien localement isotrope, car un tel champ peut être transformé en champ gaussien unitaire en le centrant et en normalisant sa variance.

1.5. MODÉLISATION

La modélisation que nous appliquons est directement tirée de Taylor et Worsley (2007). Dans cet article, l'analyse effectuée consiste à détecter les points s du cerveau S où l'épaisseur du cortex est corrélée avec une variable indépendante, le genre. Un prétraitement permet d'uniformiser les surfaces corticales des 321 sujets et de les centrer en zéro. Le modèle général appliqué sur les données prétraitées est alors

$$Y_i(s) = w_i\beta(s) + Z_i(s)\sigma(s), \tag{1.5.1}$$

où $Y_i(s)$ est l'épaisseur du cortex du sujet i, $i = 1, \ldots, n$, w_i est un p-vecteur de régresseurs connus, $\beta(s)$ est un p-vecteur de coefficients inconnus, $Z_1(s) \ldots Z_n(s)$ sont des champs aléatoires gaussiens centrés de variance unitaire indépendants et identiquement distribués et $Var(Y_i(s)|w_i) = \sigma^2(s)$ est inconnue. Noter que les champs $Z_i(s)$ ne sont pas des champs gaussiens unitaires, car leur structure de covariance spatiale est inconnue.

Ensuite, Taylor et Worsley (2007) étudient un contraste dans $\beta(s)$ en travaillant avec un champ de Student, noté $T(s)$, car la structure de variance est inconnue. Puisque $\beta(s)$ à trois dimensions, $T(s)$ à 318 degrés de liberté. Un champ T est défini de façon analogue à un champ gaussien, c'est-à-dire que ses distributions de dimension finie sont des distributions de Student multidimensionnelles. L'analyse globale porte donc sur ce champ formé par les statistiques de test qui ont été déterminées de façon univariée en chaque point. Nous appliquons le modèle

$$T(s) = \mu(s) + Z(s)\sigma(s), \tag{1.5.2}$$

où $\mu(s)$ est l'espérance du contraste, $Z(s)$ est un champ gaussien centré de variance unitaire et $Var(T(s)) = \sigma^2(s)$ est la variance du contraste au point s. Sous l'hypothèse nulle, il n'y a pas de différence entre les deux échantillons que nous comparons, donc $\mu(s) = 0$ et

$$T(s) = Z(s)\sigma(s). \tag{1.5.3}$$

1.6. Inférence topologique

Maintenant, la méthode à laquelle nous allons nous intéresser s'applique spécifiquement à des données provenant d'un échantillonnage fin d'un processus stochastique multidimensionnel continu. Il s'agit d'utiliser une caractérisation topologique du champ aléatoire sous-jacent à notre échantillon, pour déterminer un seuil qui permettra d'identifier les points pour lesquels le contraste est significativement différent de zéro. Nous nous contentons ici d'en énoncer l'idée et dans le prochain chapitre, nous verrons les notions théoriques nécessaires pour en énoncer les détails.

Dans le cas d'un champ aléatoire isotrope défini sur un espace paramétré S, nous avons

$$\alpha = \mathbb{P}(T_{max} \geq u | H_0) \approx \mathbb{E}_\varphi(A_u(S)) = \sum_{d=0}^{D} \mu_d(S)\rho_d(u), \tag{1.6.1}$$

où $A_u(S) = \{s \in S : T(s) > u\}$, formé par les points de S pour lesquels la valeur de $T(s)$ dépasse le seuil u, est dit l'ensemble d'excursion au-delà de u. Ensuite, $\mathbb{E}_\varphi(A_u(S))$ est l'espérance de la caractéristique d'Euler de l'ensemble d'excursion de S au-delà du seuil u, $\mu_d(S)$ est le d^e volume intrinsèque de S et $\rho_d(u)$ est la d^e compacité du champ aléatoire au-delà du seuil u (Taylor et Worsley, 2007).

Dans le cas particulier qui nous intéresse, la caractéristique d'Euler, notée φ, d'un ensemble tridimensionnel compte le nombre de composantes connexes moins le nombre de trous traversant l'ensemble de part en part (tunnels) plus le nombre de trous non connectés à l'extérieur. La caractéristique d'Euler sera abordée avec plus de détails à la section 2.2. Si le seuil u est suffisamment élevé, alors $A_u(S)$ est généralement vide, et $\varphi(A_u(S)) = 0$, ou occasionnellement, quand $T_{max} \geq u$, $A_u(S)$ contient une seule composante connexe, auquel cas $\varphi(A_u(S)) = 1$. C'est pourquoi l'espérance de la caractéristique d'Euler donne une approximation de la fonction indicatrice de l'événement $T_{max} \geq u$.

Les volumes intrinsèques de S sont une généralisation du concept de volume en $0, 1, \ldots, D$ dimensions, où D est la dimension de S. La compacité du champ aléatoire au-delà du seuil u, quant à elle, représente la propension du champ aléatoire à dépasser ce seuil. Sa spécification dépend donc de la distribution de probabilité du champ aléatoire étudié tel que nous le modélisons sous l'hypothèse nulle.

Dans le cas unidimensionnel où $S \subset \mathbb{R}$ et le champ unidimensionnel $T(s)$ comporte un très grand nombre de degrés de liberté, de sorte qu'il est effectivement un champ gaussien centré de variance unitaire, (1.6.1) devient

$$\mathbb{P}(T_{max} \geq u | H_0) \approx \mathbb{P}(T(s) \geq u) + \frac{1}{2\pi} \int_S \mathrm{Var}(\dot{Z}(s))^{1/2} \times \exp(-u^2/2)\, ds, \quad (1.6.2)$$

où le point dénote la dérivée spatiale. Taylor et Worsley (2007) rapportent que le même résultat est obtenu en se basant sur les dépassements de seuil (upcrossings, Rice 1945), les inégalités de Bonferroni améliorées (Hunter, 1976 ; Worsley, 1982 ; Efron, 1997), la caractéristique d'Euler (Adler, 1981), ou les formules des tubes de Weyl-Hotelling (Sun, 1993 ; Sun et Loader, 1994).

Pour ce qui est du cas général, c'est-à-dire lorsque la condition d'isotropie n'est pas satisfaite, nous décrivons au chapitre 2 deux approches permettant d'estimer $\mathbb{E}\{\varphi(A_u)\}$. La première consiste à déformer les coordonnées de l'espace paramétré S de façon à obtenir un espace isotrope pour ensuite appliquer la théorie pour le cas isotrope. La deuxième, qui ne nécessite pas l'isotropie consiste à remplacer les volumes intrinsèques par les courbures de Lipschitz-Killing.

Dans ce chapitre, nous avons abordé le problème des tests d'hypothèses multiples et introduit quelques notions sur les champs aléatoires. Dans le prochain chapitre, nous allons survoler les notions nécessaires à l'application des méthodes utilisant la caractéristique d'Euler pour associer une valeur-p au

maximum d'un champ aléatoire, puis nous allons illustrer ces méthodes au moyen d'un exemple détaillé.

Chapitre 2

NOTIONS D'INFÉRENCE TOPOLOGIQUE

Dans ce chapitre, nous introduisons certaines notions de topologie et de géométrie qui nous permettront de donner une explication informelle de la méthode d'inférence topologique. Pour une présentation théorique formelle, nous recommandons le livre d'Adler et Taylor (2007). Nous présentons ensuite une simulation visant à illustrer les concepts que nous avons introduits et le fonctionnement de la méthode d'inférence topologique.

2.1. COMPLEXES SIMPLICIAUX

Pour estimer l'espérance de la caractéristique d'Euler de l'ensemble d'excursion au-delà du seuil, nous avons besoin du volume intrinsèque ou de la courbure de Lipschitz-Killing de l'ensemble S. Dépendamment de la nature de S, il n'y a pas toujours une façon simple d'obtenir ces quantités en pratique. Le complexe simpliciel est un outil qui permet de transformer le problème en un ensemble de problèmes plus simples. En effet, la définition d'un complexe simpliciel sur notre espace nous permet d'obtenir la notion de « volume » sur notre espace total à partir du « volume » de ses composantes. La définition qui suit est tirée de Taylor et Worsley (2007).

Définition 2.1.1. *Soit un ensemble de points* $P = \{s_1 \dots s_N\}$*. Un complexe simpliciel* $\mathcal{S}$ *est un ensemble de sous-ensembles de* P *ayant la propriété que pour tout élément* $\mathcal{F}$ *de* $\mathcal{S}$*, tous les sous-ensembles de* $\mathcal{F}$ *sont aussi des éléments de* $\mathcal{S}$*.*

En particulier, un maillage triangulaire $\mathcal{S}$ constitue un complexe simpliciel si pour tout triangle $\mathcal{F}$ faisant partie de $\mathcal{S}$, toutes ses arêtes et tous ses points sont aussi dans $\mathcal{S}$.

2.2. CARACTÉRISTIQUE D'EULER

Dans cette section, nous décrivons les origines de la caractéristique d'Euler (CE) et nous expliquons sa propriété d'invariance. Ensuite, nous en donnons une définition pour des ensembles formés de polyèdres collés et pour des complexes simpliciaux de dimension quelconque. Finalement, nous donnons un aperçu de la façon de la définir pour des ensembles plus généraux, en les recouvrant d'un maillage de polyèdres.

L'origine du concept remonte à la découverte par le mathématicien Leonhard Euler que pour tous les polyèdres convexes,

$$S - A + F = 2, \tag{2.2.1}$$

où S, A et F représentent le nombre de sommets, d'arêtes et de faces du polyèdre. La quantité $S - A + F$ correspond à la caractéristique d'Euler de la surface M du polyèdre V, que nous notons $\varphi(M)$.

La caractéristique d'Euler est un invariant homotopique, c'est-à-dire qu'elle prend la même valeur pour des ensembles entre lesquels il existe une déformation lisse (homotopie). C'est aussi un invariant topologique, c'est-à-dire qu'elle prend la même valeur pour des ensembles entre lesquels il existe une application bijective continue dont la réciproque est continue (homologie), car deux objets homéomorphes appartiennent à la même classe d'homotopie.

Par exemple, un cube a 8 sommets, 12 arêtes et 6 faces, et nous avons bien $\varphi(M) = 8 - 12 + 6 = 2$. La 2-sphère $\mathbb{S}^2 = \{x \in \mathbb{R}^3 : \|x\| = r\}$ a aussi une caractéristique d'Euler de 2 puisque la surface d'un polyèdre convexe est homéomorphe à la sphère $\mathbb{S}^2$.

Ensuite, pour un solide V formé de polyèdres collés sur au moins une face,

$$\varphi(V) \equiv S - A + F - Q, \tag{2.2.2}$$

où S, A et F représentent le nombre de sommets, d'arêtes et de faces du solide et Q, le nombre de polyèdres formant le solide. Il faut noter que (2.2.2) calcule la caractéristique d'Euler du solide, et non pas celle de sa frontière. Pour un seul cube, nous avons $8 - 12 + 6 - 1 = 1$. Nous pouvons en déduire que la caractéristique d'Euler de $\mathbb{B}^3 = \{x \in \mathbb{R}^3 : \|x\| \leq r\}$ est aussi égale à 1, puisque le cube est homéomorphe à la boule. Aussi, pour deux cubes adjacents, la face, les arêtes et les sommets partagés ne sont comptés qu'une seule fois. Un solide formé de deux cubes collés a donc 12 sommets, 20 arêtes et 11 faces, menant au calcul $12 - 20 + 11 - 2 = 1$.

En général, la caractéristique d'Euler d'un complexe simpliciel est donnée par la somme alternée de ses éléments de dimension croissante

$$\varphi(\mathcal{S}) = \sum_{\mathcal{F}\in\mathcal{S}} (-1)^{\dim(\mathcal{F})}, \tag{2.2.3}$$

par laquelle nous retrouvons la formule (2.2.2) en observant que pour un solide formé de cubes, les éléments de dimension zéro sont les sommets, ceux d'une dimension les arêtes, de deux dimensions les faces et de trois dimensions les cubes.

Maintenant, une autre propriété d'invariance fort utile est que la quantité $S-A+F-Q$ ne dépend pas de la façon dont le solide a été divisé en polyèdres. En effet, si nous considérons V, un objet appartenant à la classe des variétés tridimensionnelles avec bord, nous avons la définition

$$\varphi(V) \equiv C - H + T, \tag{2.2.4}$$

où C, H et T représentent respectivement le nombre de composantes connexes, d'anses et de trous. Une anse (ou tunnel) est définie comme un trou traversant le solide de part en part, comme le « trou » d'un beigne, alors qu'un trou n'est pas connecté à l'extérieur, comme la partie creuse à l'intérieur d'une balle de tennis. L'expression (2.2.4) permet d'obtenir la caractéristique d'Euler d'un solide à partir de ses propriétés géométriques au niveau global, et il est clair que quel que soit le maillage utilisé pour le recouvrir, un solide aura toujours le même nombre de composantes connexes, de tunnels et de trous.

Finalement, nous pouvons utiliser (2.2.2) pour déterminer la caractéristique d'Euler d'un ensemble $S \subset \mathbb{R}^3$ dont la frontière est lisse. En effet, un tel ensemble peut être recouvert d'un maillage rectilinéaire et si le maillage est suffisamment fin, et la frontière suffisamment lisse, alors la caractéristique d'Euler de la plus grande structure de cubes contenue dans l'ensemble est égale à celle de l'ensemble S (voir la section 1.3 de Adler *et al.*, 2010).

2.3. VOLUMES INTRINSÈQUES

Les volumes intrinsèques sont une famille de mesures qui donnent une caractérisation à la fois géométrique et topologique de l'objet auquel ils sont appliqués. Sans en donner une définition formelle, nous illustrons ce que sont les volumes intrinsèques à travers quelques exemples simples. Nous présentons ensuite une formule pour établir le volume intrinsèque d'un complexe simpliciel.

