

# Modélisation et Statistique Spatiale

**Atelier RASMA – Université Gaston Berger**

**Saint Louis du Sénégal**

*29 novembre – 4 décembre 2010*

*Xavier Guyon – SAMM -- Université Paris 1*

- Géostatistique, modèle du second ordre, krigeage
- Donnée sur un réseau :
  - Auto-Régression Spatiale (SAR et CAR, SARX)
  - Champ de Gibbs – Markov – Auto modèle de Besag
  - Simulation par chaîne de Markov (MCMC)
- Processus ponctuel

---

# *Trois types de structures spatiales*

## Géostatistiques :

- S est un **ensemble continu** (de  $\mathbb{R}^{**2}$ ,  $\mathbb{R}^{**3}$ , ...)
- données **réelles**, uni ou multidimensionnelles
- Observations en n sites :  $s_1, s_2, \dots, s_n$  de S.

## Latticielles :

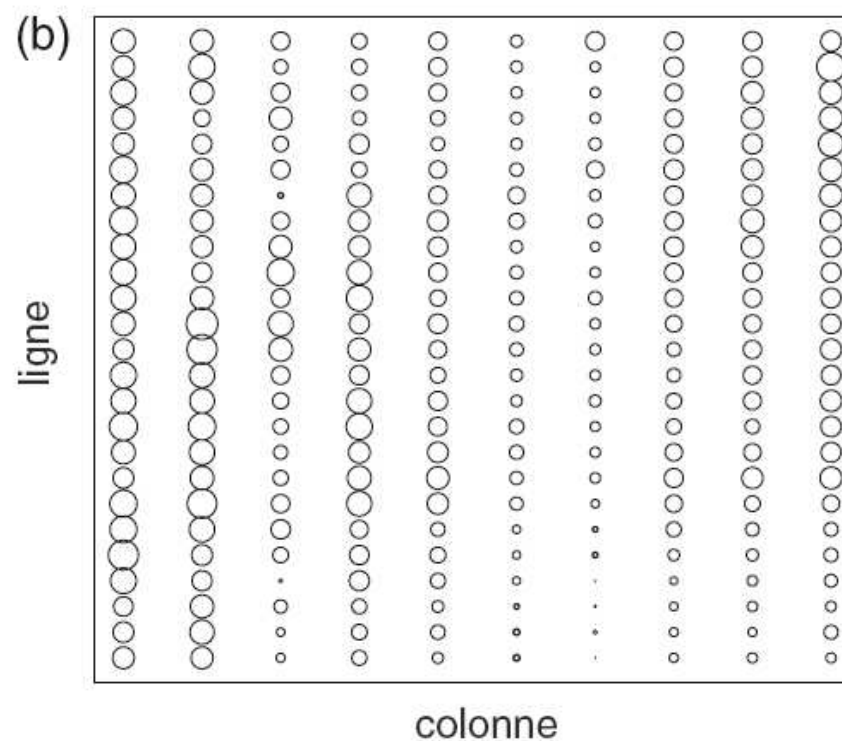
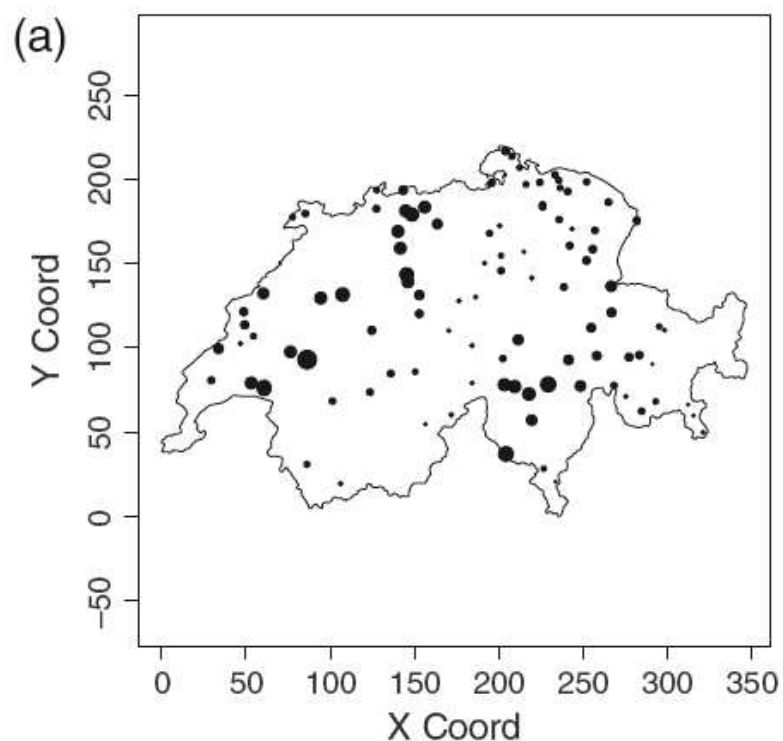
- S est un **réseau fini de sites discret** muni d'un graphe
- données réelles : modèles AR.
- ou non : champ de Markov (binaires, poisson, etc.)

## Données ponctuelles

- S est un ensemble de **points aléatoires**  $x_1, x_2, \dots, x_n$  de S, une partie de  $\mathbb{R}^{**2}$ ,  $\mathbb{R}^{**3}$ , ... (Processus Ponctuel Spatial).
- PP marqué si une marque  $m_i$  attaché à chaque  $\mathbf{x}_i$  : épicode d'un séisme et  $\mathbf{m}_i$  = intensité du séisme :  $\{(\mathbf{x}_i, \mathbf{m}_i), i=1, n\}$ .

# Données géostatistiques **X**

- (a) Cumul de pluies dans 100 stations météo suisse le jour du passage du nuage de Tchernobyl : réseau irrégulier (`sic.100` de `geoR`)
  - (b) Porosité d'un sol (`soil250` de `geoR`) : réseau régulier
- La dimension des symboles est proportionnelle à **X***



# Le logiciel **R**

Installation de **R** :

<http://cran.r-project.org/>

- **site miroir** : i.e. Toulouse
- Deux fenêtres : **R Console** (*RGui*) et **R Graphics**

Chargement du package *geoR* (données géo-stat)

# Données Porosité (*soil250*)

22 variables « chimiques » sur une grille régulière 10x25 points espacés de 5 mètres (cf. *soil250* dans la liste de *geoR*).

On sélectionne la coordonnées n°16, *ctc* (*catium exchange*)

```
> data(soil250)
> ctc <- as.geodata(soil250, data.col=16)
> plot(ctc)
```

## 4 graphiques

- 1 - les 4 quartiles (4 couleurs) de *CTC*
- 2 et 3 - les nuages (*ctc(x,y), y*) et (*x,ctc(x,y)*)
- 4 - Histogramme de répartition des 250 valeurs de *ctc*

### **Conservation d'un graphique :**

se placer dans la *fenêtre graphique* → historique → Ajouter (ou précédent, etc....)

**Autre solution :** placer la commande « `> x11()` » avant une commande graphique conservera le graphique (aller dans fenêtre, les graphiques sont numérotés séquentiellement)

# Données pluviométrie Suisse

```
> print(sic.100)  
> points(sic.100,borders=sic.borders)
```

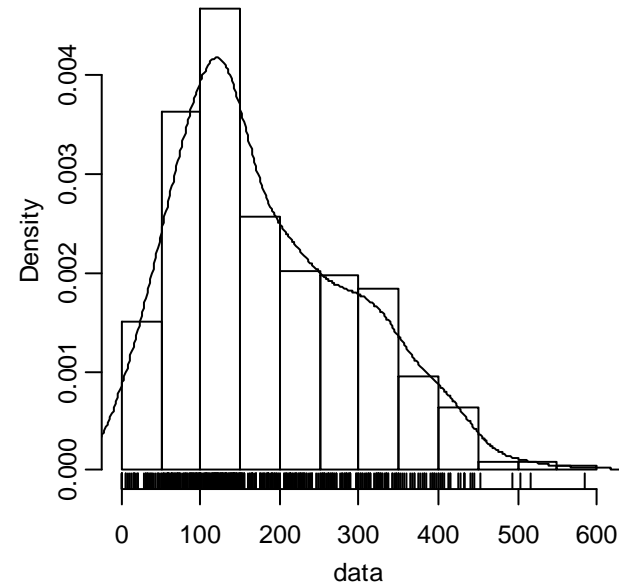
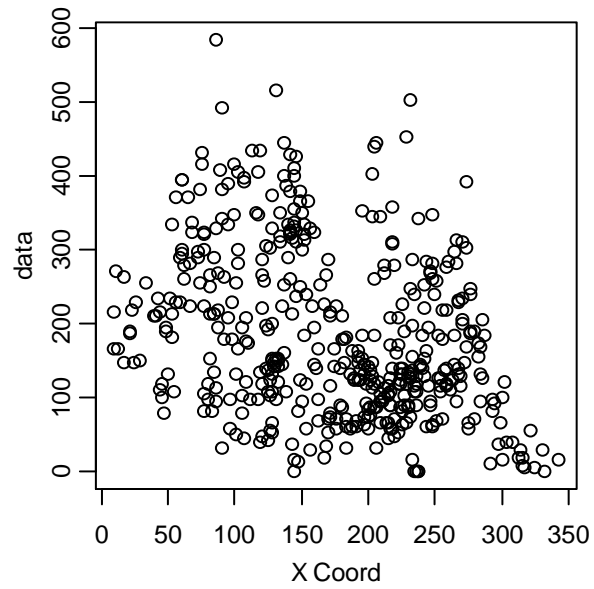
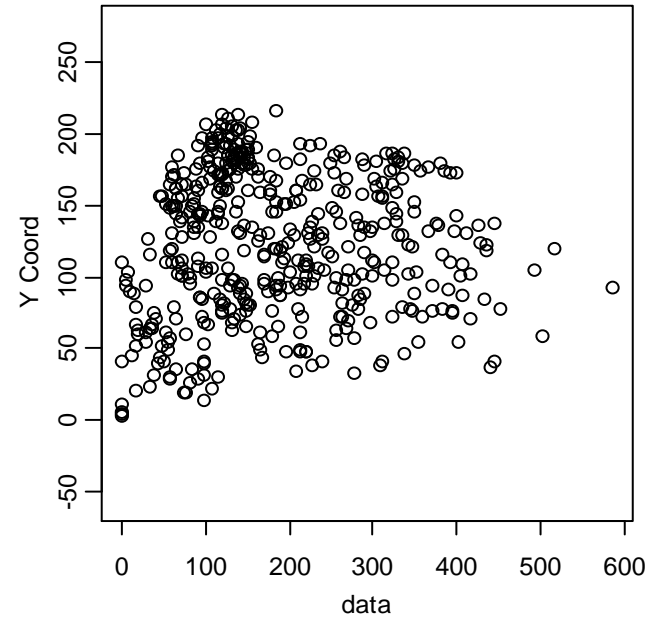
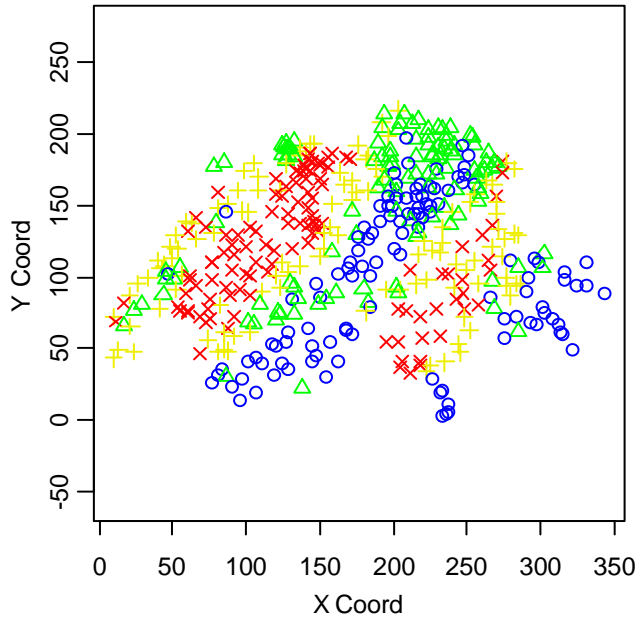
**sic.borders** : fichier frontière

