# Tests d'homogénéité spatiale basés sur les espacements

## Lionel Cucala & Christine Thomas-Agnan

**Groupe Spatial** 

Mai 2005

#### Plan

- Introduction aux processus ponctuels.
- Théorie des espacements sur [0, 1].
  - De nouvelles statistiques pour tester l'homogénéité spatiale:
- Normalité asymptotique des statistiques.
- Procédure de test multiple.
- Puissance des tests.

4 types principaux:

#### Processus de Poisson homogène



4 types principaux:

#### Processus de Poisson hétérogène



4 types principaux:

## Processus réguliers



4 types principaux:

#### Processus agrégés



 $\left( (X_1, Y_1), \cdots, (X_n, Y_n) \right) \text{ i.i.d. } \sim U([0, 1]^2).$ 

 $((X_1, Y_1), \cdots, (X_n, Y_n))$  i.i.d.  $\sim U([0, 1]^2)$ .

Hypothèse nulle car:

$$\left( (X_1, Y_1), \cdots, (X_n, Y_n) \right)$$
 i.i.d.  $\sim U([0, 1]^2)$ .

Hypothèse nulle car:pas de paramètre à estimer,

$$\left( (X_1, Y_1), \cdots, (X_n, Y_n) 
ight)$$
 i.i.d.  $\sim U([0, 1]^2)$ .

Hypothèse nulle car:
pas de paramètre à estimer,
pas de prédiction possible.

$$\left( (X_1, Y_1), \cdots, (X_n, Y_n) 
ight)$$
 i.i.d.  $\sim U([0, 1]^2)$ .

Hypothèse nulle car:
pas de paramètre à estimer,
pas de prédiction possible.

 $\Rightarrow$  Premier réflexe: tester la CSR.

 tests basés sur les quadrats, (ex.: de type Khi 2)

tests basés sur les quadrats, (ex.: de type Khi 2)
tests basés sur les distances, (ex.: fonction K de Ripley)

- tests basés sur les quadrats, (ex.: de type Khi 2)
- tests basés sur les distances, (ex.: fonction K de Ripley)
- tests basés sur la fonction de répartition empirique. (ex.: Kolmogorov-Smirnov et Cramer-Von Mises)

 $(U_1, \cdots, U_{n-1}) \in [0, 1]^{n-1}$ 

$$(U_1, \cdots, U_{n-1}) \in [0, 1]^{n-1}$$
$$\to U_{(0)} = 0 \le U_{(1)} \le \cdots \le \cdots U_{(n-1)} \le U_{(n)} = 1$$

$$(U_1, \cdots, U_{n-1}) \in [0, 1]^{n-1}$$
  

$$\to U_{(0)} = 0 \le U_{(1)} \le \cdots \le \cdots U_{(n-1)} \le U_{(n)} = 1$$
  

$$\to (D_1, \cdots, D_n), \text{ où } D_i = U_{(i)} - U_{(i-1)}.$$

$$(U_1, \dots, U_{n-1}) \in [0, 1]^{n-1}$$

$$\rightarrow U_{(0)} = 0 \leq U_{(1)} \leq \dots \leq \dots U_{(n-1)} \leq U_{(n)} = 1$$

$$\rightarrow (D_1, \dots, D_n), \text{ où } D_i = U_{(i)} - U_{(i-1)}.$$

$$H_0: \text{ Uniformité et indépendance} \Rightarrow$$

$$(D_i, i = 1, \dots, n) \text{ i.d. mais } \sum_{i=1}^n D_i = 1.$$

$$(U_1, \dots, U_{n-1}) \in [0, 1]^{n-1}$$

$$\rightarrow U_{(0)} = 0 \leq U_{(1)} \leq \dots \leq \dots \cup U_{(n-1)} \leq U_{(n)} = 1$$

$$\rightarrow (D_1, \dots, D_n), \text{ où } D_i = U_{(i)} - U_{(i-1)}.$$

$$H_0: \text{ Uniformité et indépendance } \Rightarrow$$

$$(D_i, i = 1, \dots, n) \text{ i.d. mais } \sum_{i=1}^n D_i = 1.$$

Idée: tester  $H_0$  en observant la dispersion des espacements.