Pour S un ensemble D-dimensionnel, le D^e volume intrinsèque de S, noté $\mu_D(S)$, correspond à la mesure de Lebesgue en D dimensions de S, $\mu_{D-1}(S)$ à la moitié de l'aire de la surface de S, $\mu_{D-2} \dots \mu_1$ à des mesures moyennes de la taille de sections transversales de différentes dimensions de S et $\mu_0(S)$ est la caractéristique d'Euler de S.

Par exemple, pour S un ensemble à trois dimensions convexe,

$$\mu_0(S) = \varphi(S) \tag{2.3.1}$$

$$\mu_1(S) = 2 \times \text{moyenne sur toutes les rotations de la distance entre deux plans parallèles tangents à } S \tag{2.3.2}$$

$$\mu_2(S) = 1/2 \times \text{aire de la surface de } S \tag{2.3.3}$$

$$\mu_3(S) = \text{volume de } S. \tag{2.3.4}$$

Notez que l'aire d'une surface bidimensionnelle comprise dans un espace à trois dimensions inclut les aires de ses deux faces. Pour cette raison, l'aire d'une surface plane comprise dans un espace à deux dimensions correspond à la moitié de l'aire de sa surface dans un espace à trois dimensions. Aussi, pour une variété à D dimensions imbriquée dans un espace de plus haute dimensionalité, les volumes intrinsèques de dimension supérieure à D sont nuls. Nous présentons maintenant à titre d'exemples les volumes intrinsèques de dimension zéro à trois de quelques figures géométriques dans le tableau 2.1, tiré de Taylor *et al.* (2009).

TABLE 2.1. Volumes intrinsèques de quelques figures géométriques

S	$\mu_0(S)$	$\mu_1(S)$	$\mu_2(S)$	$\mu_3(S)$
Boule (rayon r)	1	$4r$	$2\pi r^2$	$\frac{4\pi r^3}{3}$
Sphère (rayon r)	2	0	$4\pi r^2$	0
Boîte (dimensions $a \times b \times c$)	1	$a + b + c$	$ab + bc + ac$	abc

Ensuite, le d^e volume intrinsèque d'un complexe simpliciel $\mathcal{S}$ est donné par la somme alternée des volumes intrinsèques des composantes $\mathcal{F}$ de $\mathcal{S}$ qui sont de dimension égale ou supérieure à d et il est donné par :

$$\mu_d(\mathcal{S}) = \sum_{\mathcal{F} \in \mathcal{S}: \dim(\mathcal{F}) \geq d} (-1)^{\dim(\mathcal{F})-d} \times \mu_d(\mathcal{F}). \tag{2.3.5}$$

En particulier, le deuxième volume intrinsèque d'un complexe simpliciel à deux dimensions s'obtient à partir de celui des triangles qui le composent en

appliquant (2.3.5) avec $\dim(\mathcal{F}) = d = 2$

$$\mu_2(\mathcal{S}) = \sum_{\mathcal{F}\in S:\dim(\mathcal{F})=2} (-1)^{2-2} \times \mu_2(\mathcal{F}) = \sum_{\text{triangles}} \mu_2(\text{triangle}). \tag{2.3.6}$$

Par (2.3.3), $\mu_2(\text{triangle})$ est égal à l'aire du triangle. Ainsi, le deuxième volume intrinsèque d'un complexe simpliciel bi-dimensionnel est simplement la somme des aires des triangles qui le composent.

Il y a beaucoup d'autre choses intéressantes à dire sur les volumes intrinsèques. Par exemple, ils sont équivalents aux fonctionnelles de Minkowski, quoique l'ordre des mesures et leur normalisation diffèrent et ils peuvent être définis implicitement via la formule du volume des tubes de Seiner et Weyl. Nous référons le lecteur à la section 3.3 de Adler *et al.* (2010) pour plus de détails.

2.4. Compacité de la caractéristique d'Euler

Supposons que le champ aléatoire est défini sur un espace paramétré de dimension D. Une traduction littérale du terme utilisé en anglais (density) pour $\rho_d(u)$ est la d^e densité de la caractéristique d'Euler au-delà du seuil u, pour $d = 0, \ldots, D$. Cependant, pour éviter la confusion potentielle avec le terme densité désignant la distribution de probabilité d'un phénomène aléatoire, nous l'appellerons plutôt la d^e compacité de la caractéristique d'Euler au-delà du seuil u. Nous en donnerons sans explication une définition générale, puis nous énoncerons les résultats que nous utilisons pour calculer l'espérance de la caractéristique d'Euler d'un champ aléatoire associée à différents seuils.

La définition qui suit est tirée de Taylor et Worsley (2007), section A.2. Notons $T(s)$, la valeur du champ aléatoire au point $s \in S$ par T. Pour $d > 0$, la compacité de la caractéristique d'Euler d'un champ aléatoire lisse au-delà du seuil u est donnée par

$$\rho_d(u) = \mathbb{E}\{(T \geq u)\det(-\ddot{T})|\dot{T} = 0\}\mathbb{P}(\dot{T} = 0), \tag{2.4.1}$$

où le point dénote la différentiation par rapport aux premières d composantes de s, d'abord comme vecteurs ligne, puis comme vecteurs colonne.

Pour un champ aléatoire unitaire gaussien,

$$\rho_0(u) = (2\pi)^{-1/2} \int_u^\infty \exp(-u^2/2)\, du = \mathbb{P}(T \geq u), \tag{2.4.2}$$

$$\rho_1(u) = (2\pi)^{-1} \exp(-u^2/2), \tag{2.4.3}$$

$$\rho_2(u) = (2\pi)^{-3/2}\, u \exp(-u^2/2), \tag{2.4.4}$$

$$\rho_3(u) = (2\pi)^{-2} (u^2 - 1) \exp(-u^2/2). \tag{2.4.5}$$

Pour un champ gaussien isotrope non unitaire ayant la matrice de second moment $\Lambda(s) = \lambda I_D$, la compacité du champ aléatoire de dimension d est multipliée par $\lambda^{d/2}$.

À la figure 2.1, nous présentons la compacité d'un champ aléatoire unitaire gaussien en fonction du seuil utilisé pour déterminer l'ensemble d'excursion.

La compacité du champ au-delà d'un seuil donné dépend à la fois du seuil et de la distribution du champ. L'espérance de la caractéristique d'Euler d'un champ isotrope est obtenue par une somme sur les dimensions allant de zéro à d de la d^e compacité du champ multiplié par le d^e volume intrinsèque. Ainsi la compacité de la caractéristique d'Euler représente la propension du champ aléatoire à dépasser un certain seuil.

2.5. EXEMPLE

Dans l'exemple suivant, tiré de Taylor et Worsley (2007), les auteurs ont simulé un champ aléatoire gaussien unitaire défini à l'intérieur d'un cube. La figure 2.2 illustre l'évolution de la caractéristique d'Euler de l'ensemble d'excursion du champ aléatoire à mesure que le seuil augmente, en la comparant à son espérance ($\mathbb{E}\varphi$) au-delà de ce même seuil.

La caractéristique d'Euler des ensembles a été calculée en les recouvrant d'un maillage rectilinéaire fin, puis en utilisant l'équation (2.2.2)

$$\varphi(V) \equiv S - A + F - Q,$$

où S, A et F représentent le nombre de sommets, d'arêtes et de faces du solide V, et Q, le nombre de polyèdres formant le solide.

L'espérance de la caractéristique d'Euler est obtenue avec (1.6.1)

$$\mathbb{E}_\varphi(A_u) = \sum_{d=0}^{D} \mu_d(S)\rho_d(u),$$

pour une grille de $10 \times 10 \times 10$. Les volumes intrinsèques d'une telle grille, obtenus du tableau 2.1 sont $\mu_0 = 1$, $\mu_1 = 30$, $\mu_2 = 300$ et $\mu_3 = 1000$, et les

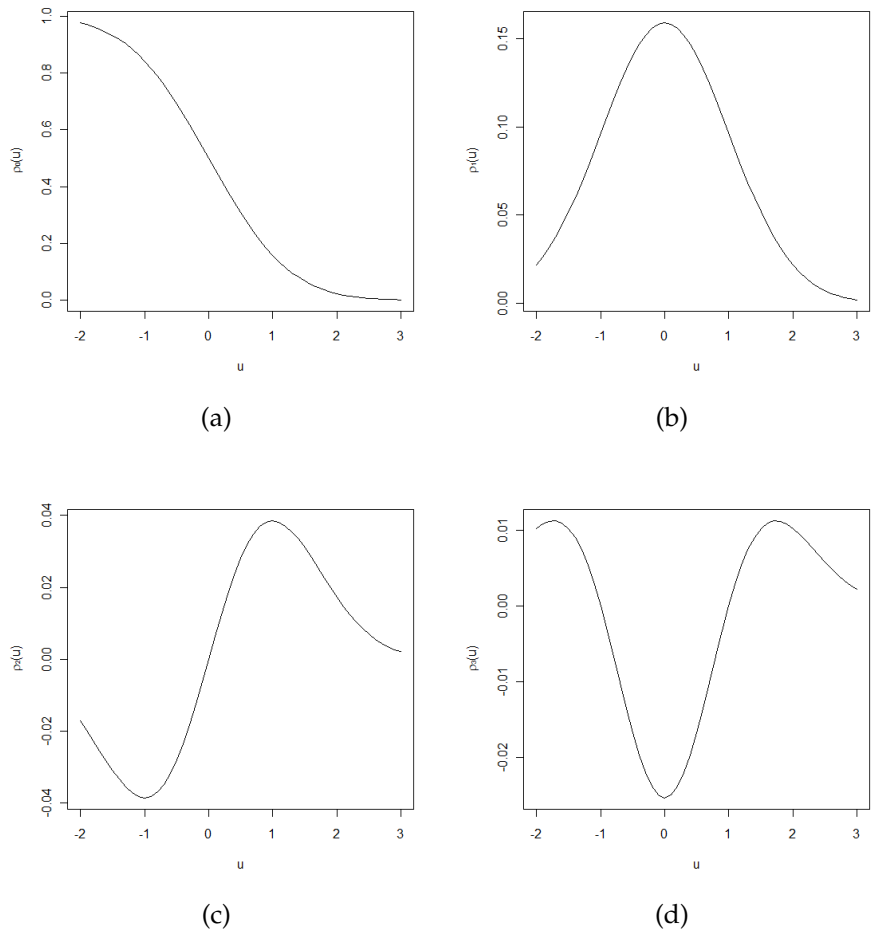
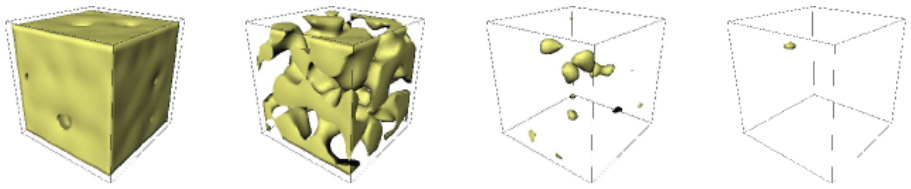


FIGURE 2.1. Compacité de dimension zéro à trois d'un champ aléatoire unitaire gaussien en fonction du seuil. (a) $\rho_0(u) = \mathbb{P}(T \geq u)$, (b) $\rho_1(u) = (2\pi)^{-1} \exp(-u^2/2)$, (c) $\rho_2(u) = (2\pi)^{-3/2} u \exp(-u^2/2)$, (d) $\rho_3(u) = (2\pi)^{-2} (u^2 - 1) \exp(-u^2/2)$.

compacités du champ aléatoire au-delà des seuils utilisés dans l'exemple sont données au tableau 2.2.

TABLE 2.2. Compacités de dimension zéro à trois d'un champ unitaire gaussien au-delà de quelques seuils choisis

Seuil	$\rho_0(S)$	$\rho_1(S)$	$\rho_2(S)$	$\rho_3(S)$
-2	0,9772	0,0215	$-0,0172$	0,0103
0	0,5000	0,1592	0,0000	$-0,0253$
2	0,0228	0,0215	0,0172	0,0103
3	0,0013	0,0018	0,0021	0,0023



(a) (b) (c) (d)

FIGURE 2.2. Ensemble d'excursion d'un champ aléatoire gaussien au-delà de quatre seuils différents. (a) Au seuil $u = -2$, l'ensemble contient des trous isolés qui contribuent chacun $+1$ pour donner $\varphi = 6$, avec $\mathbb{E}\varphi = 6{,}7$. (b) À $u = 0$, les tunnels dominent, contribuant chacun -1, pour donner $\varphi = -15$, avec $\mathbb{E}\varphi = -20{,}0$. (c) À $u = 2$, les tunnels et les trous disparaissent, laissant place à des composantes connexes isolées, contribuant chacune $+1$, donnant $\varphi = 14$ avec $\mathbb{E}\varphi = 16{,}1$. (d) À $u = 3$, il ne reste qu'une seule composante connexe (contenant la valeur maximale 3,51), donnant $\varphi = 1$, avec $\mathbb{E}\varphi = 2{,}94$ (dans l'article, nous trouvons $\mathbb{E}\varphi = 2{,}1$ et avec nos propres calculs nous obtenons $\mathbb{E}\varphi = 2{,}94$, nous avons donc supposé qu'il s'agissait d'une faute de frappe).

Cet exemple nous permet de mettre en relation les concepts que nous avons vus jusqu'à présent. Cependant, lorsque nous appliquerons la méthode de l'inférence topologique, nous ne calculerons pas la caractéristique d'Euler de l'ensemble d'excursion d'un champ aléatoire, mais uniquement son espérance. Ensuite, cet exemple présente le cas d'un champ isotrope, mais dans la pratique nous rencontrons souvent des champs anisotropes, et dans les prochaines sections nous verrons deux façons différentes de les aborder.