4 données pour chaque station : coordonnées  $(x,y)$ , *hauteur de pluie*, *altitude*

**sic.100** : 100 stations choisies au hasard dans un réseau de 367 stations

**sic.all** : toutes les 367 stations

```
> points(sic.all, borders=sic.borders)  
> plot(sic.all)
```



# Questions en Geostatistique

- **Quelle structure de corrélation spatiale ?**
  - Stationnarité (covariance) , isotropie ?
  - Non stationnarité (variogramme)
  - Modèle avec covariables (données exogènes)
- **Estimation (validation) de modèle**
- **Prédiction** partout : carte de krigeage, simulation conditionnelle
- **Outil logiciel** : *geoR*



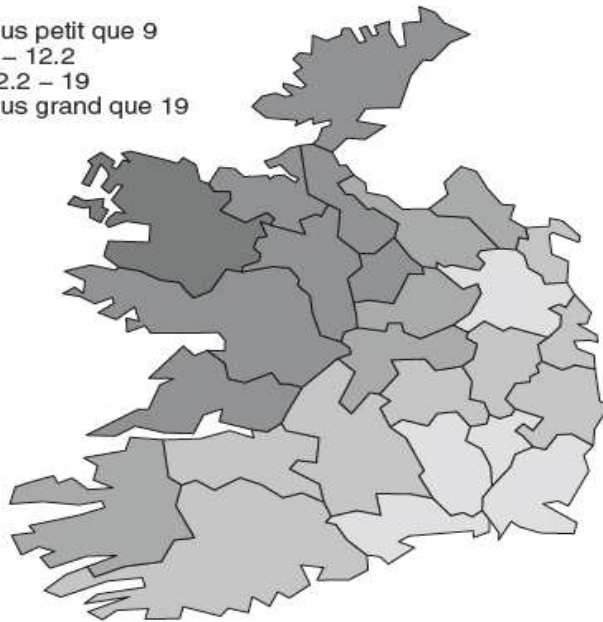
## Données **réelles** sur un réseau discret

(a) % groupe sanguin A dans 26 comtés Irlande (*eire*, *spdep*)

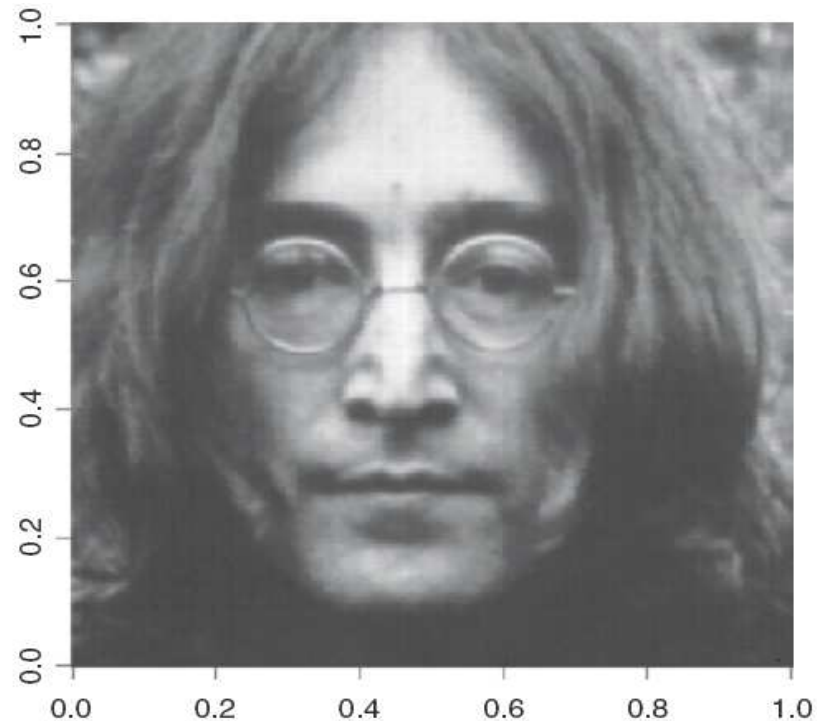
(b) Image 256 x 256 de J. Lennon (193 niveaux de gris, *lennon* du package *fields*)

→ Packages : *spdep*, *fields*, ...

□ plus petit que 9  
□ 9 – 12.2  
□ 12.2 – 19  
■ plus grand que 19



(a)



(b)

## Pourcentage du groupe sanguin A dans les 26 comtés de l'Irlande (données *eire* de *spdep*)

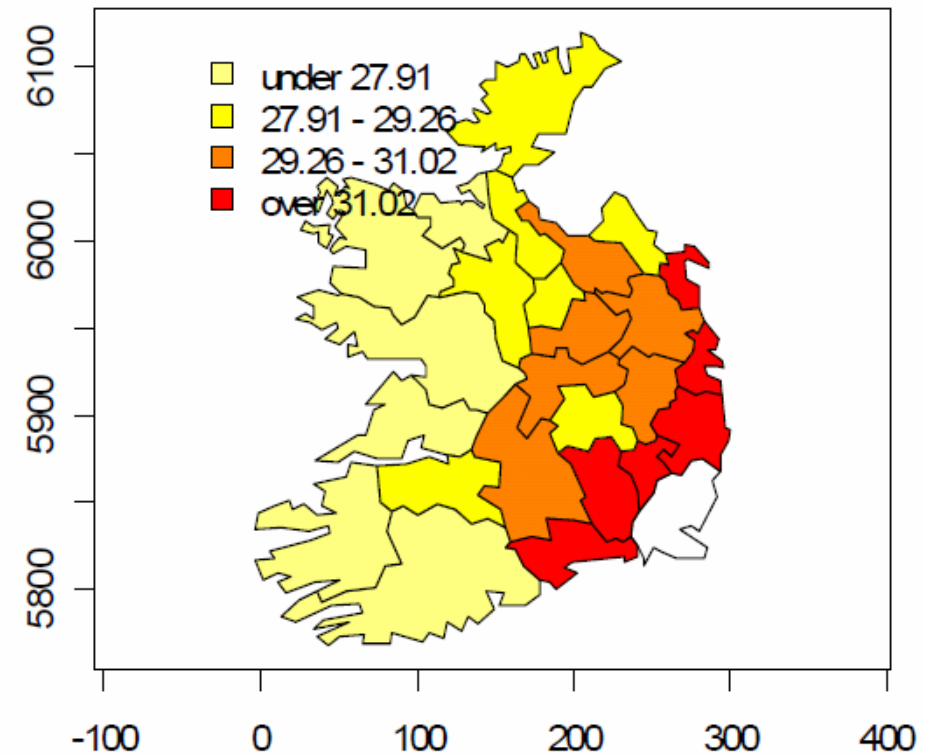
$X(s)$  = % du groupe sanguin A dans 26 comtés.

- Graphe de contiguïté spatiale,
- Indice de Moran :  $t = 4.66$  (corrélation spatiale significative).