Quelles mesures de dispersion choisir?

Quelles mesures de dispersion choisir?

• Variance 
$$\rightarrow V_n = \frac{1}{n} \sum_{i=1}^n (nD_i - 1)^2$$
.

Quelles mesures de dispersion choisir?

• Variance 
$$\rightarrow V_n = \frac{1}{n} \sum_{i=1}^n (nD_i - 1)^2$$
.

• Déviation en valeur absolue  $\rightarrow R_n = \frac{1}{n} \sum_{i=1}^n |nD_i - 1|.$ 

Quelles mesures de dispersion choisir?

Variance 
$$\rightarrow V_n = \frac{1}{n} \sum_{i=1}^n (nD_i - 1)^2$$
.

• Déviation en valeur absolue  $\rightarrow R_n = \frac{1}{n} \sum_{i=1}^n |nD_i - 1|$ .

Formule générale:

$$S_n = \frac{1}{n} \sum_{i=1}^n g(nD_i) \xrightarrow[n \to \infty]{d} \mathcal{N} \text{ sous } H_0$$
 (Le Cam, 1958).

Quelles mesures de dispersion choisir?

• Variance 
$$\rightarrow V_n = \frac{1}{n} \sum_{i=1}^n (nD_i - 1)^2$$
.

• Déviation en valeur absolue  $\rightarrow R_n = \frac{1}{n} \sum_{i=1}^n |nD_i - 1|$ .

Formule générale:

$$S_n = \frac{1}{n} \sum_{i=1}^n g(nD_i) \xrightarrow[n \to \infty]{d} \mathcal{N} \text{ sous } H_0$$
 (Le Cam, 1958).

Tests basés sur S<sub>n</sub>: puissants contre la dépendance.

GS 2005 – p.10/25





•  $A_{ij} = D_i^x D_j^y$ ,  $i = 1, \cdots, n, j = 1, \cdots, n$ .



$$A_{ij} = D_i^x D_j^y, \qquad i = 1, \cdots, n, \qquad j = 1, \cdots, n.$$
$$S_n = \frac{1}{n^{3/2}} \sum_{i=1}^n \sum_{j=1}^n g(n^2 A_{ij}).$$

Idée 1: 
$$(nD_1, \cdots, nD_n) \sim \left(\frac{E_1}{\overline{E}}, \cdots, \frac{E_n}{\overline{E}}\right)$$
,  
où  $(E_1, \cdots, E_n)$  *i.i.d.*  $\sim \mathcal{E}(1)$ .

Idée 1: 
$$(nD_1, \cdots, nD_n) \sim \left(\frac{E_1}{\bar{E}}, \cdots, \frac{E_n}{\bar{E}}\right)$$
,  
où  $(E_1, \cdots, E_n)$  *i.i.d.*  $\sim \mathcal{E}(1)$ .

$$\Rightarrow S_n \sim G_n = \frac{1}{n^{3/2}} \sum_{i=1} \sum_{j=1} g(\frac{X_i}{\bar{X}} \frac{Y_j}{\bar{Y}})$$
  
où  $(X_1, \cdots, X_n, Y_1, \cdots, Y_n)$  *i.i.d.*  $\sim \mathcal{E}(1)$ .

Idée 1: 
$$(nD_1, \cdots, nD_n) \sim \left(\frac{E_1}{\overline{E}}, \cdots, \frac{E_n}{\overline{E}}\right)$$
,  
où  $(E_1, \cdots, E_n)$  *i.i.d.*  $\sim \mathcal{E}(1)$ .

$$\Rightarrow S_n \sim G_n = \frac{1}{n^{3/2}} \sum_{i=1}^n \sum_{j=1}^n g(\frac{X_i Y_j}{\overline{X}} \overline{\overline{Y}})$$
  
où  $(X_1, \cdots, X_n, Y_1, \cdots, Y_n) i.i.d. \sim \mathcal{E}(1).$ 

Idée 2: Développement de Taylor  $\rightarrow g\left(\frac{X_i Y_j}{\overline{X} \overline{Y}}\right) \sim_{n \to \infty} g(X_i Y_j) - c(\overline{X} - 1) - c(\overline{Y} - 1),$ où  $c = Cov(g(X_1 Y_1), X_1).$
Décomposition:  $G_n = U_n + R_n$ ,