2.6. DÉFORMER VERS L'ISOTROPIE

Pour décrire la structure de variance-covariance d'un champ isotrope, il suffit de modéliser la relation de covariance entre deux points quelconques du champ en fonction de la distance qui les sépare (voir Le et Zidek (2006), section 6.4). En revanche, décrire la structure de covariance d'un champ anisotrope est beaucoup plus complexe car les relations de dépendance entre les paires de variables ne sont pas invariantes au décalage ni à la rotation. Il faut alors recourir à une modélisation plus compliquée. Le livre de Le et Zidek (2006) mentionne entre autres l'approche par convolution de Higdon *et al.* (1999) et la méthode de Fuentes (2001), qui utilise une moyenne pondérée de champs isotropes non corrélés. Une alternative consiste à déformer le champ de façon

à obtenir un champ isotrope, c'est-à-dire procéder à une mise à l'échelle multidimensionnelle de façon à obtenir un nouvel ensemble de coordonnées pour lequels l'isotropie est approchée.

Sampson et Guttorp (1992) présentent une méthode permettant d'estimer la structure de covariance spatiale de champs aléatoires anisotropes bidimensionnels à valeurs réelles. Ils utilisent le variogramme ou fonction de dispersion spatiale, soit $Var\{\mathbb{E}(\mathcal{X}(s_1) - \mathcal{X}(s_2))\}$, comme métrique pour la structure de covariance spatiale. Ils commencent par calculer une représentation bidimensionnelle des points d'échantillonnage pour laquelle une certaine fonction monotone g de la distance inter-points approxime les dispersions spatiales :

$$Var\{\mathbb{E}(\mathcal{X}(s_1) - \mathcal{X}(s_2))\} \approx \hat{g}(|s_1 - s_2|). \tag{2.6.1}$$

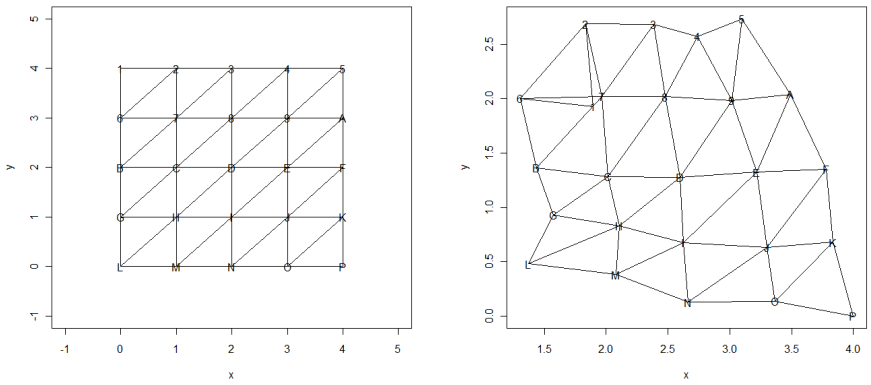
La fonction g est choisie de façon à assurer que le modèle de fonction de covariance résultant soit défini non négatif. La prochaine étape consiste à trouver de façon non paramétrique une fonction f bijective et lisse telle que le champ puisse être représenté sous la forme $\mathcal{X}(s) = g(f(s))$, où $f(g^{-1}(s))$ est approximativement isotrope. Selon Adler *et al.* (2010, section 2.4), la procédure de Sampson et Guttorp (1992) est utile mais n'a pas de généralisation évidente pour des champs ayant plus de deux dimensions.

Dans les prochaines sections, nous présentons une série de simulations visant à illustrer et à comparer deux façons d'obtenir les mesures de « volume » (de l'espace paramétré S sur lequel est défini le champ aléatoire) qui sont nécessaires pour établir le seuil de signification d'un test global sur un champ aléatoire. La première méthode consiste à modifier les coordonnées des points de façon à créer l'isotropie locale ($\Lambda(s) = \lambda I_D$) pour ensuite calculer le second volume intrinsèque de l'espace déformé. La deuxième méthode consiste à calculer l'estimateur de la seconde courbure de Lipschitz-Killing $\mathcal{L}_2(S)$, qui généralise le concept de volume intrinsèque pour les cas où le champ de covariance n'est pas isotrope.

2.7. DÉFORMER VERS L'ISOTROPIE LOCALE

Au départ, les points sont placés sur un graphe représentant leur disposition géométrique ou géographique, comme à la figure 2.3(a). Après la déformation, les nouvelles coordonnées des points sont telles que la structure de covariance du champ aléatoire est isotrope, comme à la figure 2.3(b). Ensuite, nous calculons le volume intrinsèque de l'espace déformé, noté $\mu_2(S^*)$, où S^* est l'espace déformé. La méthodologie employée pour ce faire est tirée de Worsley *et. al* (1999). Notez que la figure 2.3 est présentée à fin d'illustration

uniquement et que l'effet de la déformation sur le champ est décrit à la section 2.10 à l'aide de plusieurs exemples détaillés.



(a) (b)

FIGURE 2.3. Complexe simpliciel (a) avant et (b) après la déformation

En travaillant avec l'isotropie locale, nous nous intéressons à la relation de dépendance entre chaque paire de points reliés par une arête, plutôt qu'aux relations entre toutes les paires de points. Selon le théorème de plongement de Nash (Nash, 1956), pour un champ aléatoire défini sur une variété à D dimensions, il suffit de $D+\frac{D(D+1)}{2}$ dimensions pour que l'isotropie exacte soit possible. Ainsi, pour un champ aléatoire défini sur une variété bidimensionnelle, cinq dimensions suffisent. Cependant, nous nous intéresserons uniquement au cas d'une déformation de $\mathbb{R}^2$ vers $\mathbb{R}^2$, afin de pouvoir visualiser le complexe simpliciel déformé.

Nous décrivons d'abord la méthode en utilisant la terminologie d'imagerie médicale de l'article, puis nous faisons un lien avec la condition d'isotropie locale telle que nous l'avons définie au premier chapitre.

En résumé, la méthode consiste à normaliser les résidus, puis à estimer la rugosité (γ) du champ aléatoire le long de chacune des arêtes du graphe pour tenter de déformer l'espace S de façon à ce que γ soit égal à 1 pour toutes les arêtes. La rugosité du champ aléatoire correspond à l'écart-type de la dérivée du bruit divisé par l'écart-type du bruit lui-même.

Soit $\{u_{ji}\}$ la matrice $N \times n$ des résidus normalisés et $u_1 \dots u_N$ ses lignes, qui correspondent aux n observations d'un même point. Les résidus normalisés sont définis en chaque point de chacune des observations comme

$$u_{ij} = \frac{r_{ji}}{\|r_{ji}\|} = \frac{r_{ji}}{\sqrt{\sum_{j=1}^{n} r_{ji}^2}}, \tag{2.7.1}$$

où r_{ji} correspond au résidu observé au point i de l'observation j. Nous numérotons les points aux deux extrémités de chaque arête par 1 et 2, et nous calculons

$$\Delta u = \|u_1 - u_2\| = \sqrt{\sum_{j=1}^{n} (u_{j1} - u_{j2})^2}. \tag{2.7.2}$$

Alors, l'estimateur de la rugosité du champ aléatoire le long de l'arête est donné par :

$$\hat{\gamma} = \Delta u / \Delta x, \tag{2.7.3}$$

où Δx est la longueur de l'arrête. L'estimateur utilisé est sans biais (Worsley et al., 1999).

Maintenant, nous allons voir que les manipulations précédentes sont équivalentes à diviser les résidus par leur écart-type séparément pour chaque point, pour ensuite estimer l'écart-type de la dérivée spatiale discrète le long de l'arête. En effet, en normalisant les résidus, nous procédons, à une constante près, à la division par l'écart-type échantillonnal. Ensuite, la dérivée directionnelle discrète du champ le long de l'arête est donnée par $\Delta u/\Delta x$. En effet, en posant $u^*_{ji} = \dfrac{r_{ji}}{\sqrt{\frac{1}{n}\sum_{j=1}^{n} r_{ji}^2}} = \sqrt{n}\, u_{ji}$, alors

$$\hat{\gamma} = \frac{\Delta u}{\Delta x} = \sqrt{\frac{\sum_{i=1}^{n} \frac{1}{n}(u^*_{1i} - u^*_{2i})^2}{(\Delta x)^2}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} \left(\frac{\Delta u^*}{\Delta x}\right)^2} \quad = \sqrt{\hat{\mathrm{Var}}\left(\frac{\Delta u^*}{\Delta x}\right)}.$$

Ainsi, $\gamma = \sqrt{\lambda}$, où λ est le second moment spectral évalué le long de l'arête. En effet, la matrice du second moment spectral, $\Lambda(s) = \mathrm{Var}(\dot{Z}(s)) = \lambda I_D$, devient un simple scalaire λ car nous estimons $\Lambda(s)$ dans seulement une direction, celle de l'arête.

Nous voulons maintenant déformer les coordonnées des points de façon à ce que $\gamma = \sqrt{\lambda} \approx 1$ pour chaque arête, ce que nous pouvons aussi écrire $\lambda \approx 1$. Il suffit que λ soit constant pour remplir la condition d'isotropie, mais nous choisissons $\lambda = 1$, qui correspond au second moment spectral d'un champ aléatoire gaussien unitaire.

La méthode de déformation consiste à minimiser un critère de stress pour un point à la fois en gardant les autres fixes, et itérer jusqu'à convergence. Le critère de stress est le suivant :

$$\mathrm{Stress} = \sum_{\text{arêtes}} ([\Delta \tilde{x}]^2 - [\Delta u]^2)^2, \tag{2.7.4}$$

où $\Delta\tilde{x}$ représente la longueur de l'arête après le déplacement du point. La déformation prend fin lorsque pour cinq itérations successives, la différence entre le résultat obtenu à l'itération courante et celui de la précédente est inférieure à $0.001 \times \mathcal{L}_2(S)$, où $\mathcal{L}_2(S)$ est la seconde courbure de Lipschitz-Killing de S, notion qui sera définie à la prochaine section. Nous avons choisi ce critère d'arrêt proportionnel pour faciliter la comparaison entre des simulations de tailles différentes.

Finalement, nous calculons le 2^e volume intrinsèque en additionnant l'aire de tous les triangles formant le maillage. Notez que les triangles peuvent se chevaucher après la déformation, comme à la figure 2.3(b). La somme des aires des triangles n'est alors pas égale à l'aire du polygone formé par la frontière du maillage déformé.

2.8. Remplacer les volumes intrinsèques par les courbures de Lipschitz-Killing

Les courbures de Lipschitz-Killing dépendent à la fois des « volumes » de S et de la structure de covariance du champ aléatoire, mais uniquement à travers son second moment spectral, une mesure de corrélation de courte portée. Le second moment spectral est donné par $\Lambda(s) = Var(\dot{\mathcal{X}}(s))$, où le point dénote la dérivée spatiale. La matrice $\Lambda(s)$ est de dimension $D \times D$, où D représente la dimension de l'espace paramétré sur lequel le champ est défini, et ses éléments représentent la variance de la dérivée spatiale du champ dans D directions, par exemple celles données par les vecteurs de la base canonique de $\mathbb{R}^D$.

Dans le cas isotrope, $\Lambda(s) = \lambda I_D$ et $\mathcal{L}_d(S, \Lambda(s)) = \mu_d(S)$. En effet, la courbure de Lipschitz-Killing constitue une généralisation du volume intrinsèque pour les champs anisotropes. Pour S un intervalle,

$$\mathcal{L}_1(S, \Lambda(s)) = \int_S \Lambda(s) ds = \int_S Var(\dot{\mathcal{X}}(s))^{1/2} ds, \tag{2.8.1}$$

et en général,

$$\mathcal{L}_d(S, \Lambda(s)) = \int_S \det\{\Lambda(s)\}^{1/2} \, ds. \tag{2.8.2}$$

Nous pouvons voir $\mathcal{L}_d(S, \Lambda(s))$ comme une somme sur l'espace S d'éléments de « volume » en d-dimensions donnés par le déterminant de la matrice des dérivées directionnelles en s. Ces « volumes » sont représentatifs de la propension du champ à dépasser le seuil u car plus la corrélation entre les points de S est forte, plus le nombre d'observations effectivement indépendantes diminue, et plus les « volumes » sont petits.

Maintenant, pour les champs anisotropes, plutôt que de procéder à une déformation, nous pouvons remplacer (1.6.1) par

$$\mathbb{P}(T_{max} \geq u) \approx \mathbb{E}_{\varphi}(A_u(S)) = \sum_{d=0}^{D} \mathcal{L}_d(S, \Lambda(s))\, \rho_d(u), \tag{2.8.3}$$

où $\rho_d(u)$ est la compacité d'un champ aléatoire unitaire gaussien, puisque maintenant la structure de covariance du champ est reflétée dans $\mathcal{L}_d(S, \Lambda(s))$.

Ce résultat est valide sur des espaces paramétrés suffisamment réguliers, si le champ aléatoire est dérivé d'un champ gaussien par une fonction lisse (Chamandy *et al.*, 2008). Pour une présentation détaillée, voir le chapitre 10 du livre de Adler et Taylor (2007).

2.9. Les courbures de Lipschitz-Killing

Nous présentons maintenant les estimateurs des courbures de Lipschitz-Killing de dimension zéro à deux tirés de Taylor et Worsley (2007), section 5.2.

La méthode d'estimation consiste à calculer les volumes intrinsèques en remplaçant les coordonnées des points par les résidus normalisés observés en ces points, ce que nous noterons $\widehat{\mathcal{L}}_d(S, \Lambda(s)) = \mu_d(\widetilde{S})$, où $\widetilde{S}$ représente l'espace où les coordonnées sont remplacées par les résidus. Pour justifier cette méthode d'estimation, Taylor et Worsley (2007) proposent deux idées différentes. Premièrement, l'isotropie échantillonnale exacte peut être atteinte en utilisant autant de dimensions qu'il y a d'observations, soit en remplaçant les coordonnées des points par les valeurs des résidus normalisés qui y sont observées. Il n'est alors pas nécessaire de calculer une bijection entre l'espace initial et l'espace isotrope, il suffit de savoir que celle-ci existe. Cette étape d'estimation étant évitée, nous disons que cette méthode donne une représentation exactement isotrope de l'échantillon. La deuxième idée consiste à remplacer les distances provenant de la métrique euclidienne sur les coordonnées des points par une métrique sur la structure de covariance du champ.