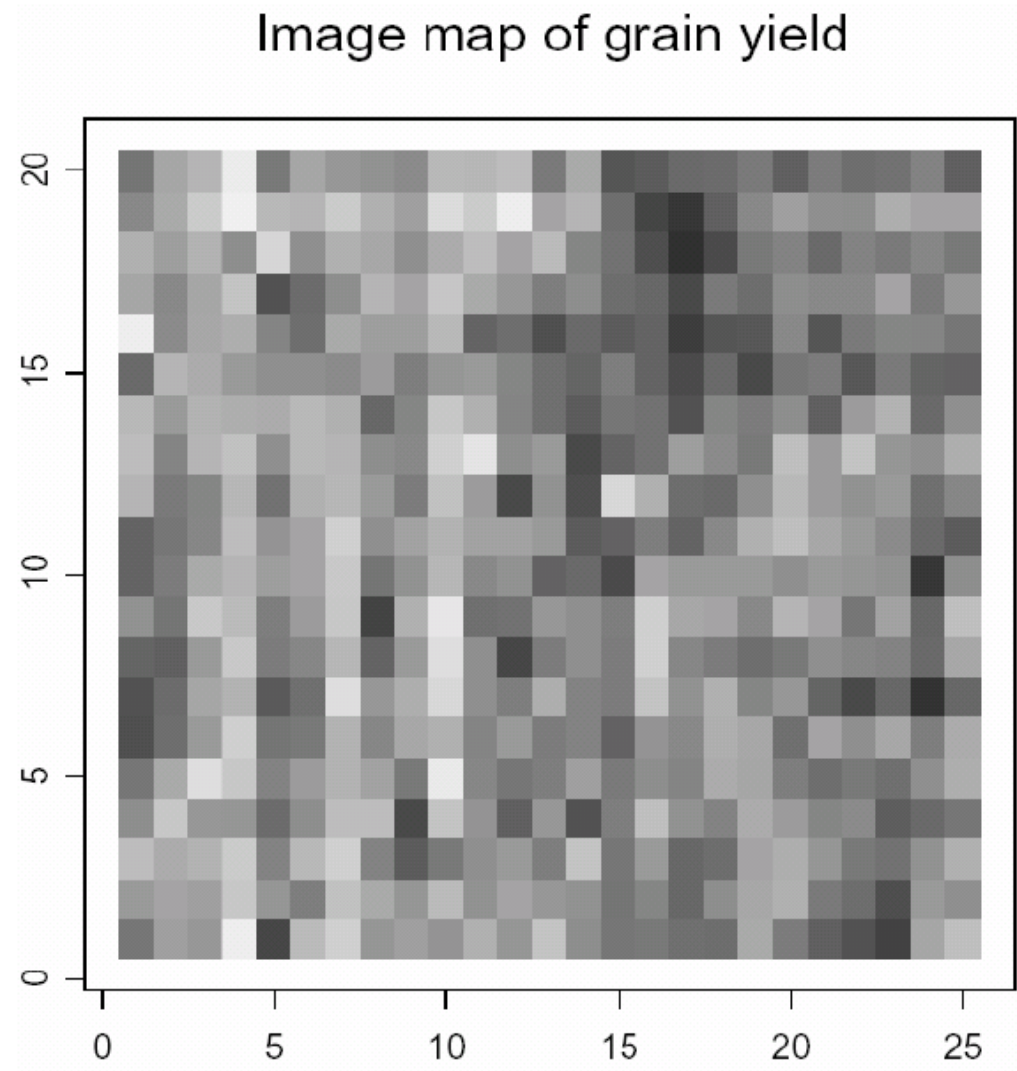
### Covariables

- *Towns* : densité urbaine,
- *Pale* : indique si le comté était sous contrôle anglo normand ou non.

Percentage with blood group A in Eire

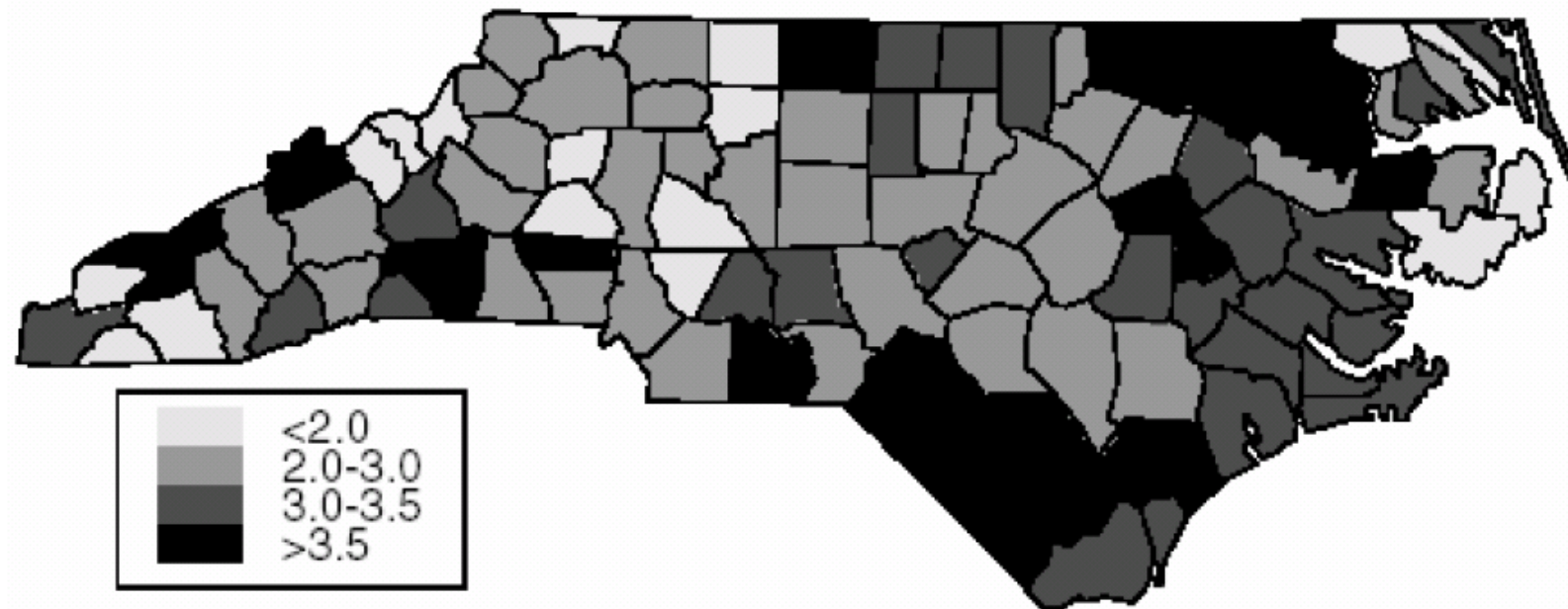


**Données de Mercer et Hall** : image en pixels de niveau de gris.



## Données latticielles : « la mort subite du nourrisson »

Nombre de cas dans 100 comtés de Caroline du nord entre 1974-1978 (données *sids* de *R*; Cressie,1993)



# Questions

- **Quel modèle ?**
  - voisinages d'influence pour chaque site
  - SAR ou CAR
  - stationnaire ou non
  - avec variables exogènes (SARX)
- **Estimation et validation de modèle**
- **Tests sur les paramètres ....**
- **Outils logiciel : *spdep, fields, ...***

## Données SIDS : la mort subite du nourrisson (suite)

**Données :**  $X(s)$  = taux de msn pour le canton  $s$  dans 100 cantons de north carolina, (1974-78), réelles.

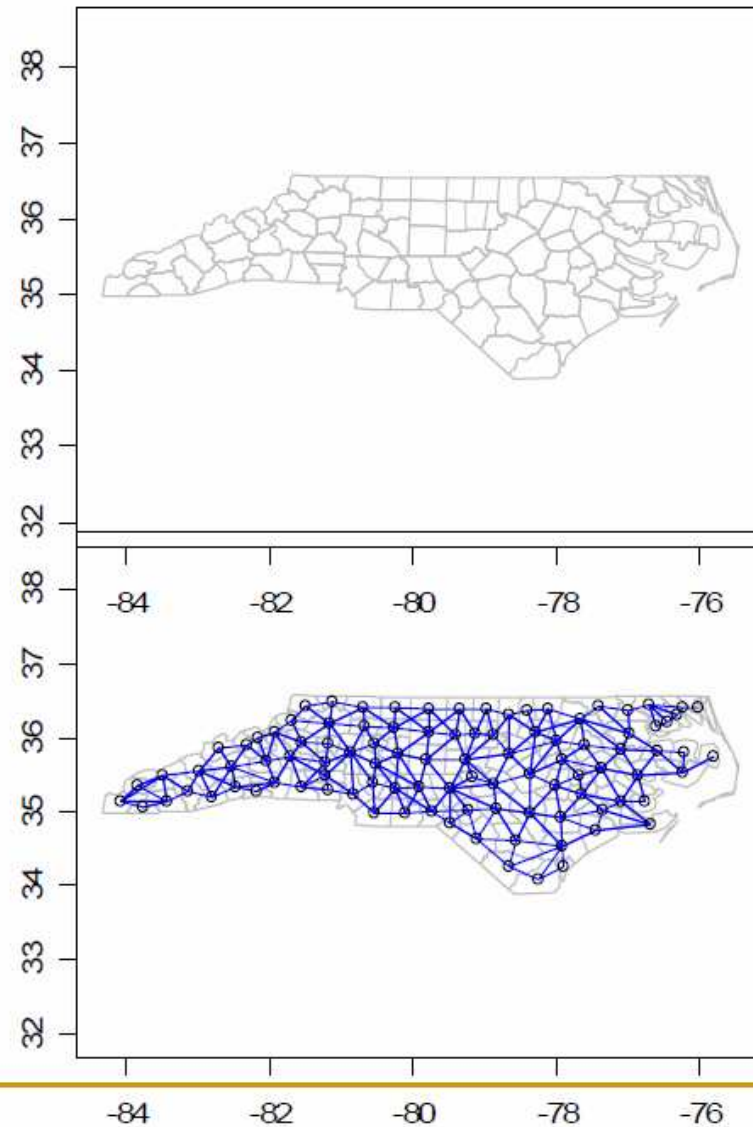
**Cantons et graphe de voisinage**

**Covariables (en  $s$ ) :**

- Taux  $X(ds)$  au voisinage  $s$ ,
- Nombre total de naissances,
- Pourcentage par communauté, etc

**Questions :**

- Auto corrélation spatiale ?
- Modélisation régression spatiale, SAR, régression AR, SARX (avec eXogènes)....