Décomposition: 
$$G_n = U_n + R_n$$
,  
où  $U_n = \frac{1}{n^{3/2}} \sum_{i=1}^n \sum_{j=1}^n [g(X_i Y_j) - c(X_i - 1) - c(Y_j - 1)]$ , et

Décomposition: 
$$G_n = U_n + R_n$$
,  
où  $U_n = \frac{1}{n^{3/2}} \sum_{i=1}^n \sum_{j=1}^n [g(X_i Y_j) - c(X_i - 1) - c(Y_j - 1)]$ , et  
 $R_n =$ 

$$\frac{1}{n^{3/2}} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[ g(\frac{X_i Y_j}{\bar{X}} - g(X_i Y_j) + c(X_i + Y_j - 2)) \right].$$

Décomposition: 
$$G_n = U_n + R_n$$
,  
où  $U_n = \frac{1}{n^{3/2}} \sum_{i=1}^n \sum_{j=1}^n [g(X_i Y_j) - c(X_i - 1) - c(Y_j - 1)]$ , et  
 $R_n =$ 

$$\frac{1}{n^{3/2}} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[ g(\frac{X_i Y_j}{\bar{X}} - g(X_i Y_j) + c(X_i + Y_j - 2)) \right].$$

 $U_n$  est une U-statistique généralisée  $\Rightarrow U_n \xrightarrow[n \to \infty]{d} \mathcal{N}$ .

Décomposition: 
$$G_n = U_n + R_n$$
,  
où  $U_n = \frac{1}{n^{3/2}} \sum_{i=1}^n \sum_{j=1}^n [g(X_i Y_j) - c(X_i - 1) - c(Y_j - 1)]$ , et  
 $R_n =$ 

$$\frac{1}{n^{3/2}} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[ g(\frac{X_i Y_j}{\bar{X}} - g(X_i Y_j) + c(X_i + Y_j - 2)) \right].$$

 $U_n$  est une U-statistique généralisée  $\Rightarrow U_n \xrightarrow[n \to \infty]{d} \mathcal{N}$ .

$$\mathbb{E}(R_n^2) \xrightarrow[n \to \infty]{} 0.$$

Décomposition: 
$$G_n = U_n + R_n$$
,  
où  $U_n = \frac{1}{n^{3/2}} \sum_{i=1}^n \sum_{j=1}^n [g(X_i Y_j) - c(X_i - 1) - c(Y_j - 1)]$ , et  
 $R_n =$ 

$$\frac{1}{n^{3/2}} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[ g(\frac{X_i Y_j}{\bar{X}} - g(X_i Y_j) + c(X_i + Y_j - 2)) \right].$$

 $U_n \text{ est une } U\text{-statistique généralisée} \Rightarrow U_n \xrightarrow[n \to \infty]{d} \mathcal{N}.$ 

$$\mathbb{E}(R_n^2) \xrightarrow[n \to \infty]{} 0.$$

$$\Rightarrow S_n \xrightarrow[n \to \infty]{d} \mathcal{N}.$$

Variance  $\rightarrow V_n = \frac{1}{n^{3/2}} \sum_{i=1}^n \sum_{j=1}^n \{ (n^2 A_{ij} - 1)^2 - 3 \} \xrightarrow[n \to \infty]{d} \mathcal{N}(0, 32).$ 

Variance  

$$\rightarrow V_n = \frac{1}{n^{3/2}} \sum_{i=1}^n \sum_{j=1}^n \{ (n^2 A_{ij} - 1)^2 - 3 \} \xrightarrow[n \to \infty]{d} \mathcal{N}(0, 32).$$

#### Déviation en valeur absolue

$$\to R_n = \frac{1}{n^{3/2}} \sum_{i=1}^n \sum_{j=1}^n \{ |n^2 A_{ij} - 1| - \mu \} \xrightarrow[n \to \infty]{d} \mathcal{N}(0, \sigma^2).$$

Variance  

$$\rightarrow V_n = \frac{1}{n^{3/2}} \sum_{i=1}^n \sum_{j=1}^n \{ (n^2 A_{ij} - 1)^2 - 3 \} \xrightarrow[n \to \infty]{d} \mathcal{N}(0, 32).$$

Déviation en valeur absolue  $\rightarrow R_n = \frac{1}{n^{3/2}} \sum_{i=1}^n \sum_{j=1}^n \{ |n^2 A_{ij} - 1| - \mu \} \xrightarrow[n \to \infty]{d} \mathcal{N}(0, \sigma^2).$ 

Limitation:  $n \leq 100 \Rightarrow$  fractiles empiriques sensiblement différents des fractiles de la loi limite.