D'après (2.3.5),

$$\widehat{\mathcal{L}}_d(S, \Lambda(s)) = \sum_{\mathcal{F} \in \mathcal{S}: \dim(\mathcal{F}) \geq d} (-1)^{\dim(\mathcal{F}) - d} \times \mu_d(\widetilde{\mathcal{F}}), \tag{2.9.1}$$

et en particulier,

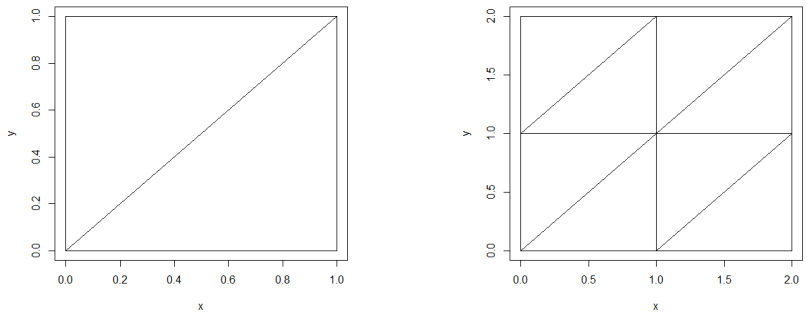
$$\widehat{\mathcal{L}}_0(S,\Lambda(s)) = \mu_0(\widetilde{S}) = \sum_{\text{points}} \mu_0(\widetilde{\text{points}}) - \sum_{\text{arêtes}} \mu_0(\widetilde{\text{arête}}) + \sum_{\text{triangles}} \mu_0(\widetilde{\text{triangle}}) \tag{2.9.2}$$

$$= \sum_{\text{points}} 1 - \sum_{\text{arêtes}} 1 + \sum_{\text{triangles}} 1,$$

$$\widehat{\mathcal{L}}_1(S,\Lambda(s)) = \mu_1(\widetilde{S}) = \sum_{\text{arêtes}} \mu_1(\widetilde{\text{arête}}) - \sum_{\text{triangles}} \mu_1(\widetilde{\text{triangle}}), \tag{2.9.3}$$

$$\widehat{\mathcal{L}}_2(S,\Lambda(s)) = \mu_2(\widetilde{S}) = \sum_{\text{triangles}} \mu_2(\widetilde{\text{triangle}}). \tag{2.9.4}$$

L'estimateur de $\mathcal{L}_0(S,\Lambda(s))$ est donné par le nombre de composantes de dimension zéro (points) moins le nombre de composantes d'une dimension (arêtes), plus le nombre de composantes de dimension deux (triangles) du complexe simplicial $\mathcal{S}$ qui recouvre S (par (2.2.3) ou (2.9.2)). Cependant, il n'est pas nécessaire de faire de calculs puisque $\mathcal{L}_0$ correspond à la caractéristique d'Euler. Or, pour les espaces paramétrés que nous utilisons, $\widehat{\mathcal{L}}_0(S,\Lambda(s)) = 1$ car il s'agit de surfaces planes connexes. En effet, comme nous pouvons le voir à la figure 2.4, pour 4 points, $\widehat{\mathcal{L}}_0(S,\Lambda(s)) = 4-5+2 = 1$ et pour 9 points, $\widehat{\mathcal{L}}_0(S,\Lambda(s)) = 9-16+8 = 1$.



(a) (b)

FIGURE 2.4. Complexes simpliciaux de (a) quatre et (b) neuf points

Ensuite, $\mu_1(\widetilde{\text{arête}})$ correspond à la longueur de l'arête dans l'espace des résidus normalisés. En numérotant les points aux deux extrémités de chaque arête par 1 et 2, $\mu_1(\widetilde{\text{arête}}) = \|u_1 - u_2\|$. De même, puisque le volume intrinsèque d'un

triangle correspond à la moitié de son périmètre, en numérotant les trois sommets du triangle par 0, 1 et 2, $\mu_1(\widetilde{\text{triangle}}) = (\|u_0-u_1\|+\|u_1-u_2\|+\|u_2-u_0\|)/2$.

Finalement, $\mu_2(\widetilde{\text{triangle}})$ correspond à la surface du triangle dans l'espace des résidus. Soit $\Delta u_{triangle} = (u_1 - u_0, u_2 - u_0)$, nous obtenons $\mu_2(\widetilde{\text{triangle}})$ en calculant $\frac{1}{2}\sum_{triangles} |\Delta u'_{triangle}\Delta u_{triangle}|^{1/2}$.

2.10. GRILLE DE QUATRE POINTS

Dans cette section, nous présentons une étude des effets de la déformation vers l'isotropie locale pour des grilles de quatre points. Nous simulons un échantillon de n observations des valeurs prises par un champ aléatoire gaussien en un nombre fini de points, $i = 1, \ldots, 4$ situés sur un maillage triangulaire. L'analyse porte sur les résidus car généralement les données sont pré-traitées en ajustant par exemple un modèle linéaire. Par contre, pour nos exemples, nous simulons directement ces résidus en générant des réalisations de lois normales multidimensionnelles centrées. Leur matrice de covariance (Σ) correspond à la structure de corrélation spatiale du champ aléatoire sous-jacent duquel nous simulons les échantillons. En effet, par définition, un champ gaussien est tel que pour tout sous-ensemble de points pris dans l'espace paramétré sur lequel il est défini, leurs variables aléatoires associées ont conjointement une distribution normale multidimensionnelle. Nous traitons quatre structures de covariance. Les trois premières sont isotropes, soit l'indépendance, une corrélation faible et une corrélation forte. La dernière est une structure de covariance anisotrope que nous appelons « horizontale ».

Les coordonnées initiales des points 1 à 4 sont respectivement (0,1), (1,1), (0,0) et (1,0). Le complexe simplicial est formé par les composantes de dimension zéro, soit les quatre points, celle d'une dimension, soit les arêtes (1,2), (1,3), (2,3), (2,4) et (3,4), et celles à deux dimensions soit les triangles reliant les points (1,2,3) et (2,3,4). Les coordonnées initiales et le complexe simplicial sont représentés à la figure 2.5, et les distances entre les points sont données au tableau 2.3.

TABLE 2.3. Distance entre les points

	1	2	3	4
1	0	1	1	$\sqrt{2}$
2	1	0	$\sqrt{2}$	1
3	1	$\sqrt{2}$	0	1
4	$\sqrt{2}$	1	1	0

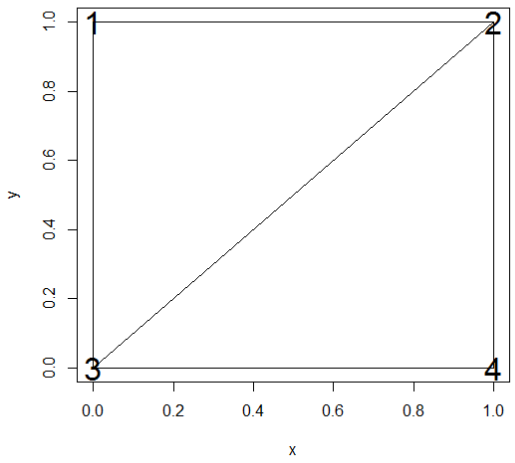


FIGURE 2.5. Coordonnées et composantes du complexe simplicial de quatre points

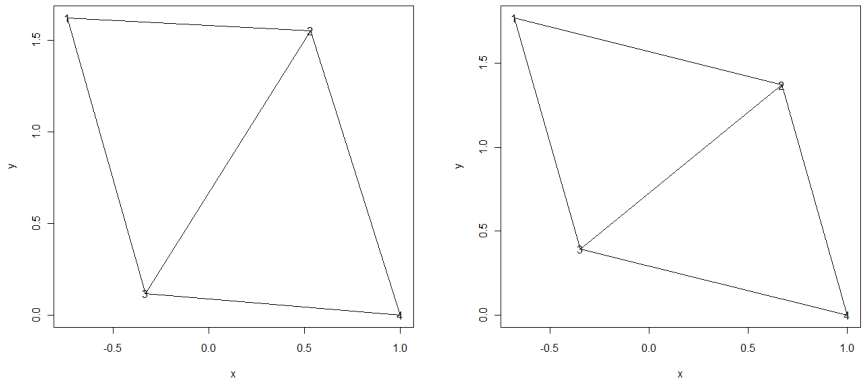
2.10.1. Indépendance

Prenons d'abord l'exemple de la grille de 4 points indépendants. Nous générons chaque grille de points conjointement à partir d'une normale multidimensionnelle avec la matrice de covariance $\Sigma = I_N$.

L'objectif de la déformation est de trouver un nouvel ensemble de coordonnées faisant en sorte que $\lambda = \Delta u/\Delta x = 1$ pour chaque arête, alors que le fait d'avoir λ uniforme sur toutes les arêtes est suffisant pour remplir la condition d'isotropie. Ainsi, même une grille isotrope sera déformée.

Aussi, la longueur des arêtes n'est pas la même pour toutes les arêtes du complexe simpliciel, alors que la différence de résidus sera asymptotiquement la même entre chaque paire de points. Par conséquent, la déformation tendra, lorsque la taille de l'échantillon est grande, à produire un losange formé par deux triangles équilatéraux. La figure 2.6 présente le complexe simpliciel après la déformation pour deux simulations de taille 30 et 1 000. La longueur des arêtes après la déformation pour ces même simulations est présentée aux tableaux 2.4 (a) et (b).

Nous constatons que pour l'échantillon de taille 1 000, la forme de l'espace déformé est très proche d'un losange formé de deux triangles équilatéraux , alors qu'avec un échantillon de taille 30, il y a moins d'uniformité dans la longueur des arêtes.



(a) (b)

FIGURE 2.6. Résultat de la déformation pour une simulation de taille (a) 30 et (b) 1 000.

TABLE 2.4. Longueur des arêtes après la transformation pour les simulations de taille (a) 30 et (b) 1 000

(b)

	1	2	3	4
1	-	1,272	1,556	-
2	1,272	-	1,672	1,620
3	1,556	1,672	-	1,340
4	-	1,620	1,340	-

(a)

	1	2	3	4
1	-	1,408	1,418	-
2	1,408	-	1,411	1,412
3	1,418	1,411	-	1,405
4	-	1,412	1,405	-

2.10.2. Corrélations isotropes

Par définition, pour les structures isotropes, la corrélation entre deux points dépend seulement de la distance qui les sépare, notée Δx. Nous présentons deux structures de covariance de la forme $C(\Delta x) = \rho/\Delta x$, $C(0) = 1$, soit une corrélation faible, avec $\rho = 0{,}2$, et une corrélation forte, avec $\rho = 0{,}6$. Les matrices de covariance correspondantes sont de la forme :

$$\Sigma_{C(\Delta x)} = \begin{pmatrix} 1 & \rho & \rho & \frac{\rho}{\sqrt{2}} \\ \rho & 1 & \frac{\rho}{\sqrt{2}} & \rho \\ \rho & \frac{\rho}{\sqrt{2}} & 1 & \rho \\ \frac{\rho}{\sqrt{2}} & \rho & \rho & 1 \end{pmatrix}$$

Nous générons des échantillons de taille 1 000 afin de nous rapprocher du comportement asymptotique. Nous présentons à la figure 2.7, sur une même échelle, la configuration initiale des points, ainsi que la configuration après la

déformation pour l'indépendance, la corrélation faible et la corrélation forte. Une plus grande corrélation entre les points de la grille entraîne une plus petite différence de résidus entre les paires de points aux extrémités des arêtes. La déformation mène donc à une courte distance entre ces paires de points.

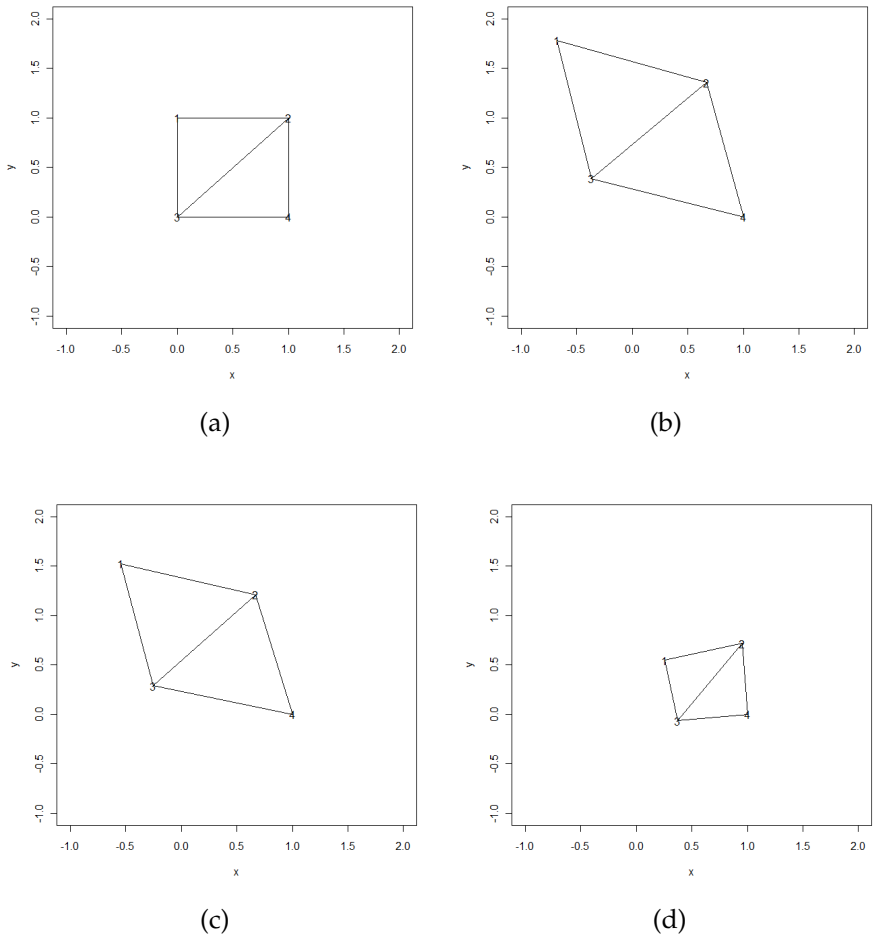


FIGURE 2.7. Déformation d'un complexe simpliciel de quatre points pour trois structures de corrélation. (a) Position initiale $Aire(S) = 1$ (b) Indépendance $\mu_2(S^*) = 1{,}71$ (c) Corrélation faible $\mu_2(S^*) = 1{,}41$ (d) Corrélation forte $\mu_2(S^*) = 0{,}45$

Puisque la corrélation entre chaque paire de points est inversement proportionnelle à la longueur de l'arête, nous pouvons imaginer que la forme du complexe simpliciel devrait rester intacte, et son aire diminuer lorsque ρ augmente. Nous observons que pour la corrélation faible, la forme du complexe après la déformation se rapproche du losange formé de deux triangles équilatéraux, tandis que pour la corrélation forte, la forme se rapproche de la forme initiale, un carré.