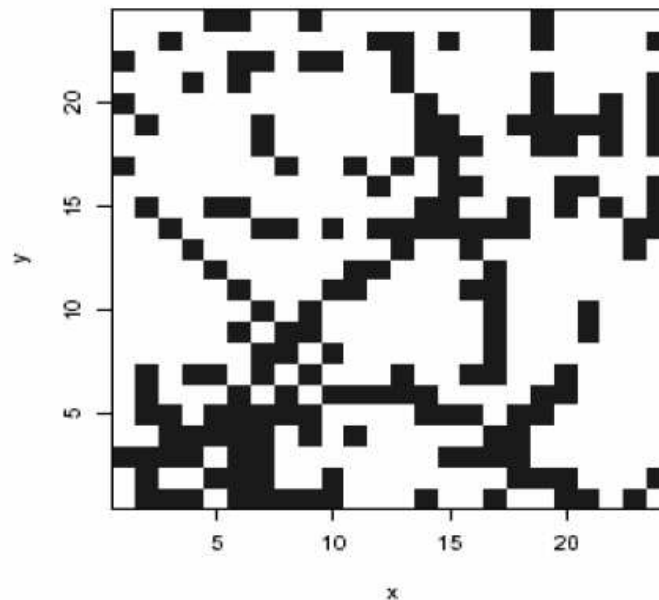
# Modèle de Gibbs - Markov

Ex : répartition spatiale d'une espèce végétale

présence / absence de la *grande laîche*

Modèle de Auto - Logistique  $\{0, 1\}$

Voisinage de dépendance? Estimation, validation, tests ? Simulation



(a) Présence (■) ou absence (□) de grande laîche.

# Données Ponctuelles

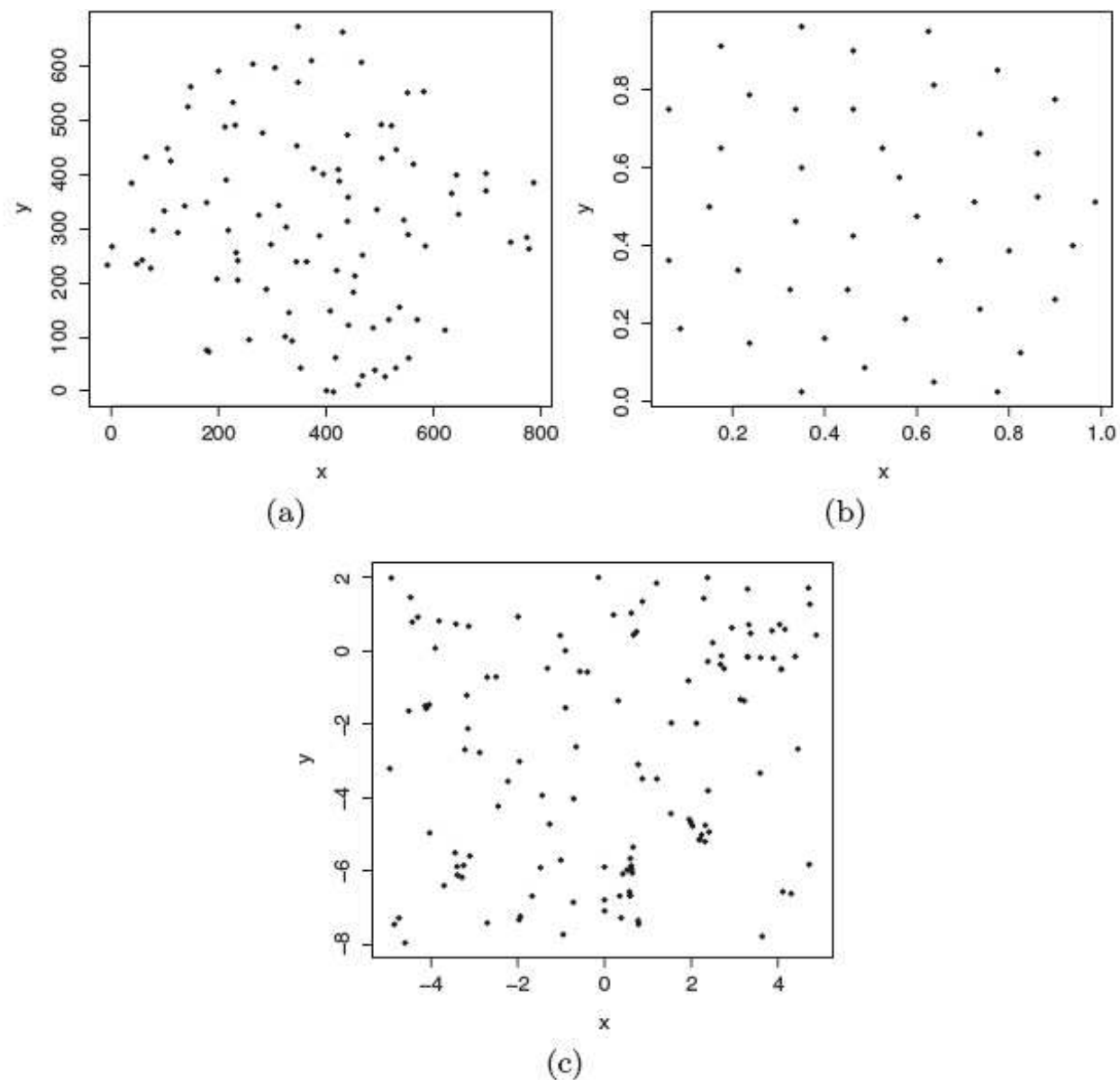
$\mathbf{x}$  = configuration spatiale de  $n$  points

3 exemples

- (a) – 97 fourmilières : données `ants` de `spatstat`
- (b) – 42 centres de cellules d'une coupe histologique (`cells`)
- (c) -- 126 pins d'une forêt finlandaise (`finpines`)

Package `spatstat`





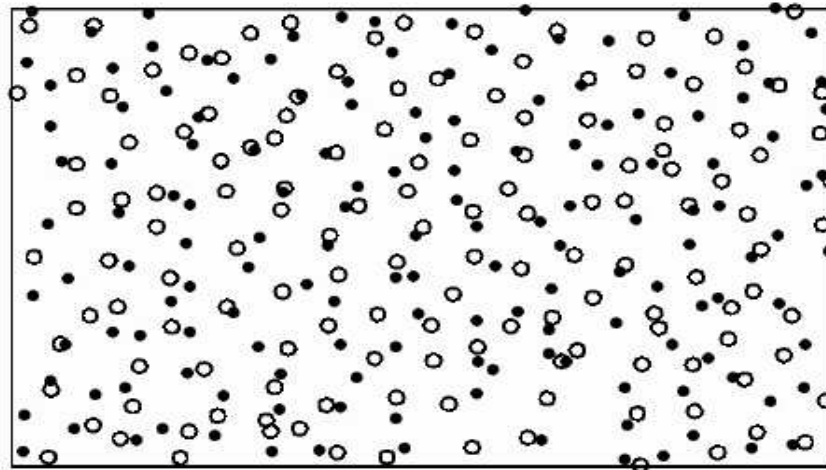
**Fig. 3.1.** Exemples de répartition ponctuelle : (a) 97 fourmilières (données `ants` du `package spatstat`); (b) 42 centres de cellules d'une section histologique observée au microscope (données `cells` de `spatstat`); (c) 126 pins d'une forêt finlandaise (données `finpines` de `spatstat`).

# PP bivarié : 2 types de cellules de la rétine du lapin (294 en tout)

> `data(betacells)` et > `plot(betacells)`

« on » (°) active à l'ouverture de la lumière, « off » (.) à la fermeture.

*Question* : ces cellules sont-elles sur une même couche ?



# Questions sur les Processus ponctuels

- Répartition spatiale *au hasard* (P.P.de Poisson = CSR pour Complete Spatial Randomness))
- Ou non :
  - *avec compétition* (chaque centre de cellule développe une zone d'influence)
  - *avec coopération* (i.e. agrégats autour d'un père)
- *Homogénéité* spatiale ou non
- Quels *modèles explicatifs* ?
- Statistique

---

# Géostatistique

## Modèles, prédiction, estimation

---

Terminologie proposée par Matheron (1962, École des mines de Fontainebleau)

Initialement pour *l'évaluation des réserves minières*.

*Aujourd'hui utilisée dans des domaines variés* : environnement, épidémiologie, science de la terre, ....

Wackernagel (1995), Chiles et Delfiner (1999), Diggle et Ribeiro (2006).

# Objectifs de la géostatistique

## Modélisation

- Modèle au second ordre, covariance, stationnarité
- Modèle intrinsèque : accroissements stationnaires
- Régularité : continuité, dérivabilité
- Prédiction à covariance connue : le Krigeage (simple, universel)

## Statistique

- Nuée variographique
- Variogramme empirique
- Estimation d'un modèle paramétrique
- Validation de modèle : validation croisée, bootstrap paramétrique

# Champ du second ordre $X$ sur $S$ ( $L^{**2}$ )

- Domaine d'étude : sites  $s$  de  $S$ , sous ensemble de  $R^{**2}$
- Observation  $X(s)$  *réelle* et de variance finie :  $Var(X(s)) < \infty$
- $X$  caractérisé par ses lois finies dimensionnelles  
Moyenne :  $m(s) = E(X(s))$   
Covariance :  $c(s,t) = cov(X(s), X(t))$
- Le plus souvent, modèle gaussien (pas une nécessité)

# Différents Bruits Blancs (BB)

- **BB fort** : variables  $\{e(s)\}$  i.i.d.
- **BB faible** : variables centrées et de même variances
- **BB gaussien** : BB faible gaussien
- **BB coloré** : variables centrées même variances mais corrélées

# Caractérisation d'une covariance : la semi définie positivité (sdp)

$$\forall a \in \mathbb{R}^m \text{ et } \forall (s_1, s_2, \dots, s_m) \in S^m : \sum_{i=1}^m \sum_{j=1}^m a_i a_j c(s_i, s_j) \geq 0$$

$$\text{Var} \left( \sum_{i=1}^m a_i X_{s_i} \right) = \sum_{i=1}^m \sum_{j=1}^m a_i a_j c(s_i, s_j) \geq 0.$$



# Champ gaussien **X**

Si toute combinaison linéaire est gaussienne

**X** spécifié par sa moyenne **m(.)** et sa covariance **c(.,.)**

$a = (a_s, s \in \Lambda)$ ,  $\sum_{s \in \Lambda} a_s X_s$  est une variable gaussienne.