#### Premier exemple:

#### Premier exemple:

•  $\mathbf{U}^x = \{U_1^x, \cdots, U_{n-1}^x\}$  i.i.d. U([0, 1]).

#### Premier exemple:

# • $\mathbf{U}^x = \{U_1^x, \cdots, U_{n-1}^x\}$ i.i.d. U([0, 1]). • $\mathbf{U}^y = \mathbf{U}^x$ .

# Premier exemple: $\mathbf{U}^x = \{U_1^x, \cdots, U_{n-1}^x\}$ i.i.d. U([0, 1]). $\mathbf{U}^y = \mathbf{U}^x$ .



#### Second exemple:

#### Second exemple:

# Domaine D du processus ponctuel non rectangulaire.

#### Second exemple:

# Domaine D du processus ponctuel non rectangulaire.



Processus U. Axes d'origine: X et Y.  $\omega = 0$ .

Axes  $X_{\omega}$  et  $Y_{\omega}$ . Construire  $R_{\omega}$  plus grand rectangle inclus dans D de côtés parallèles aux axes.

- Axes  $X_{\omega}$  et  $Y_{\omega}$ . Construire  $R_{\omega}$  plus grand rectangle inclus dans D de côtés parallèles aux axes.
- Repérer les m 1 points de U inclus dans  $R_{\omega}$  et en déduire la statistique  $V_{m,\omega}$ .

- Axes  $X_{\omega}$  et  $Y_{\omega}$ . Construire  $R_{\omega}$  plus grand rectangle inclus dans D de côtés parallèles aux axes.
- Repérer les m 1 points de U inclus dans  $R_{\omega}$  et en déduire la statistique  $V_{m,\omega}$ .
- $\square$   $p_{\omega}$  est la p-valeur associée.

- Axes  $X_{\omega}$  et  $Y_{\omega}$ . Construire  $R_{\omega}$  plus grand rectangle inclus dans D de côtés parallèles aux axes.
- Repérer les m 1 points de U inclus dans  $R_{\omega}$  et en déduire la statistique  $V_{m,\omega}$ .
- $\square$   $p_{\omega}$  est la p-valeur associée.
- $\omega = \omega + \pi/(2a).$

Processus U. Axes d'origine: X et Y.  $\omega = 0$ .

Axes  $X_{\omega}$  et  $Y_{\omega}$ . Construire  $R_{\omega}$  plus grand rectangle inclus dans D de côtés parallèles aux axes.

Repérer les m - 1 points de U inclus dans  $R_{\omega}$  et en déduire la statistique  $V_{m,\omega}$ .

 $\square$   $p_{\omega}$  est la p-valeur associée.

 $\omega = \omega + \pi/(2a).$ 

Itérer jusqu'à  $\omega = \pi/2$ .

*a* p-valeurs:  $p_0, \dots, p_{\pi/2-\pi/(2a)}$ .

a p-valeurs:  $p_0, \cdots, p_{\pi/2-\pi/(2a)}$ .  $\rightarrow p_{(1)} \leq \cdots \leq p_{(a)}$ .

 $a \text{ p-valeurs: } p_0, \cdots, p_{\pi/2-\pi/(2a)}.$  $\rightarrow p_{(1)} \leq \cdots \leq p_{(a)}.$ 

Test de niveau nominal  $\alpha$ . On rejette  $H_0$  ssi:

a p-valeurs:  $p_0, \cdots, p_{\pi/2-\pi/(2a)}$ .  $\rightarrow p_{(1)} \leq \cdots \leq p_{(a)}$ .