2.10.3. Corrélation horizontale

Finalement, la covariance « horizontale » est une structure où deux points voisins selon l'axe horizontal ont une covariance de ρ_1 alors que les autres paires de points ont une covariance nulle. La matrice de covariance est la suivante :

$$\Sigma_{\text{horizontale}} = \begin{pmatrix} 1 & \rho_1 & 0 & 0 \\ \rho_1 & 1 & 0 & 0 \\ 0 & 0 & 1 & \rho_1 \\ 0 & 0 & \rho_1 & 1 \end{pmatrix}.$$

Afin d'illustrer le fonctionnement de la déformation pour cette structure de covariance, nous simulons quatre échantillons de taille 1 000 avec $\rho_1 = \{0{,}2\,;\ 0{,}4\,;\ 0{,}6\,;\ 0{,}8\}$.

Les complexes simpliciaux transformés sont présentés à la figure 2.8, et le tableau 2.5 présente la longueur des arêtes après la déformation pour $\rho_1 = 0{,}2$ et $\rho_1 = 0{,}8$. Nous nous attendons à ce que les arêtes (1,2) et (3,4) deviennent

TABLE 2.5. Longueur des arêtes après la transformation pour $\rho_1 = 0{,}2$ et $\rho_1 = 0{,}8$

(a)

	1	2	3	4
1	-	1,288	1,363	-
2	1,288	-	1,421	1,397
3	1,363	1,421	-	1,249
4	-	1,397	1,249	-

(b)

	1	2	3	4
1	-	0,642	1,395	-
2	0,642	-	1,404	1,409
3	1,395	1,404	-	0,611
4	-	1,409	0,611	-

plus courtes à mesure que ρ_1 augmente, tandis que les autres arêtes prennent la longueur que nous retrouvons lorsque les deux points sont indépendants (environ 1,4), voir le tableau 2.4.

2.11. COMPARAISON DE $\mu_2(S^*)$ ET $\mathcal{L}_2(S)$

Pour comparer les deux méthodes, nous simulons des grilles de taille 4, 9, 16 et 25 disposées sur des maillages carrés pour chacune des quatre structures de covariance décrites précédemment soit l'indépendance et les covariances faible, forte et horizontale ($\rho_1 = 0{,}5$). La simulation comporte 100 répétitions de chacun des 16 scénarios, et nous utilisons $n = 30$. Nous ne calculons pas

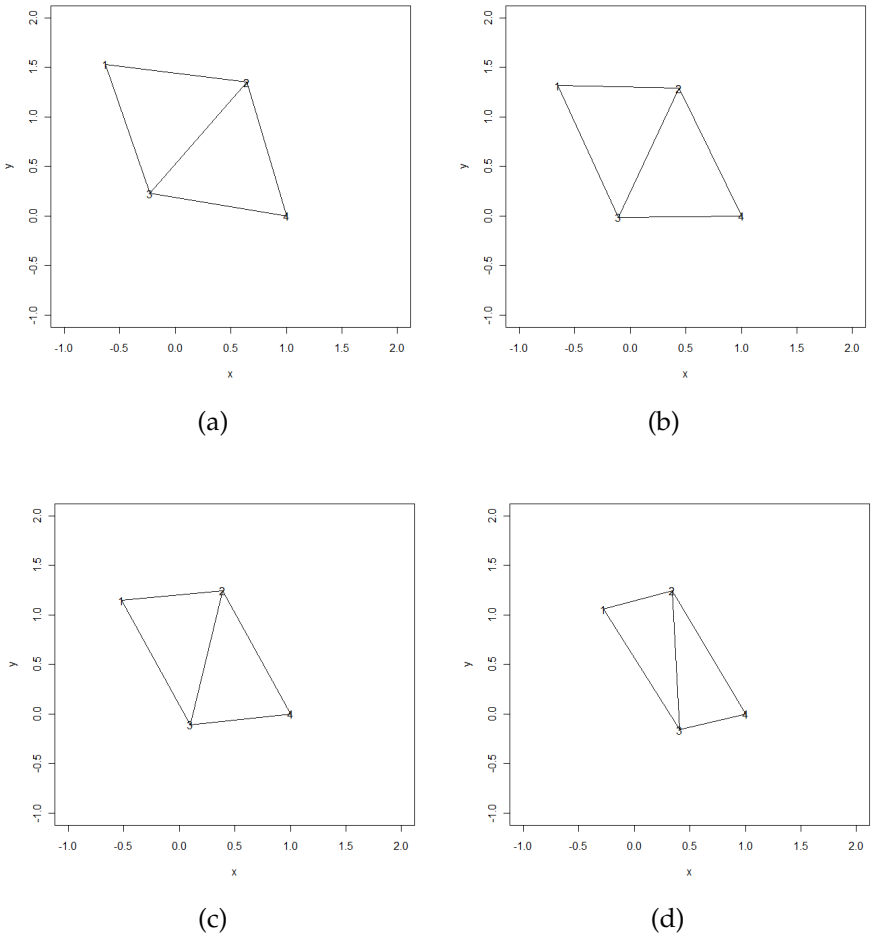


FIGURE 2.8. Déformation d'un complexe simpliciel de quatre points pour la covariance horizontale. (a) $\rho_1 = 0{,}2$ $\mu_2(S^*) = 1{,}58$ (b) $\rho_1 = 0{,}4$ $\mu_2(S^*) = 1{,}43$ (c) $\rho_1 = 0{,}6$ $\mu_2(S^*) = 1{,}19$ (d) $\rho_1 = 0{,}8$ $\mu_2(S^*) = 0{,}85$

les seuils, nous contentant d'étudier les mesures à deux dimensions, qui représentent le terme dominant des développements (1.6.1) et (2.8.3) pour S bidimensionnel.

Comme nous l'avons mentionné précédemment (section 2.9), calculer $\mathcal{L}_2(S)$ est équivalent à calculer $\mu_2(S^*)$, à condition que la déformation mène à une configuration des points parfaitement isotrope pour l'échantillon observé. Ce n'est pas le cas pour les grilles de plus de quatre points et par conséquent, les deux mesures ne coïncident pas. Cela est dû au fait que l'espace vers lequel est déformé S comporte seulement deux dimensions, alors qu'il en faudrait cinq pour garantir que la déformation qui donne l'isotropie existe.

Nous présentons des résumés graphiques et numériques de la distribution du nombre d'itérations du processus de déformation avant l'atteinte du critère d'arrêt, de l'écart relatif entre $\mu_2(S^*)$ et $\mathcal{L}_2(S)$ au moment de l'atteinte de ce même critère, ainsi que de la relation entre les deux.

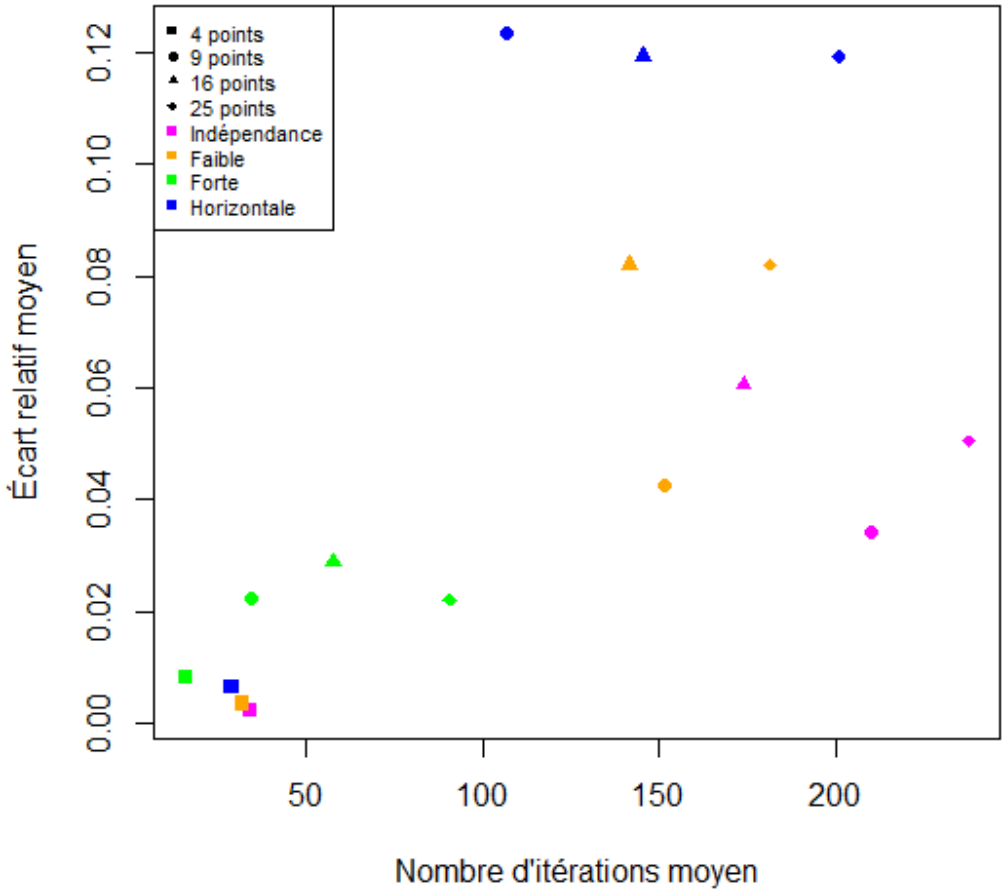


FIGURE 2.9. Écart relatif entre $\mu_2(S^*)$ et $\mathcal{L}_2(S)$ pour les structures de covariance indépendance, faible, forte et horizontale, pour 4, 9, 16 et 25 points.

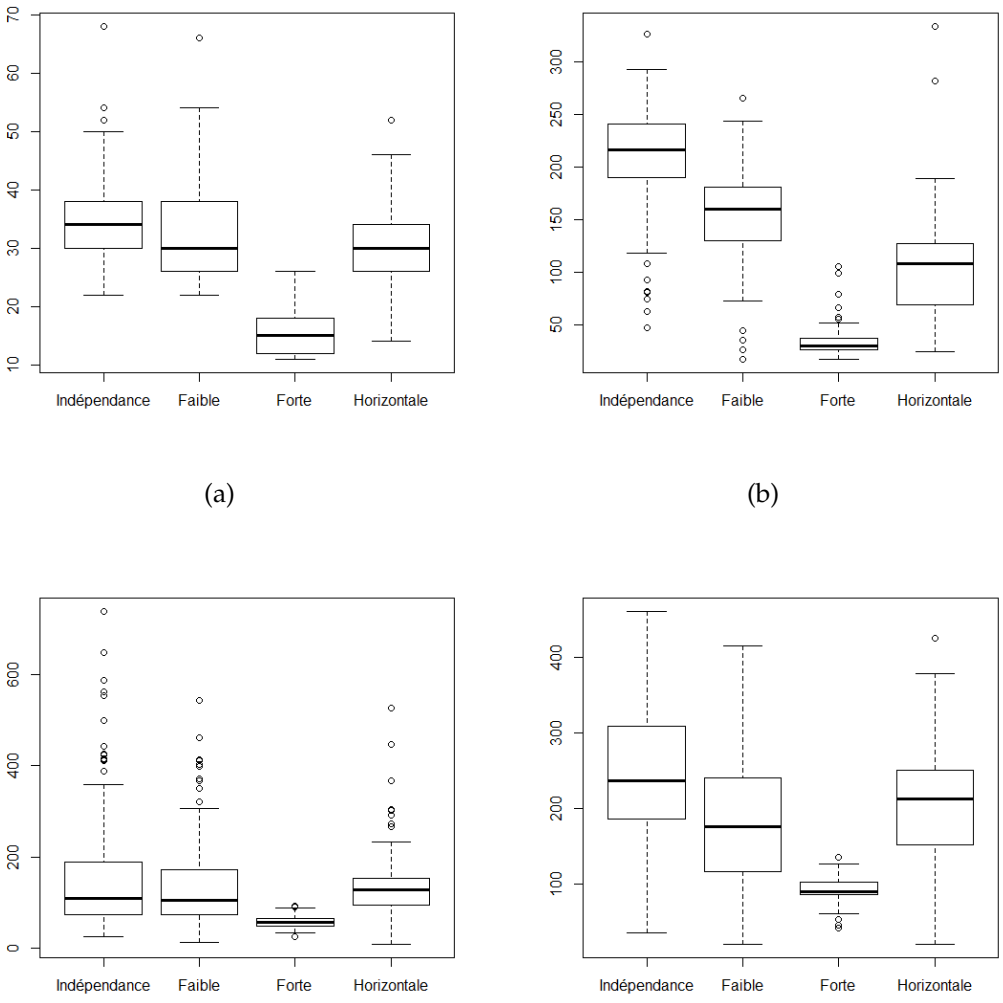
La figure 2.9 permet d'apprécier la relation entre l'écart relatif moyen après la déformation et le nombre d'itérations moyen, et ce pour chaque nombre de points et chaque structure de covariance. Ainsi, nous voyons que pour 4 points, la déformation nécessite peu d'itérations et donne un écart relatif toujours en-dessous de 0,02. Ensuite, nous voyons que peu importe le nombre de points, la structure de covariance isotrope forte mène à de petits écarts relatifs,

pour un petit nombre d'itérations lorsque comparée aux trois autres structures de covariance. Nous l'expliquons par le fait que la corrélation étant forte, la relation d'isotropie entre les résidus est bien marquée. Nous pouvons aussi constater que les structures d'indépendance et de corrélation faible présentent un comportement semblable, ce qui s'explique par la similarité de la structure de dépendance des observations, puisque la corrélation faible ne produit pas une structure de dépendance claire, elle est plus proche de l'indépendance que de la corrélation forte. Finalement, nous observons que la déformation de la structure horizontale, au-delà de 4 points, mène à des écarts relatifs autour de 0,12, clairement plus élevés que ceux des autres structures de covariance.