$$f_{\Lambda}(x_{\Lambda}) = (2\pi)^{-\#\Lambda/2} (\det \Sigma_{\Lambda})^{-1/2} \exp \left\{ -1/2 {}^t(x_{\Lambda} - m_{\Lambda}) \Sigma_{\Lambda}^{-1} (x_{\Lambda} - m_{\Lambda}) \right\}$$

# Champ stationnaire

- Moyenne *constante*
- Covariance *invariante par translation* :

$$c(s, t) = Cov(X_s, X_t) = C(t - s)$$

# Champ isotrope

covariance invariante par isotropie :

$$c(s, t) = C_0(\|s - t\|) = C(s - t).$$

# Propriétés d'une covariance stationnaire $C$

- $C$  est *semi-définie positive*
- $|C(h)| \leq C(0)$
- $X(As)$  est stationnaire si  $s \rightarrow As$  est linéaire
- Une somme pondérée à coefficients  $>0$  de covariances est encore une covariance
- Si  $C$  est continue en  $0$ , alors  $C$  est uniformément continue partout

# Quelques covariances isotropiques

Portée  $a > 0$  et Variance  $\sigma^{**2} > 0$

- **Pépitique**:  $C(0) = \sigma^{**2}$  et  $C(h) = 0$  sinon
- **Exponentielle** :  $C(h) = \sigma^{**2} \exp(-a \|h\|)$
- **Sphérique** si  $d \leq 3$

$$C(h) = \sigma^2 \left\{ 1.5 \|h\|/a - 0.5 (\|h\|/a)^3 \right\} \text{ si } \|h\| \leq a$$

$C(h) = 0$  sinon

- **Gaussienne** :  $C(h) = \sigma^2 \exp(-(\|h\|/a)^2)$

(cf. liste assez complète dans `> cov.spatial`)

# Modèle(s) de Matern

- Plus un paramètre  $\nu$  contrôle la régularité de  $C$  en  $0$  ( $K$  est la fonction de Bessel de première espèce)
- $\nu = 1/2 \rightarrow$  cov. exponentielle
- $\nu = \infty \rightarrow$  cov. Gaussienne
- Plus  $\nu$  augmente, plus  $C(h)$  est régulière en  $0$  et plus  $X$  est régulier (en moyenne quadratique)

$$C(h) = \sigma^2 2^{1-\nu} (\|h\| / a)^\nu \mathcal{K}_\nu(\|h\| / a) / \Gamma(\nu)$$

# Champ intrinsèque et variogramme

- Considérer le champ des *h*-accroissements :

$$X_{s+h} - X_s : s \in S$$

- *X* intrinsèque si ses *h*-accroissements sont stationnaires

- Variogramme en *h* :

$$2\gamma(h) = \text{Var}(X_{s+h} - X_s)$$

# Stationnaire ou intrinsèque ?

- Stationnaire  $\rightarrow$  intrinsèque :  $2\gamma(h) = 2(C(0) - C(h))$

- Intrinsèque  $\not\Rightarrow$  stationnaire :

*Exemple* : le mouvement brownien,  $\gamma(h) = |h|$

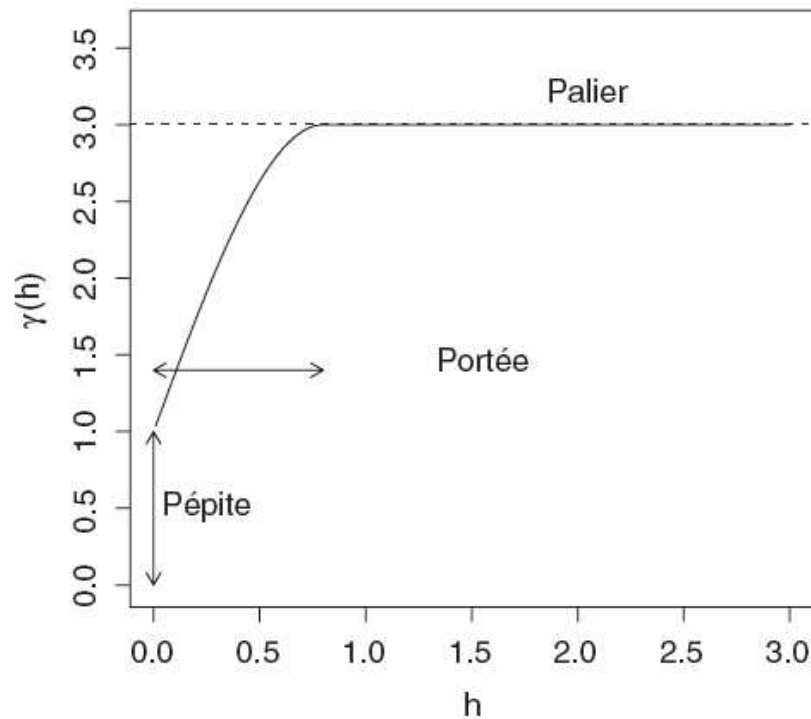
- Un variogramme n'est pas toujours borné :

$$\gamma(h; b, c) = b\|h\|^c, \quad 0 < c \leq 2.$$

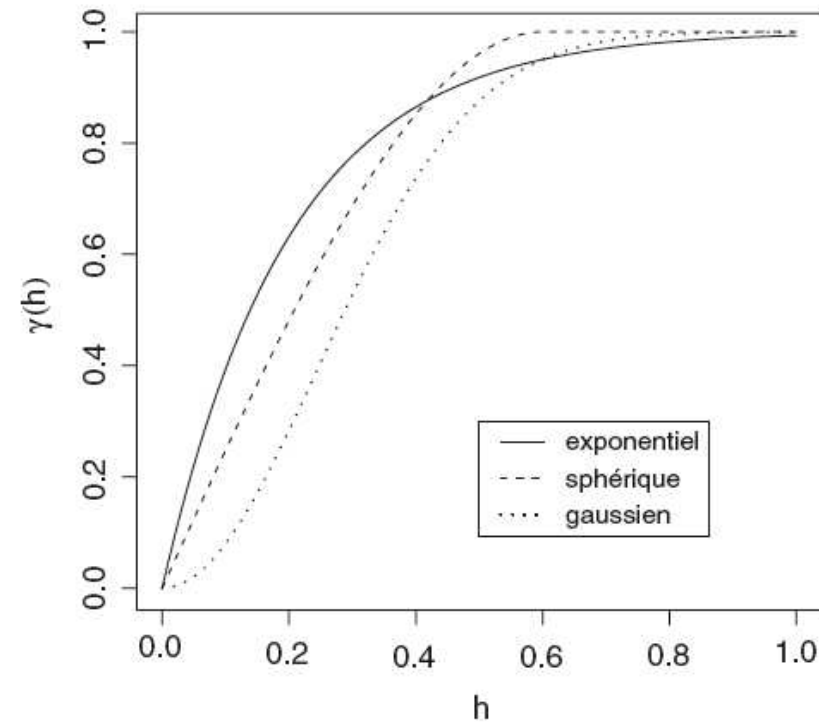
*Exemple* : vario puissance et auto-similarité ( $c = 1$  pour le mouvement brownien).

# Portée, palier, effet pépité d'un variogramme

- (a) Les 3 caractéristiques d'un variogramme
- (b) Variog. expo.(1), sphérique (2) et gaussien (3) : *régularité* en 0 *linéaire* pour (1-2) et *parabolique* pour (3)



(a)

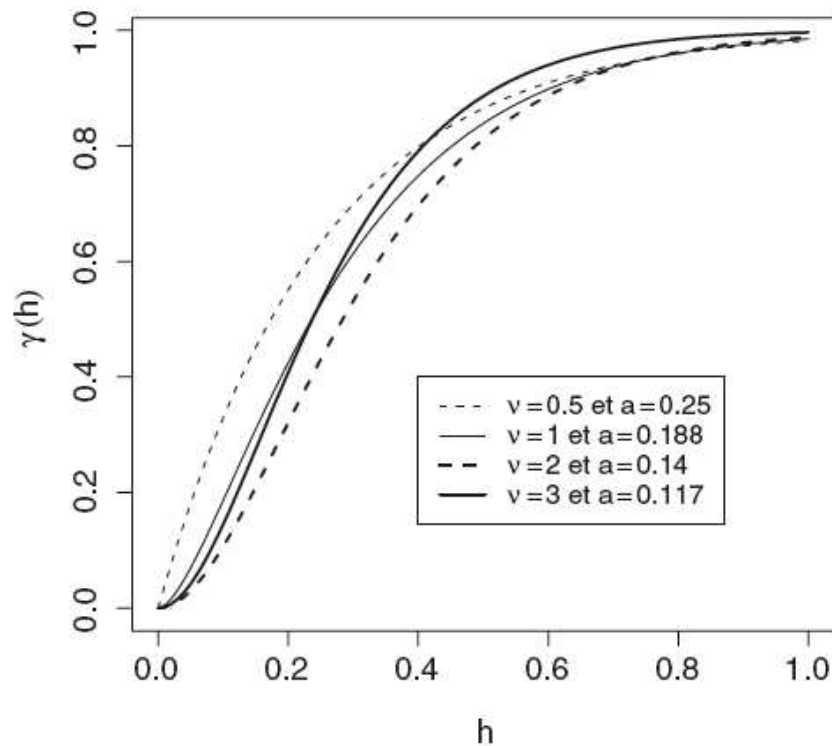


(b)