Test de niveau nominal  $\alpha$ . On rejette  $H_0$  ssi:  $p_{(1)} < \alpha/a$  (Bonferroni).

a p-valeurs:  $p_0, \cdots, p_{\pi/2-\pi/(2a)}$ .  $\rightarrow p_{(1)} \leq \cdots \leq p_{(a)}$ .

Test de niveau nominal  $\alpha$ . On rejette  $H_0$  ssi:  $p_{(1)} < \alpha/a$  (Bonferroni).  $\exists i = 1, \cdots, a$   $p_{(i)} < i\alpha/a$  (Simes).

Bonferroni: Erreur  $1^{\text{ère}}$  espèce  $\leq \alpha$ .

Bonferroni: Erreur 1<sup>ère</sup> espèce ≤ α.
 Simes: puissance supérieure.

- Bonferroni: Erreur  $1^{\text{ère}}$  espèce  $\leq \alpha$ .
- Simes: puissance supérieure.
- Estimation de l'erreur de 1<sup>ère</sup> espèce pour  $\alpha = 0.04$ :
Description de la procédure multiple

Bonferroni: Erreur  $1^{\text{ère}}$  espèce  $\leq \alpha$ .

Simes: puissance supérieure.

Estimation de l'erreur de 1<sup>ère</sup> espèce pour  $\alpha = 0.04$ :

a	Erreur l estimée:								
	$BV_n$	$SV_n$	$BR_n$	$SR_n$					
5	0.037	0.038	0.0415	0.042					
10	0.0345	0.0355	0.0375	0.0375					

Description de la procédure multiple

Bonferroni: Erreur  $1^{\text{ère}}$  espèce  $\leq \alpha$ .

Simes: puissance supérieure.

Estimation de l'erreur de 1<sup>ère</sup> espèce pour  $\alpha = 0.04$ :

a	Erreur l estimée:								
	$BV_n$	$SV_n$	$BR_n$	$SR_n$					
5	0.037	0.038	0.0415	0.042					
10	0.0345	0.0355	0.0375	0.0375					

Nous adoptons Simes avec a = 10.

4 jeux de données respectivement considerés homogène, agrégé, régulier et hétérogène.

# 4 jeux de données respectivement considerés homogène, agrégé, régulier et hétérogène.

Statistique	Resultats pour les données suivantes								
	Japanese pines	Redwoods	<b>Biological cells</b>	Scouring rushes					
$V_n$	< 0.002	0.042	0.064	0.024					
$R_n$	< 0.002	< 0.002	0.832	0.044					
$MV_n$	< 0.02	<0.02	0.06	0.248					
$MR_n$	< 0.02	< 0.02	0.24	0.54					
$ar{\omega}^2$	0.712	0.692	0.006	0.004					
$D_n$	0.26	0.908	0.014	0.044					
T	0.915	< 0.001	< 0.001	0.936					
Li	0.918	< 0.002	< 0.002	0.102					
$L_m$	0.90	< 0.01	< 0.01	0.32					

### Japanese Pines:



#### Japanese Pines:



Processus de Poisson hétérogène (grille) d'intensité  $\lambda(x, y) = \lambda_1(x)\lambda_2(y)$ où  $\lambda_1(x) = \max_{x_1, \dots, x_m} \exp(-c_x |x - x_i|)$ et  $\lambda_2(y) = \max_{y_1, \dots, y_l} \exp(-c_y |y - y_j|).$ 

# Processus régulier:

$\epsilon$	Puissance estimée:								
	$L_m$	$ar{\omega}^2$	$D_n$	$V_n$	$R_n$	$SV_n$	$SR_n$		
0.03	0.467	0.053	0.047	0.059	0.045	0.064	0.054		
0.05	1	0.121	0.099	0.078	0.062	0.117	0.089		
0.07	1	0.324	0.279	0.126	0.101	0.204	0.154		

### Processus régulier:

$\epsilon$	Puissance estimée:								
	$L_m$	$ar{\omega}^2$	$D_n$	$V_n$	$R_n$	$SV_n$	$SR_n$		
0.03	0.467	0.053	0.047	0.059	0.045	0.064	0.054		
0.05	1	0.121	0.099	0.078	0.062	0.117	0.089		
0.07	1	0.324	0.279	0.126	0.101	0.204	0.154		