Maintenant, la figure2.9 permet d'apprécier la relation entre la précision du résultat et le nombre d'itérations nécessaires pour y arriver, mais afin de présenter ces mêmes résultats de façon plus précise, nous présentons des diagrammes en boîte et des tableaux du nombre d'itérations et de l'écart relatif entre $\mu_2(S^*)$ et $\mathcal{L}_2(S)$ à l'arrêt de la déformation de façon séparée. La figure 2.10 présente les diagrammes en boîte du nombre d'itérations et la figure 2.11 présente les diagrammes en boîte de l'écart relatif entre $\mu_2(S^*)$ et $\mathcal{L}_2(S)$. Ensuite les tableaux 2.6 et 2.7 présentent la moyenne et l'écart-type du nombre d'itérations et de l'écart relatif à l'arrêt de la déformation. Nous voyons à la figure 2.10 et au tableau 2.6 que la structure de corrélation forte nécessite nettement moins d'itérations que les trois autres, peu importe le nombre de points. Aussi, pour 4 points, le critère d'arrêt de la déformation est atteint en moins de 50 itérations, alors que pour 9, 16 et 25 points, il en faut généralement plus de 200. Finalement, puisque $\mathcal{L}_2(S)$ représente la valeur que prendrait $\mu_2(S^*)$ si la déformation était parfaite, l'écart relatif entre $\mu_2(S^*)$ et $\mathcal{L}_2(S)$ nous donne une bonne mesure du succès de la déformation en terme de précision. Nous voyons à la figure 2.11 et au tableau 2.7 que la précision de la déformation diminue lorsque le nombre de points augmente. En effet, pour 4 points, l'écart relatif moyen reste inférieur à 1%, pour 9 points il demeure en-dessous de 5%, à l'exception de la covariance horizontale, et pour 16 et 25 points, il est légèrement plus élevé.

TABLE 2.6. Moyenne et écart-type du nombre d'itérations avant l'arrêt de la déformation

Covariance	Moyenne	Écart-type
4 points		
Indépendance	34,27	8,04
Faible	32,22	7,82
Forte	16,08	4,06
Horizontale	29,19	6,74
9 points		
Indépendance	209,73	50,03
Faible	151,67	44,45
Forte	34,79	14,25
Horizontale	106,77	47,29
16 points		
Indépendance	173,85	147,36
Faible	141,54	110,87
Forte	57,96	13,30
Horizontale	145,61	77,61
25 points		
Indépendance	237,57	91,12
Faible	181,51	85,35
Forte	90,92	19,01
Horizontale	200,89	86,18



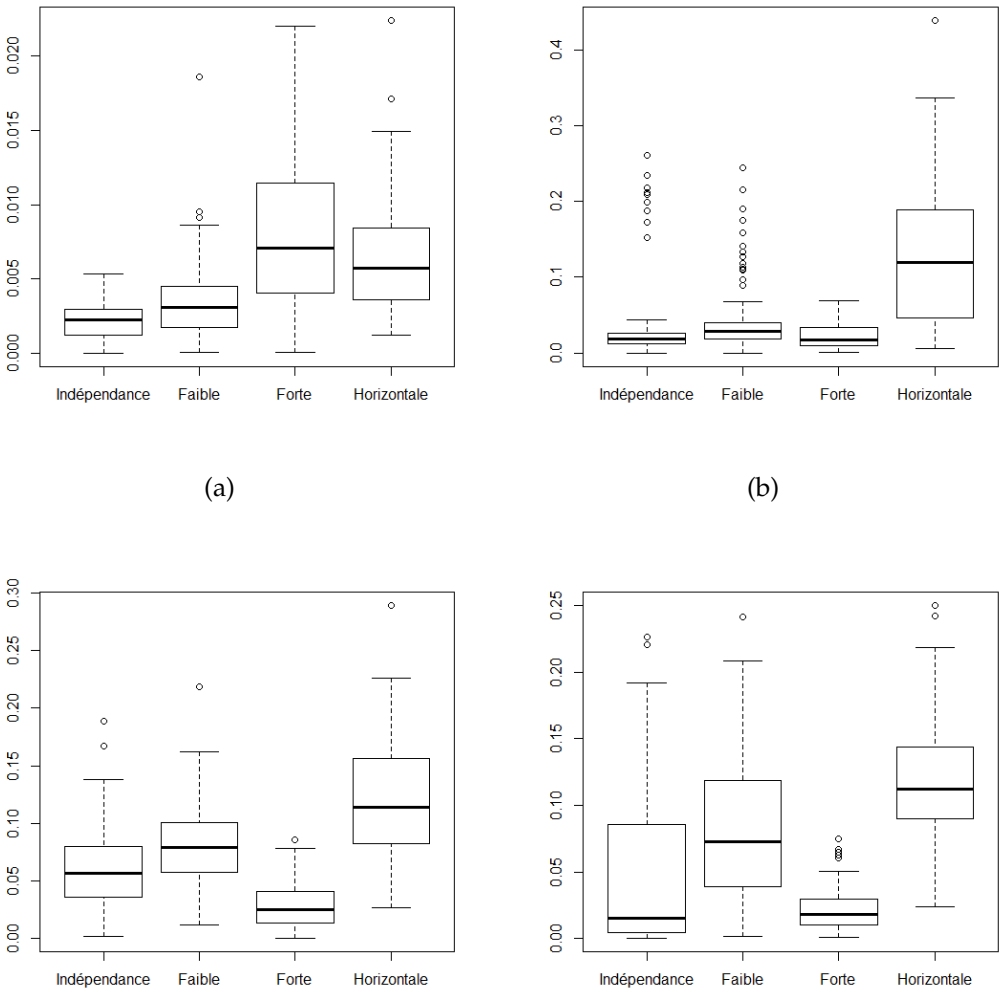
(a) (b)

(c) (d)

FIGURE 2.10. Nombre d'itérations de la déformation pour (a) 4 points (b) 9 points (c) 16 points et (d) 25 points.

TABLE 2.7. Moyenne et écart-type de l'écart relatif entre $\mu_2(S^*)$ et $\mathcal{L}_2(S)$ à l'arrêt de la déformation

Covariance	Moyenne	Écart-type
4 points		
Indépendance	0,00216	0,00121
Faible	0,00341	0,00253
Forte	0,00805	0,00508
Horizontale	0,00654	0,00373
9 points		
Indépendance	0,03426	0,05546
Faible	0,04260	0,04612
Forte	0,02213	0,01628
Horizontale	0,12340	0,08769
16 points		
Indépendance	0,06053	0,03927
Faible	0,08201	0,03380
Forte	0,02888	0,02135
Horizontale	0,11928	0,04780
25 points		
Indépendance	0,05042	0,05848
Faible	0,08211	0,05154
Forte	0,02206	0,01660
Horizontale	0,11927	0,04411



(a) (b)

(c) (d)

FIGURE 2.11. Écart relatif entre $\mu_2(S^*)$ et $\mathcal{L}_2(S)$ après la déformation pour (a) 4 points (b) 9 points (c) 16 points et (d) 25 points.

Dans ce chapitre, nous avons présenté divers concepts permettant d'expliquer la méthode d'inférence topologique. Nous avons ensuite présenté quelques exemples afin d'illustrer la méthode de déformation de l'espace pour pouvoir appliquer la théorie isotrope. Nous avons aussi comparé la méthode de déformation à l'utilisation de la courbure de Lipschitz-Killing. Dans le prochain chapitre, nous présentons une étude de niveau et de puissance comparant la méthode d'inférence topologique avec la correction de Bonferroni et le FDR. Nous présentons aussi l'application de la méthode d'inférence topologique pour décrire l'évolution du changement climatique sur le territoire du Québec.

Chapitre 3

NIVEAU, PUISSANCE ET APPLICATION AU CHANGEMENT CLIMATIQUE

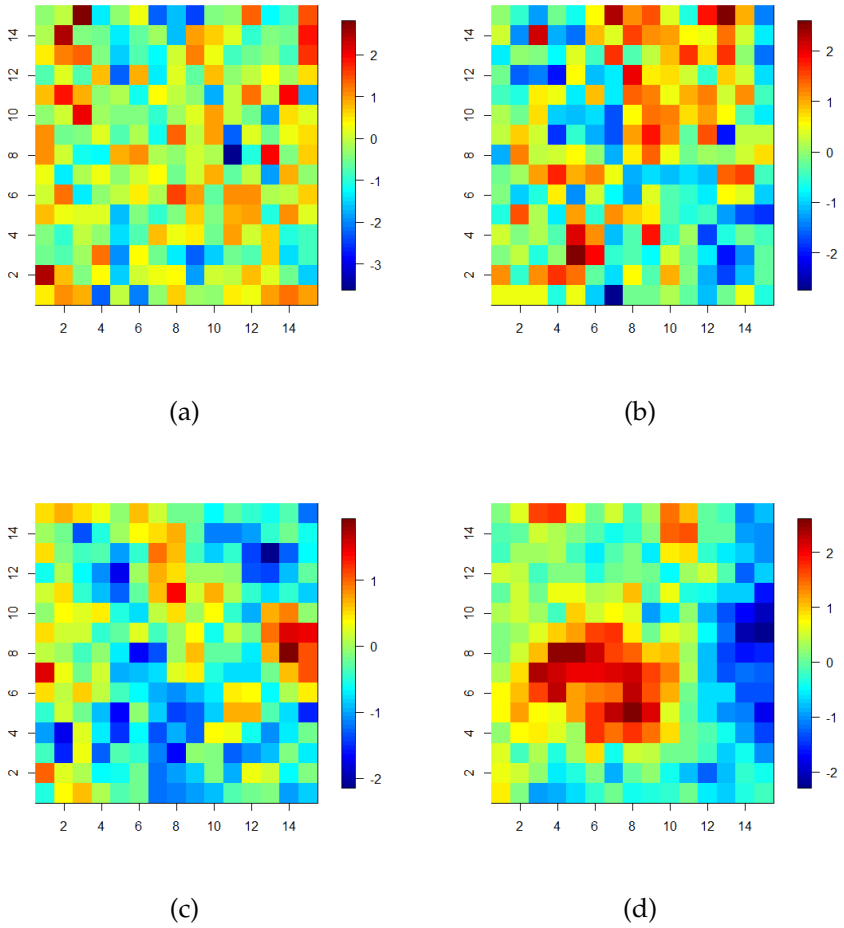
Dans ce chapitre, nous présentons une étude de niveau et de puissance en comparant la méthode de l'inférence topologique avec Bonferroni, le FDR et l'absence de correction pour la multiplicité. Ensuite, nous présentons une application de la méthode d'inférence topologique à un jeu de données de températures simulées.

3.1. NIVEAU ET PUISSANCE

Dans cette section, nous comparons au moyen d'une simulation le niveau et la puissance du test utilisant l'inférence topologique par la courbure de Lipschitz-Killing (CLK) avec ceux de Bonferroni, du FDR, et l'absence de correction pour la multiplicité.

Les champs aléatoires présentés à la figure 3.1 ont été obtenus avec la fonction sim.rf() du package fields en utilisant la fonction de covariance exponentielle où le paramètre θ prend successivement les valeurs 0,01, 0,1, 0,5 et 0,8. Nous voyons à la figure 3.1 que plus le paramètre θ augmente, plus la corrélation de courte portée est grande, ce que nous pouvons observer par le fait qu'il y a davantage de points voisins ayant des valeurs similaires.

À chaque itération de la simulation, nous générons 60 réalisations indépendantes d'un champ aléatoire de dimension 15 par 15, avec le paramètre θ fixé. Les trente premières réalisations $X_{a,1} \dots X_{a,30}$ représentent le groupe a et les suivantes, $X_{b,1} \dots X_{b,30}$, le groupe b sous l'hypothèse nulle ($\mu_a = \mu_b$). Pour simuler l'hypothèse alternative, nous ajoutons aux réalisations du groupe b un champ $\mu_b(s)$ et nous obtenons $X_{b,1} + \mu_b(s) \dots X_{b,30} + \mu_b(s)$, le groupe b sous l'hypothèse alternative ($\mu_a \neq \mu_b$). Le champ $\mu_b(s)$ prend successivement les valeurs $1/15, 2/15, \dots 1$ en allant de la gauche vers la doite et les valeurs sont



(a) (b)

(c) (d)

FIGURE 3.1. Une réalisation d'un champ aléatoire avec le paramètre θ égal à (a) 0,01 (b) 0,1 (c) 0,5 et (d) 0,8.

constantes le long de l'axe vertical, comme représenté à la figure 3.2. Pour obtenir les niveaux et les puissances empiriques, nous comparons les deux groupes a et b au moyen de 225 tests-T au niveau 5%, c'est-à-dire un pour chaque point du champ aléatoire. Les résultats de niveau et de puissance que nous présentons maintenant représentent la moyenne sur les 10 000 répétitions de la simulation. Les niveaux sont présentés au tableau 3.1, et les puissances, à la figure 3.3.