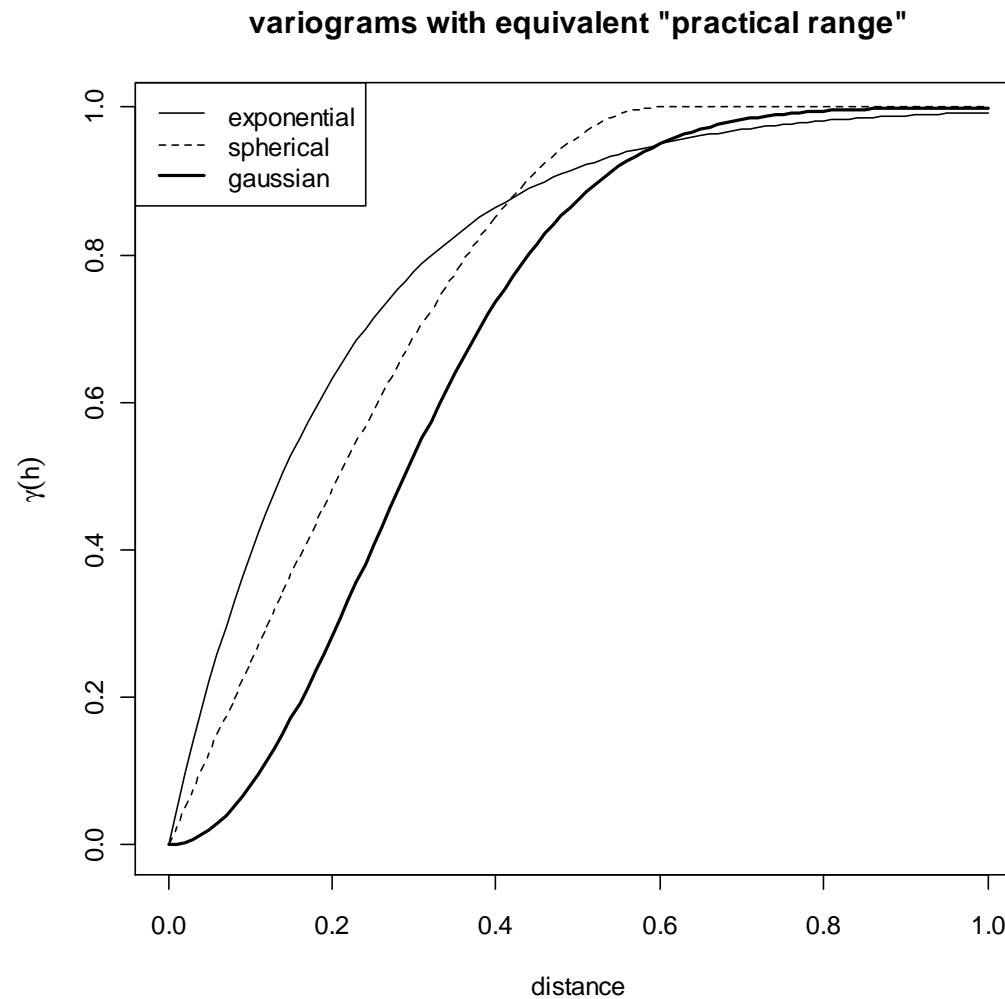
# Régularité du variogramme de Matern

- $\nu = 1/2$  donne le variogramme exponentiel
- $\nu \uparrow$ , plus de régularité en  $0$
- $\nu \geq 2$ ,  $C$  dérivable en  $0$  à dérivée nulle



`cov.spatial` de R → principales covariances spatiales

**Exemple** : expo., sphérique, gauss de même portée pratique.



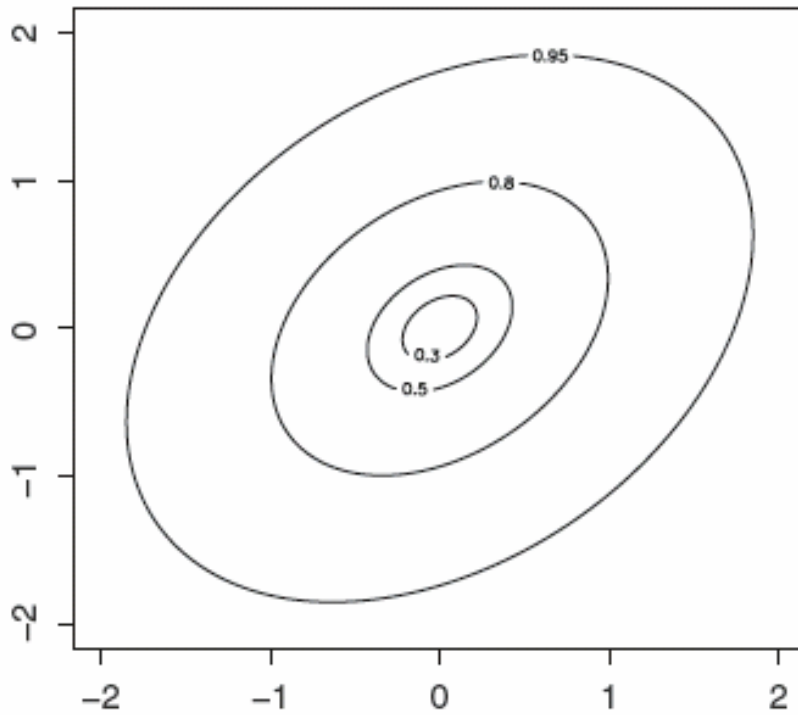
# Anisotropies

Variogrammes différents suivant les directions

- Anisotropie géométrique :  $\gamma(h) = \gamma_0(\|Ah\|)$
- Anisotropie zonale:  $\gamma(h) = \gamma_1(\sqrt{h_1^2 + h_2^2}) + \gamma_2(|h_2|)$

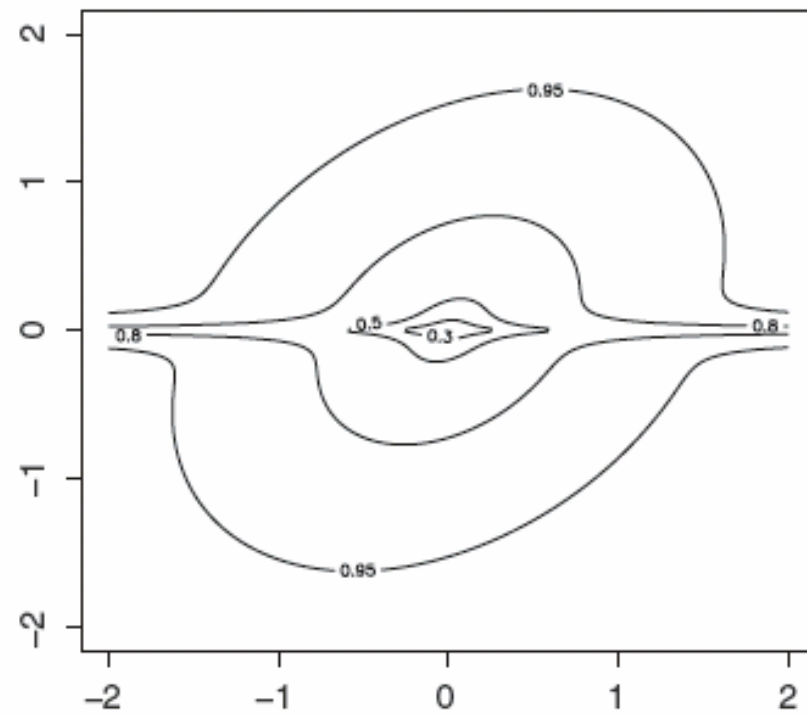
# Anisotropies

(a) : géométrique



(a)

(b) zonale



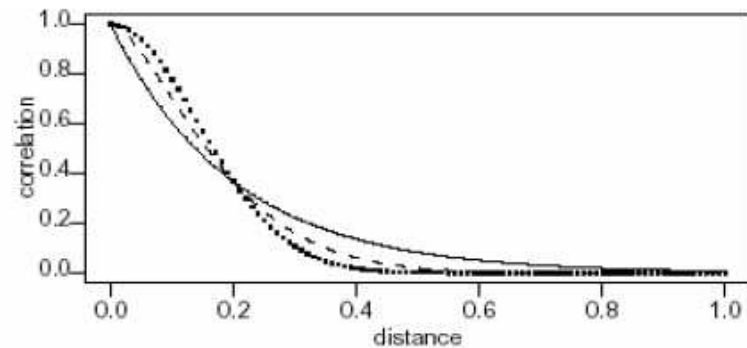
(b)

# La régularité de la covariance $C$ en $0$ règle la régularité en m.q. de $X$ partout

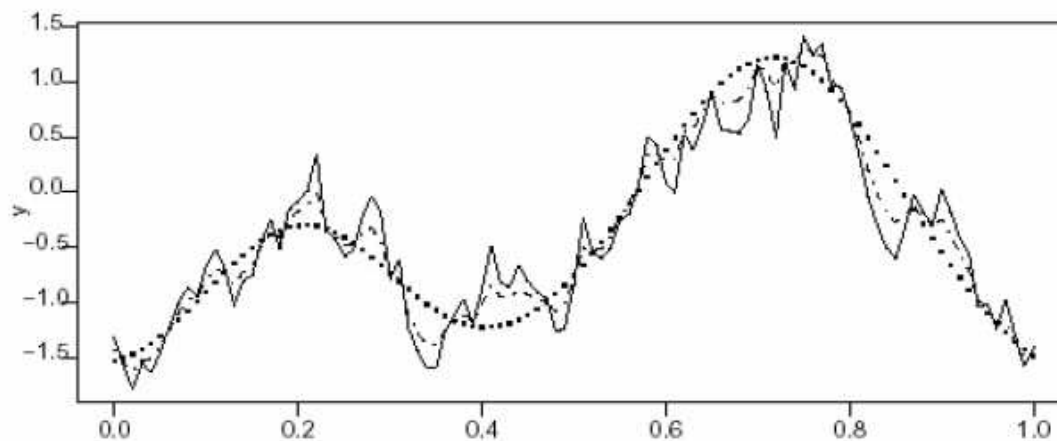
- $C$  continue en  $0 \rightarrow X$  continu partout
- $C''$  existe en  $0 \rightarrow X$  dérivable partout
- Idem en remplaçant  $C$  par le variogramme  $\gamma$
- Importance : régularité d'une carte de prédiction (Krigage) est fonction du choix de  $C$  (ou  $\gamma$ )

**Covariance « exponentielle de puissance »**  $C(h) = \exp(-(h/\phi)^{\kappa})$ .

$C(h)$  et simulations sur un intervalle pour  $\phi=0.2$  et  $\kappa=1$  (solide), 1.5 (tirets) et 2 (pointillés). **La régularité en 0 de  $h \rightarrow C(h)$  augmente avec  $\phi$ .**



**Figure 3.3.** Three examples of the powered exponential correlation function with  $\phi = 0.2$  and  $\kappa = 1$  (solid line),  $\kappa = 1.5$  (dashed line) and  $\kappa = 2$  (dotted line).