#### Processus agrégé:

$\mu$	ρ	t	Puissance estimée:						
			Li	$ar{\omega}^2$	$D_n$	$V_n$	$R_n$	$SV_n$	$SR_n$
10	10	0.15	0.998	0.849	0.768	0.817	0.712	0.932	0.864
10	10	0.25	0.752	0.716	0.605	0.469	0.371	0.556	0.433
20	5	0.3	0.869	0.852	0.760	0.659	0.541	0.702	0.564

### Processus hétérogène (tendance plane):

$ heta_1$	$ heta_2$	Puissance estimée:							
		Li	$ar{\omega}^2$	$D_n$	$V_n$	$R_n$	$SV_n$	$SR_n$	
4	4	0.230	0.795	0.766	0.191	0.151	0.129	0.096	
6	6	0.378	0.937	0.914	0.326	0.237	0.229	0.164	
8	4	0.339	0.906	0.888	0.293	0.222	0.191	0.150	

#### Processus hétérogène (tendance plane):

$ heta_1$	$ heta_2$	Puissance estimée:							
		Li	$ar{\omega}^2$	$D_n$	$V_n$	$R_n$	$SV_n$	$SR_n$	
4	4	0.230	0.795	0.766	0.191	0.151	0.129	0.096	
6	6	0.378	0.937	0.914	0.326	0.237	0.229	0.164	
8	4	0.339	0.906	0.888	0.293	0.222	0.191	0.150	

#### Processus hétérogène (grille d'angle 0):

m	С	Puissance estimée:							
		Т	$ar{\omega}^2$	$D_n$	$V_n$	$R_n$	$SV_n$	$SR_n$	
5	25	0.361	0.026	0.045	0.680	0.724	0.302	0.431	
5	30	0.582	0.024	0.055	0.893	0.916	0.581	0.748	
7	30	0.095	0.028	0.030	0.309	0.450	0.083	0.201	
7	40	0.162	0.029	0.036	0.695	0.862	0.291	0.626	

### Processus hétérogène (grille d'angle $\pi/3$ ):

m	С	Puissance estimée:								
		Т	$ar{\omega}^2$	$D_n$	$V_n$	$R_n$	$SV_n$	$SR_n$		
7	40	0.638	0.029	0.033	0.039	0.038	0.350	0.711		
7	50	0.737	0.050	0.050	0.045	0.039	0.493	0.870		
9	50	0.227	0.047	0.049	0.037	0.029	0.130	0.632		
9	60	0.309	0.042	0.044	0.041	0.046	0.157	0.779		

### Processus hétérogène (grille d'angle $\pi/3$ ):

m	С	Puissance estimée:							
		Т	$ar{\omega}^2$	$D_n$	$V_n$	$R_n$	$SV_n$	$SR_n$	
7	40	0.638	0.029	0.033	0.039	0.038	0.350	0.711	
7	50	0.737	0.050	0.050	0.045	0.039	0.493	0.870	
9	50	0.227	0.047	0.049	0.037	0.029	0.130	0.632	
9	60	0.309	0.042	0.044	0.041	0.046	0.157	0.779	

#### p-valeurs minimales:



Les méthodes basées sur les espacements sont utiles pour tester l'origine d'une forêt.

Les méthodes basées sur les espacements sont utiles pour tester l'origine d'une forêt.

Estimation des paramètres du processus.

- Les méthodes basées sur les espacements sont utiles pour tester l'origine d'une forêt.
- Estimation des paramètres du processus.
- Différentes alternatives à la CSR ⇒ différents tests pour les détecter.

- Les méthodes basées sur les espacements sont utiles pour tester l'origine d'une forêt.
- Estimation des paramètres du processus.
- Différentes alternatives à la CSR ⇒ différents tests pour les détecter.
- Généralisation à la 3D.

- Les méthodes basées sur les espacements sont utiles pour tester l'origine d'une forêt.
- Estimation des paramètres du processus.
- Différentes alternatives à la CSR ⇒ différents tests pour les détecter.
- Généralisation à la 3D.
- Généralisation aux espacements d'ordre supérieur.