TABLE 3.1. Niveaux empiriques

θ	CLK	Bonf	FDR	Aucun
0,01	0,0200	0,0544	0,0560	1,0000
0,1	0,0224	0,0548	0,0576	0,9996
0,5	0,0348	0,0304	0,0356	0,8948
0,8	0,0380	0,0212	0,0276	0,7440

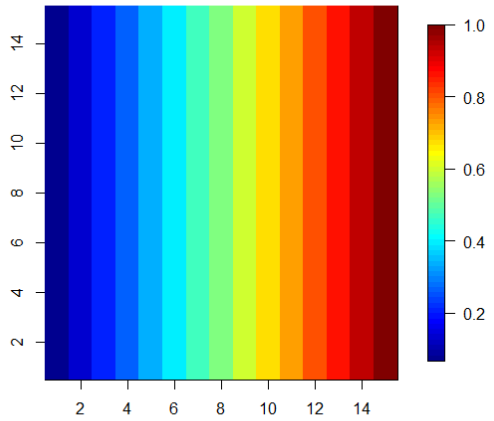
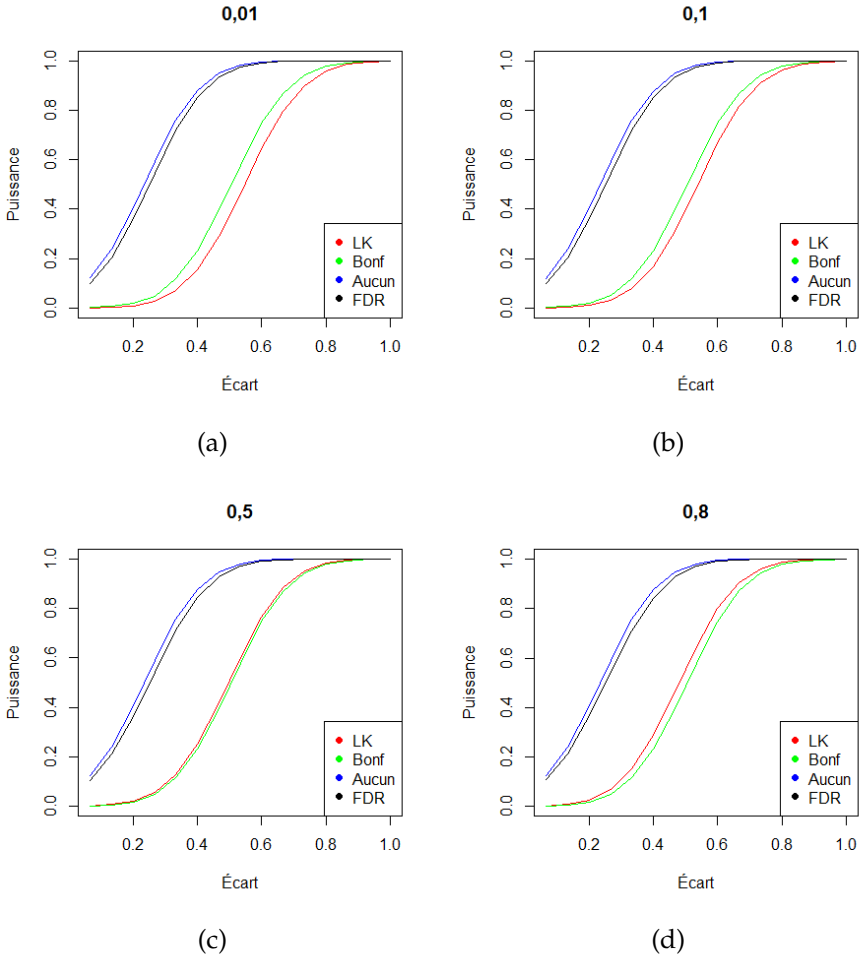


FIGURE 3.2. Centre de la distribution du groupe b sous l'hypothèse alternative

Au tableau 3.1, nous voyons que le niveau empirique de l'inférence topologique augmente avec la corrélation, alors que pour les trois autres méthodes, il diminue. En effet, plus la corrélation est élevée, plus la probabilité de dépasser un seuil donné diminue (il y a moins d'observations effectivement indépendantes). Parmi les méthodes que nous comparons, l'inférence topologique est la seule à tenir compte de la dépendance des observations, à travers la courbure de Lipschitz-Killing. Nous voyons aussi que CLK contrôle le niveau en-dessous de la valeur théorique de 5%. Comme le mentionne Worsley (2003), Bonferroni est conservateur pour détecter le contraste aux points d'échantillonnage, alors que la méthode de l'inférence topologique contrôle le niveau sur l'espace continu entre les points. En pratique, les logiciels d'imagerie utilisent le minimum des seuils de Bonferroni et de l'inférence topologique. Finalement, nous notons qu'en l'absence de correction pour la multiplicité, le niveau empirique observé est de 1 dans le cas indépendant, ce qui signifie que dans chacune des simulations, nous trouvons au minimum une fausse découverte.

La figure 3.3 présente les courbes de puissance, où l'axe horizontal représente l'écart entre le centre des distributions des deux échantillons comparés et l'axe vertical représente la proportion de résultats de test positifs.

Il est important de noter que le FDR contrôle le niveau au sens faible seulement, c'est-à-dire que la procédure est construite de façon à ce que l'espérance de la proportion de fausses découvertes soit égale à α. Sous l'hypothèse nulle globale, cela correspond au niveau, soit la probabilité d'avoir au moins une



(a) (b)

(c) (d)

FIGURE 3.3. Courbes de puissance avec le paramètre θ égal à (a) 0,01 (b) 0,1 (c) 0,5 et (d) 0,8.

fausse découverte. Cependant, lorsqu'une proportion importante des hypothèses ne sont pas nulles, alors ces deux quantités sont différentes. Par conséquent, la courbe de puissance du FDR, comme celle pour l'absence de correction pour la multiplicité, est présentée uniquement à titre informatif et n'est pas comparable aux courbes de puissance de Bonferroni et de l'inférence topologique.

Nous voyons que lorsque θ vaut 0,01, la correction de Bonferroni donne plus de puissance que l'inférence topologique, et lorsque θ vaut 0,1, Bonferroni est toujours plus puissant mais par une marge plus petite. Ensuite, lorsque θ vaut 0,5, les puissances sont pratiquement égales pour les deux méthodes. Finalement, lorsque θ vaut 0,8, la méthode de l'inférence topologique donne

une plus grande puissance. Ceci suggère que pour des champs aléatoires présentant une corrélation de courte portée plus élevée, le gain en puissance par rapport à la méthode de Bonferroni sera encore plus important.

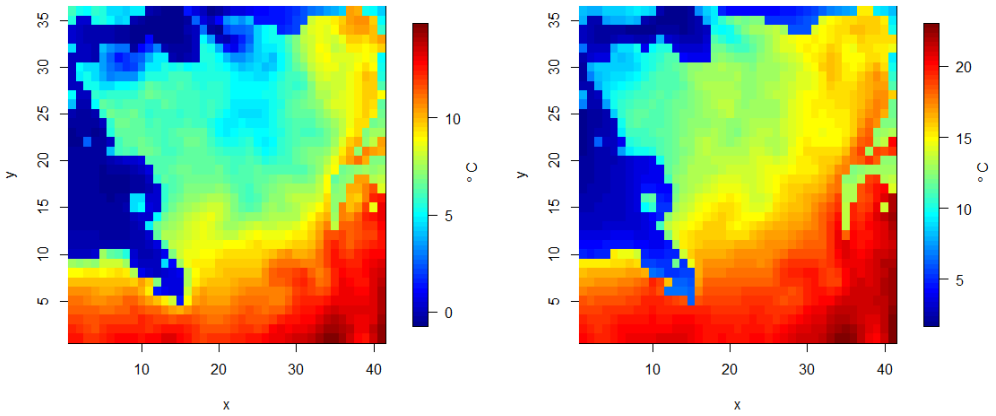
3.2. Présentation des données

Nous présentons maintenant une étude de l'évolution du changement climatique en fonction du temps en procédant à une série de tests utilisant la méthode de la courbure de Lipschitz-Killing. Les données que nous utilisons sont des températures simulées générées et fournies par l'équipe Simulations climatiques d'Ouranos, via la page Web de distribution de données du Centre canadien de la modélisation et de l'analyse climatique (`http://www.cccma.ec.gc.ca/french/data/crcm423/crcm423_aet_sresa2.shtml`). Plus précisément, il s'agit de données mensuelles de la version 4.2.3 de la simulation du modèle régional climatique canadien (MRCC4.2.3) pour les années de 1961 à 2100 (simulation aet). La simulation est pilotée par la version 3 du modèle couplé du climat du globe (MCCG3), suivant le scénario « observé du 20è siècle » du groupe d'experts intergouvernemental sur l'évolution du climat (GIEC) pour les années de 1961 à 2000 et le scénario SRES A2 du GIEC pour les années de 2001 à 2100 (membre #4) sur le domaine nord américain. Le modèle possède une résolution horizontale de 45 km, 29 niveaux verticaux et le pilotage spectral des vents de grande échelle (Music et Caya, 2007).

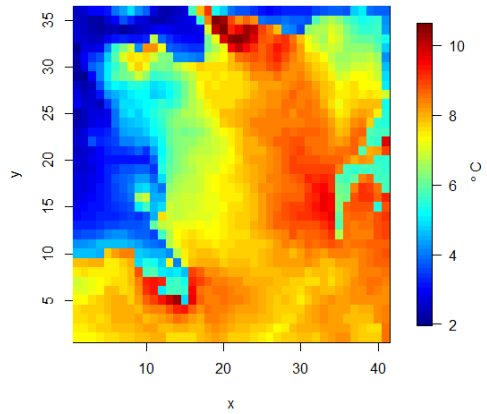
La variable que nous analysons est la température moyenne pour le mois de juillet à deux mètres du sol pour les années 1961 à 2100. Les simulations sont présentées sur une grille de dimensions 182 par 174 représentant l'ensemble de l'Amérique du Nord. De celle-ci, nous avons retenu une sous-grille de dimensions 41 par 36, correspondant grossièrement au territoire du Québec.

À titre d'exemple, nous présentons à la figure 3.4 la température moyenne au mois de juillet pour les années 1961 et 2100 ainsi que le résultat de la soustraction des températures moyennes en 1961 à celles de 2100. Les axes x et y représentent respectivement les axes est-ouest et nord-sud du territoire étudié et les unités sont les indices de ligne et de colonne de la grille de températures. Nous voyons d'abord à la figure 3.4(c) que la différence de température est positive, c'est-à-dire que les températures en 2100 sont plus élevées que celles de 1961 pour l'ensemble du territoire. Ensuite, nous voyons que la différence est plus marquée dans la portion est du territoire.

Maintenant, afin d'évaluer l'évolution du changement climatique, nous comparons une période de référence, soit la première période de trente ans (1961 à



(a) (b)



(c)

FIGURE 3.4. Température moyenne sur le territoire du Québec (a) au mois de juillet 1961 (b) au mois de juillet 2100 et (c) la différence entre juillet 2100 et juillet 1961

1990) avec la suivante (1991-2020), puis nous répétons avec (1992-2021), (1993-2022)...(2071-2100). Nous utilisons des périodes de trente ans, comme cela est fait généralement en climatologie, pour diminuer l'effet des variations annuelles puisque c'est la tendance à long terme qui nous intéresse. Pour chacune des 81 paires d'horizons de temps, nous comparons ainsi deux grilles de 1476 points de façon globale en utilisant la théorie des champs aléatoires et plus spécifiquement, la méthode de la courbure de Lipschitz-Killing.

La première étape consiste à ajuster un modèle linéaire afin d'éliminer une partie de la variabilité due à certains facteurs qui ne font pas partie de notre

analyse, soit la latitude et la longitude. En fait, nous ne travaillons pas directement avec les coordonnées géographiques, nous utilisons plutôt les indices des 36 lignes (x) et des 41 colonnes (y) de la grille. Notez que même s'il n'y avait pas de facteur à éliminer, nous devrions tout de même ajuster un modèle aux données car la méthode utilisant la courbure de Lipschitz-Killing est basée sur l'étude des résidus.

Nous ajustons donc un modèle de régression linéaire multiple avec comme variables explicatives l'indice de ligne (x) et l'indice de colonne (y), et comme variable expliquée la température et nous obtenons

$$\text{Température} = 10{,}28 + 0{,}24x - 0{,}28y - 0{,}0009xy. \tag{3.2.1}$$

Les valeurs prédites par le modèle pour chaque point de la grille sont données à la figure 3.5.

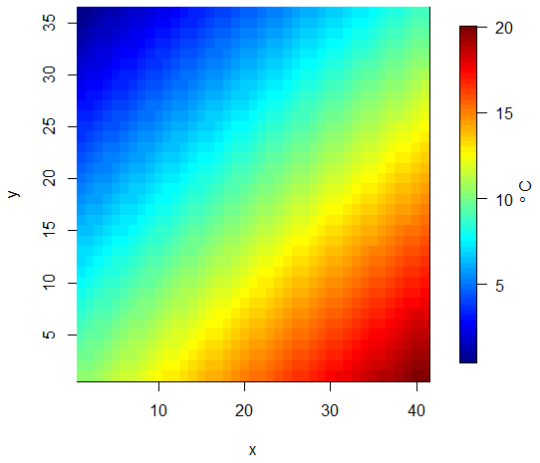


FIGURE 3.5. Valeurs prédites

À la figure 3.6, nous voyons le résultat des 81 comparaisons où les axes x et y sont les mêmes qu'à la figure 3.4, c'est-à-dire qu'il s'agit des axes est-ouest et nord-sud du territoire étudié, avec comme unités les indices de ligne et de colonne de la grille de températures. Quant à l'axe z, il représente la première année de l'horizon de temps ayant été comparé avec la période de référence. Cette figure donne seulement une idée globale du résultat des tests, nous voyons qu'autour de 1991, certains points présentent une différence significative et qu'à partir de 2021, c'est le cas pour tous les points visibles.