# Simulation d'un champ gaussien

le package `RandomFields`

`GaussRF` : simule un champ spatial ou spatio-temporel stationnaire

Il faut déclarer :

- la fonction de covariance

- la grille de simulation

- la tendance si il y en a une

- la méthode de simulation retenue

(cf. 1<sup>er</sup> exemple de champ stable pour différentes grilles)

`CovarianceFct` : donne liste des covariances/variogramme spatiaux ou spatio-temporel

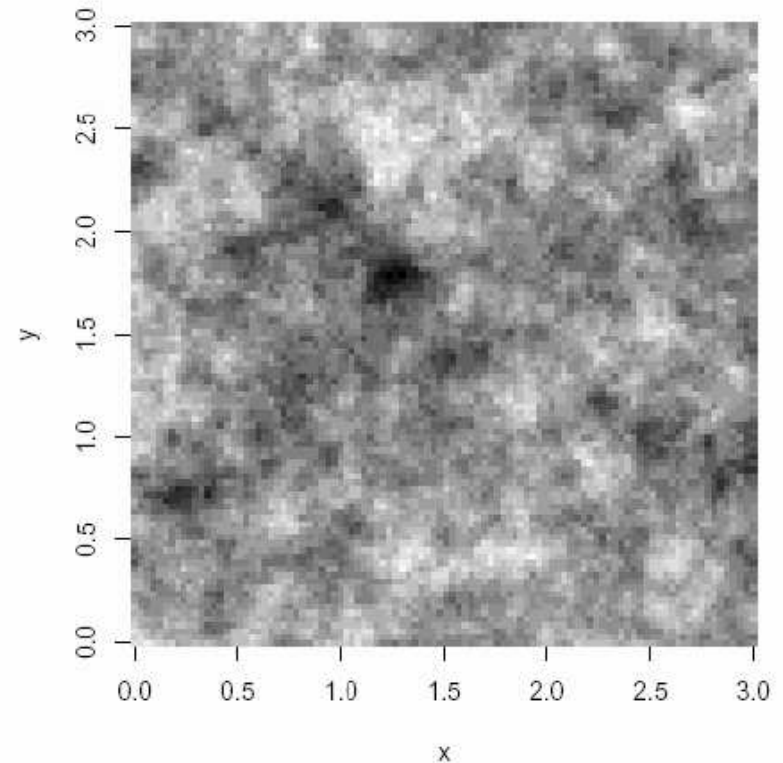
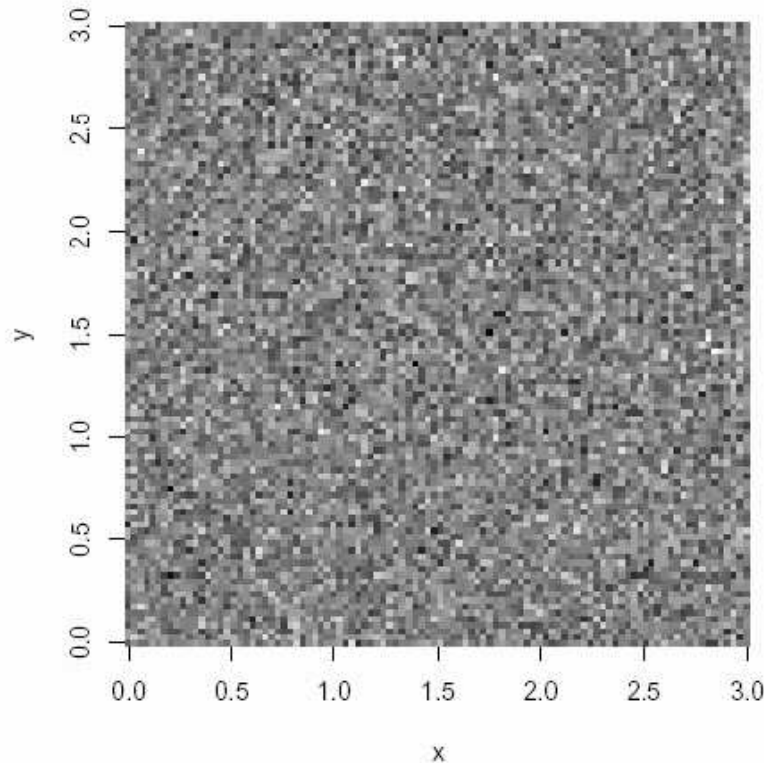
`CondSimu` : réalise la simulation conditionnelle d'un champ gaussien en dehors des sites d'observation

`ShowModels` : démonstration interactive de simulation de modèles

## Réalisations d'un champ gaussien isotropique

à gauche pépitique (discontinu);

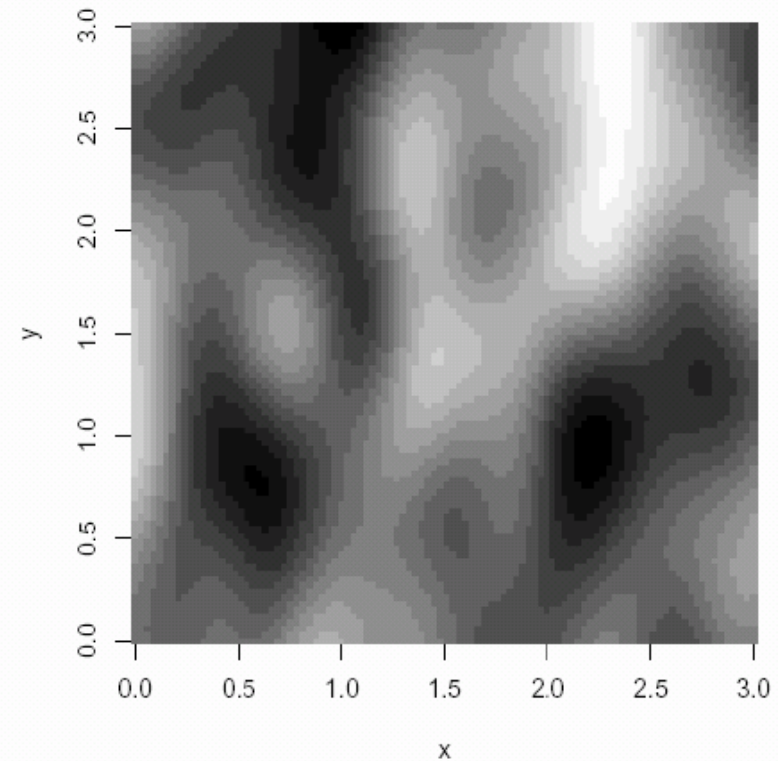
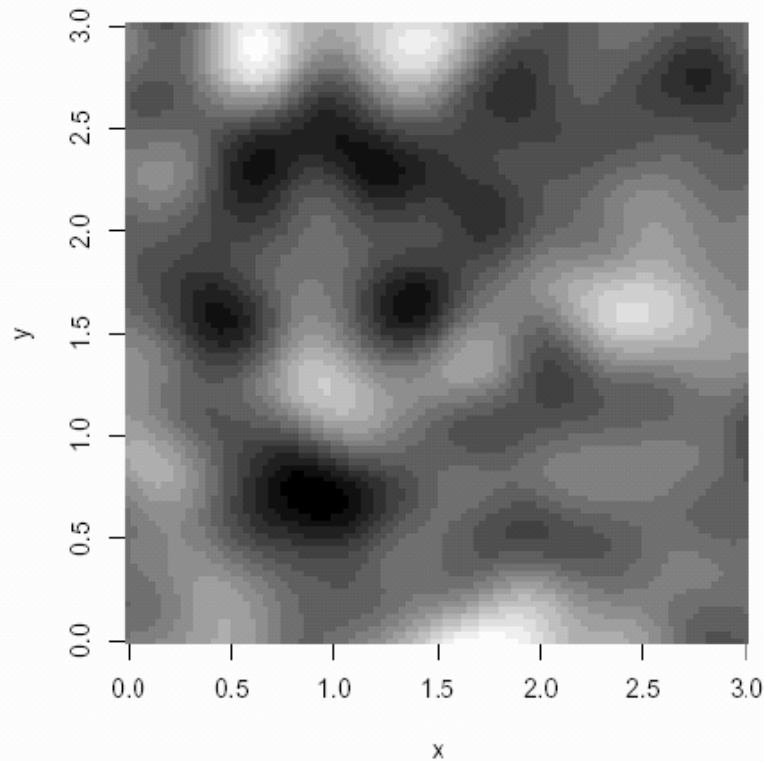
à droite exponentiel (continu mais non dérivable).





(suite) Champ gaussien isotropique à **variogramme gaussien**  
à **gauche**, isotropique; à **droite**, anisotropique.

La surface  $s \rightarrow X(s)$  est dérivable.