Ensuite, à la figure 3.7, nous présentons le résultat des tests d'hypothèses pour six périodes prises à intervalles de 10 ans entre 1991 à 2041. L'année indiquée représente le début de la période de trente ans qui est comparée à la période de référence (1961-1990). Les axes x et y représentent encore une fois les axes est-ouest et nord-sud du territoire, et les unités, les indices de ligne et de colonne de la grille. L'image en couleurs représente la température moyenne observée pour l'année du début de la période, soit 1991, 2001, 2011, 2021, 2031 et 2041, alors que les points de la grille pour lesquels une différence significative est détectée sont indiqués par des cercle noir superposés à l'image en couleur. Nous pouvons voir que pour 1991-2020, le changement est significatif seulement pour quelques points situés dans la région du golfe du fleuve Saint-Laurent. Le changement s'étend graduellement vers le nord et le sud pour occuper toute la partie est du territoire pour la période 2011-2040. Ensuite, le changement s'étend vers l'ouest, et nous voyons que pour 2041-2070 presque tous les points montrent un changement significatif.

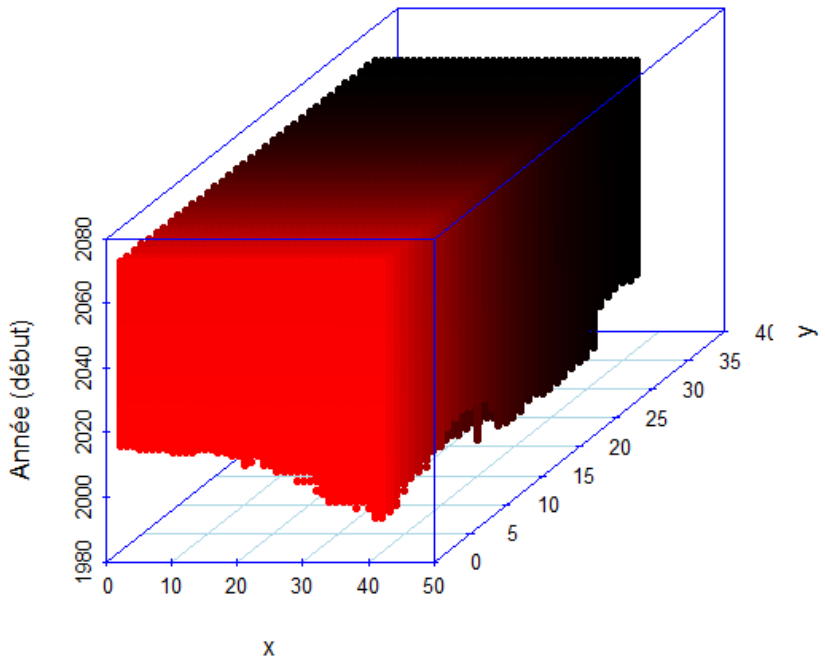


FIGURE 3.6. Distribution géographique du changement climatique en fonction du temps

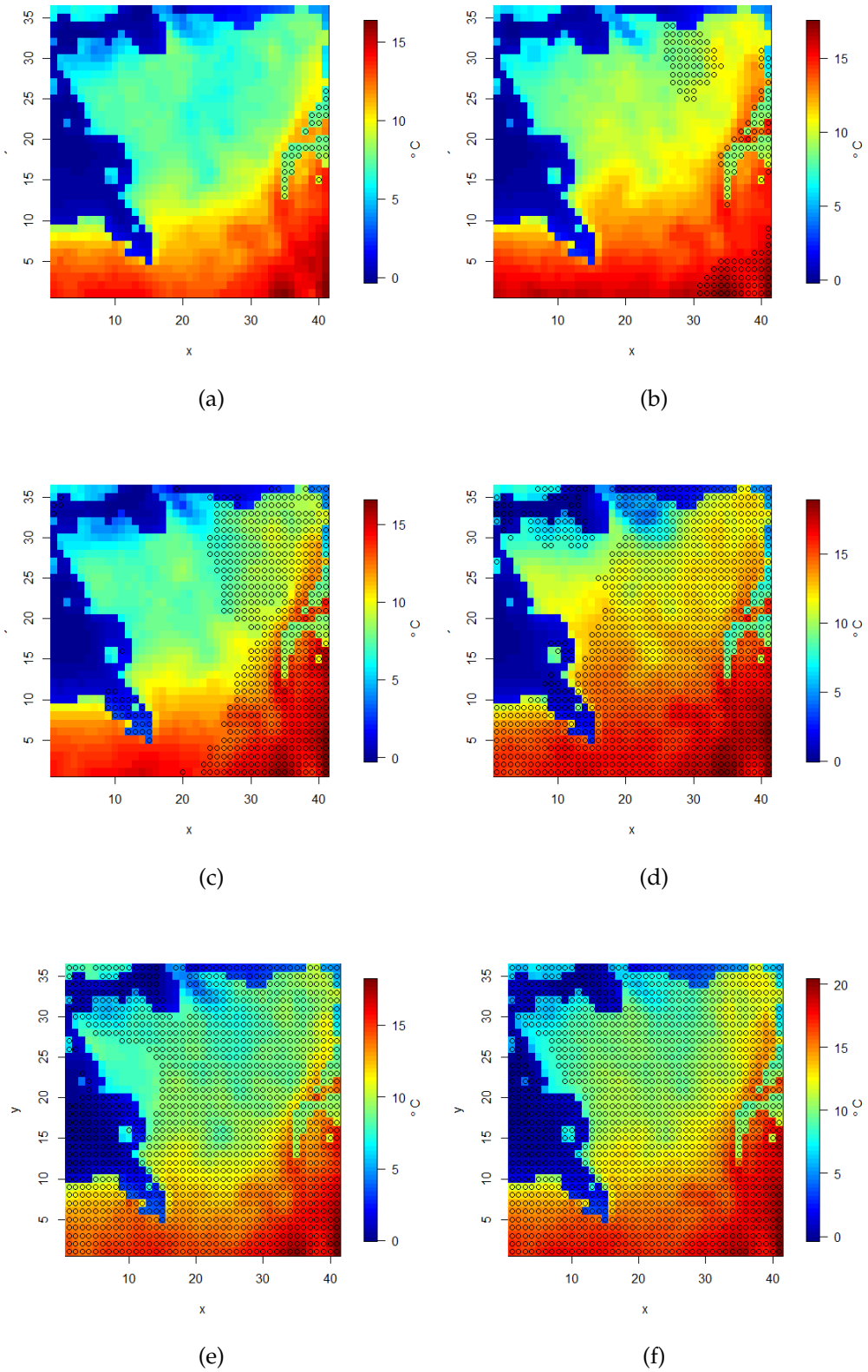


FIGURE 3.7. Points auxquels nous détectons une différence de température significative pour les années (a) 1991-2020 et (b) 2001-2030 (c) 2011-2040 (d) 2021-2050 (e) 2031-2060 et (f) 2041-2070.

CONCLUSION

Dans ce mémoire, nous nous sommes intéressés aux problèmes où les données proviennent d'un échantillonnage fin d'un processus aléatoire continu. Nous avons décrit et utilisé la méthode de l'inférence topologique, permettant de trouver un seuil de signification commun pour un ensemble de tests d'hypothèses, basée sur l'estimation de l'espérance de la caractéristique d'Euler de l'ensemble d'excursion du champ aléatoire au-delà du seuil.

Dans un premier temps, nous avons abordé le problème des comparaisons multiples et nous avons présenté quelques notions reliées à l'étude des champs aléatoires. Nous avons adopté le modèle du champ gaussien unitaire pour les données standardisées. Ensuite, nous avons décrit les concepts nécessaires pour expliquer l'utilisation de la méthode de l'inférence topologique et nous avons donné un exemple illustrant la relation entre le second volume intrinsèque de l'espace déformé vers l'isotropie et la fonction de covariance entre les points d'échantillonnage. Nous avons aussi montré comment le volume intrinsèque de l'espace déformé est égal à la courbure de Lipschitz-Killing de l'espace non déformé lorsqu'une configuration isotrope existe. Puis, nous avons procédé à une étude de niveau et de puissance en comparant l'inférence topologique avec la correction de Bonferroni. Nous avons vu que le niveau empirique de l'inférence topologique est inférieur au niveau théorique pour tous les modèles que nous avons utilisés, mais qu'à mesure que la corrélation de courte portée augmente, le niveau empirique s'élève aussi. Comme nous pouvons nous y attendre, pour un modèle de covariance proche de l'indépendance, la correction de Bonferroni était plus puissante que l'inférence topologique. Cependant, à mesure que la corrélation augmente, la puissance relative de l'inférence topologique se rapproche de celle générée par l'utilisation de la correction de Bonferroni. Finalement, pour illustrer notre approche, nous avons employé l'inférence topologique pour analyser l'évolution du changement climatique sur le territoire du Québec entre les années 1961 et 2100, en utilisant des données simulées de températures.

Afin de pousser plus loin la recherche, il serait intéressant de tester la méthode de l'inférence topologique dans des domaines où les données ne sont pas des images, mais peuvent être représentées sous forme d'images dans un sens plus abstrait. Par exemple, nous pourrions représenter des données épidémiologiques sur un espace paramétré en utilisant comme paramètres des variables cliniques ayant un lien avec le phénomène étudié, comme l'âge, ou certains marqueurs génétiques. De plus, il serait intéressant de chercher à déterminer à l'intérieur de quelles limites, en ce qui a trait à la fonction de covariance et du nombre de points d'échantillonnage (le nombre de pixels dans le cas d'une image), cette méthode est fiable et peut permettre de maximiser la puissance des tests tout en contrôlant le niveau de façon globale.

BIBLIOGRAPHIE

[1] ADLER, R. J., TAYLOR, J. E. ET WORSLEY, K. J. (2009), *Applications of Random Fields and Geometry Foundations and Case Studies*. `http://webee.technion.ac.il/people/adler/hrf.pdf`.

[2] ADLER, R. J., TAYLOR, J. E. (2007), *Random Fields and Geometry*, Springer (Monography in Mathematics), 448 p.

[3] ADLER, R. J. (1981), *The geometry of Random Fields*, Chichester, U. K. :Wiley.

[4] BENJAMINI, Y. (2010), *Discovering the False Discovery Rate*, Journal of the Royal Statistical Society, **B 72**, 405-416.

[5] BENJAMINI, Y. ET HOCHBERG, Y. (1995), *Controlling the False discovery Rate : A Practical ans Powerful approach to Multiple Testing*, Journal of the Royal Statistical Society, **B 57**, 289-300.

[6] CHAMANDY, N., WORSLEY, K.J., TAYLOR, J. ET GOSSELIN, F. (2008), *Titled Euler Characteristic Densities for Central Limit Random Fields, with Application to "Bubbles"*, The Annals of Statistics, **36** (5), 2471-2507.

[7] EFRON, B. (1997), *The Length Heuristic for Simultaneous Hypothesis Tests*, Biometrika, **84**, 143-157.

[8] FUENTES, M. (2001), *A High Frequency Kriging approache for non-stationary environmental processes*, Environmetrics, **12**, 469-484.

[9] FURRER,R., NYCHKA, D. ET SAIN, S (2013), *fields : Tools for spatial data. R package version 6.8*, `http://CRAN.R-project.org/package=fields`.

[10] HIGDON, D., SWALL, J., KERN, J. (1999), *Non-stationary spatial modelling*, Bayesian Statistics **6** (Editeurs) J. M. Bernardo *et al.*. Oxford :Oxford University Press, 761-768.

[11] HUNTER, D. (1976), *An Upper Bound for the Probability of a Union*, Journal of Applied Probability, **13**, 597-603.

[12] LE, N. D. ET ZIDEK, J. V. (2006). *Statistical Analysis of Environmental Space-Time Processes.* , Springer, 341 p.

[13] MUSIC, B., ET D. CAYA (2007). *Evaluation of the Hydrological Cycle over the Mississippi River Basin as Simulated by the Canadian Regional Climate Model.* , Journal of Hydrometeorology, **8** (5), 969-988.

[14] NASH, J. (1956). *The imbedding problem for Riemannian manifolds*, Annals of Mathematics,**63** (1) : 20–63.

[15] RICE, S. O. (1945), *Mathematical Analysis of Random Noise*, Bell System Technical Journal, **24**, 46-156.

[16] SAMPSON, P. D. ET GUTTORP, P. (1992), *Nonparametric Estimation of Nonstationary Spatial Covariance Structure*, Journal of the American Statistical Association, **87**, 108-119.

[17] SUN, J. ET LOADER, C. R. (1994), *Simultaneous Confidence Bands for Linear Regression and Smoothing*, The Annals of Statistics**22**, 1328-1345.

[18] SUN, J. (1993), *Tail Probabilities of the Maxima of Gaussian Random Fields*, The Annals of Probability, **21**, 34-71.

[19] TAYLOR, J., EVANS, A ET FRISTON, K (2009), *A Tribute to : Keith Worsley—1951–2009*, NeuroImage, doi :10.1016/j.neuroimage.2009.04.026.

[20] TAYLOR, J. E. ET WORSLEY, K. J. (2007), *Detecting Sparse Signals in Random Fields, With an Application to Brain Mapping*, Journal of the American Statistical Association, **102** (479), 913-928.

[21] WORSLEY,K. J. (2003), *Detecting Activation in fMRI Data* , Statistical Methods in Medical Research **12**, 401-418.

[22] WORSLEY, K. J., ANDERMANN, M., KOULIS, T., MACDONALD, D. ET EVANS, A.C. (1999), *Detecting Changes in Nonisotropic Images*, Human Brain Mapping **8**, 98-101.

[23] WORSLEY, K. J. (1996) *The Geometry of Random Images* , *Chance*, **9**, 27-40.

[24] WORLSEY, K. J. (1982), *An Improved Bonferroni Inequality and Applications*, Biometrika **69**, 297-302.