# Prédiction à moyenne et covariance connue : le Krigeage simple

- Vecteur des  $n$  observations  $X = (X(s(1)), X(s(2)), \dots, X(s(n)))$
- Prédire  $X(s(0))$  (*carte de krigeage*) partout sur  $S$
- Choix d'une covariance connue (variogramme)  $C$  :  
 $\Sigma = \text{cov}(X)$  et  $c = \text{cov}(X(s(0)), X)$
- Reconstruction par MCO : minimiser

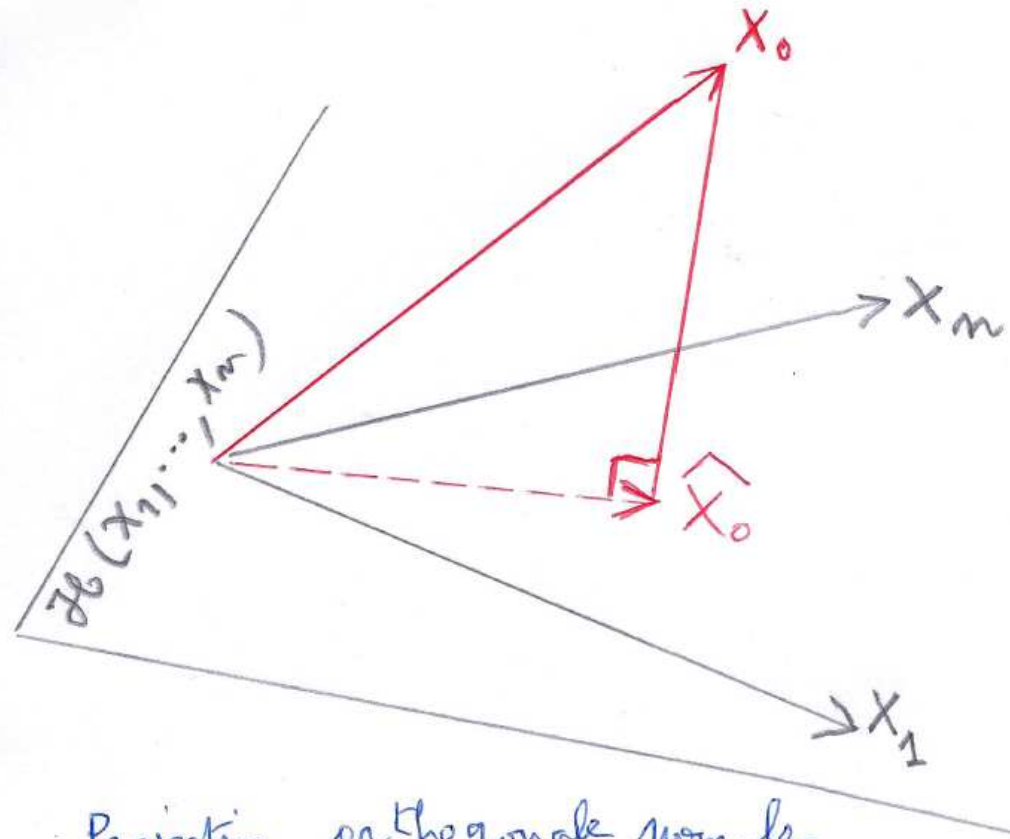
$$EQM(s_0) = E\{(X_0 - \hat{X}_0)^2\}.$$

- si  $X$  de moyenne  $m$  connue, **krigeage simple** en  $s(0)$

*Prévision BLUP* et *Variance de prédiction* :

$$\hat{X}_0 = {}^t c \Sigma^{-1} X, \quad \tau^2(s_0) = \sigma_0^2 - {}^t c \Sigma^{-1} c.$$

# Krigeage = Prédiction = Projection orthogonale pour le produit scalaire de la covariance



Projection orthogonale pour le  
produit scalaire de la covariance

# Le Krigeage universel : $m$ inconnue

- $X$  suit un *modèle de régression* : covariables  $Z$ , paramètre  $\delta$  inconnu,  $\varepsilon$  résidu de covariance  $\Sigma$

$$X = Z\delta + \varepsilon$$

- *Le krigeage* :
  1. Estimer  $\delta$  par *MCG*
  2. Krigeage simple sur résidu  $\varepsilon$

$$\hat{X}_0 = {}^t z_0 \hat{\delta} + c \Sigma^{-1} (X - Z \hat{\delta}), \quad \text{avec}$$

$$\hat{\delta} = ({}^t Z \Sigma^{-1} Z)^{-1} {}^t Z \Sigma^{-1} X.$$

## Krigeage / prédiction avec `geoR`

- `krige.conv`

Effectue la prédiction partout (en fait sur une grille à définir) pour un modèle de variogramme donné (ou un modèle estimé par `variofit` ou par `likfit`)

Option pour le krigeage simple, ordinaire ou universel.

- `Output.control`

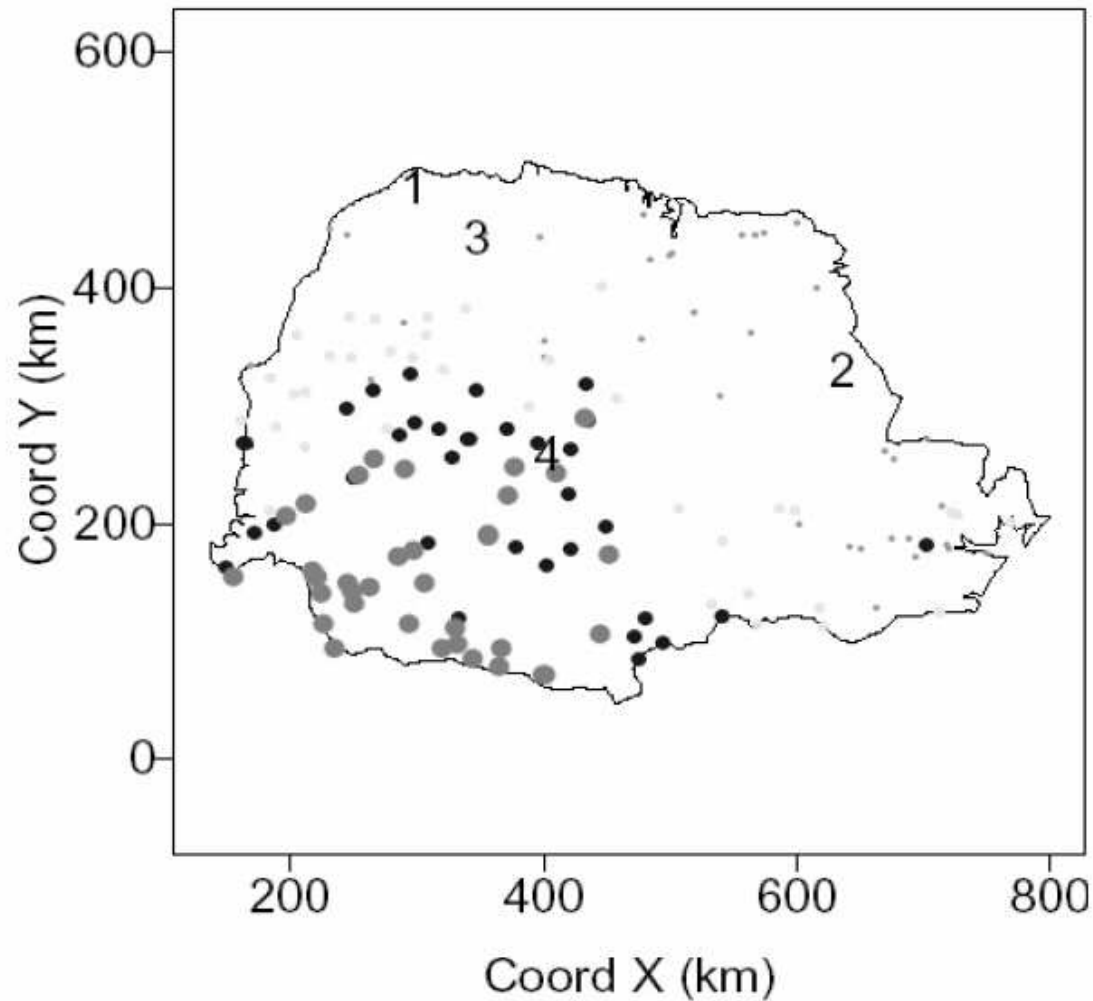
Réalise des simulations conditionnelles; permet d'évaluer la carte des probabilités de dépasser un seuil.

- Voir également les packages

`Fields` (`Krig` et `sim.Krig.grid`) et

`RandomFields` (`Kriging` et `CondSim`)

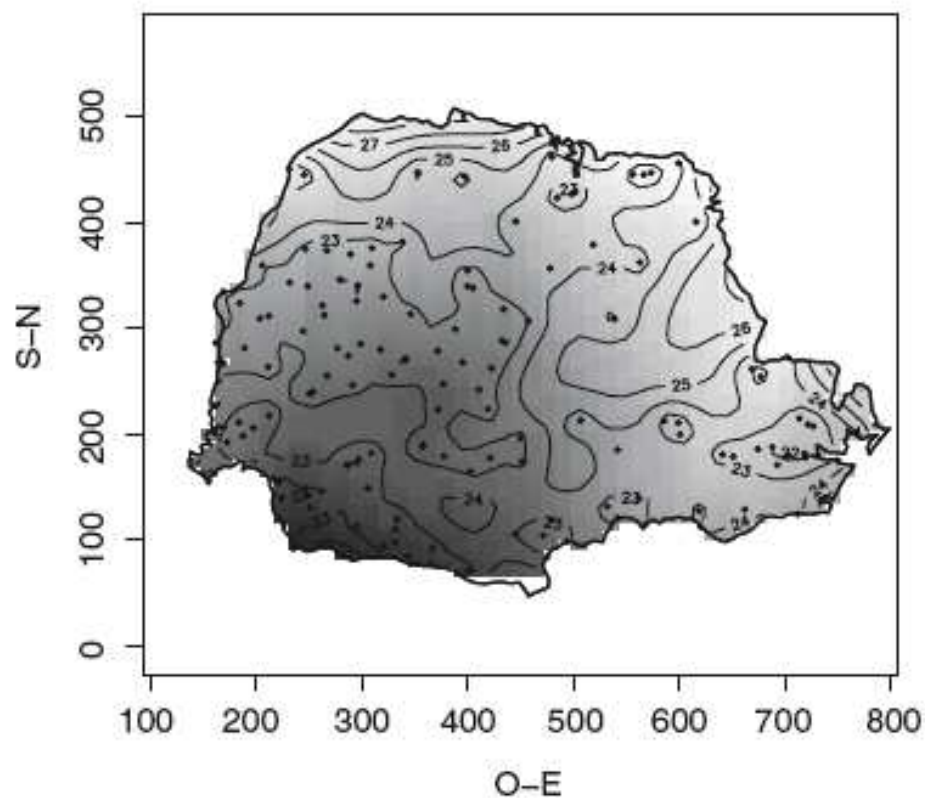
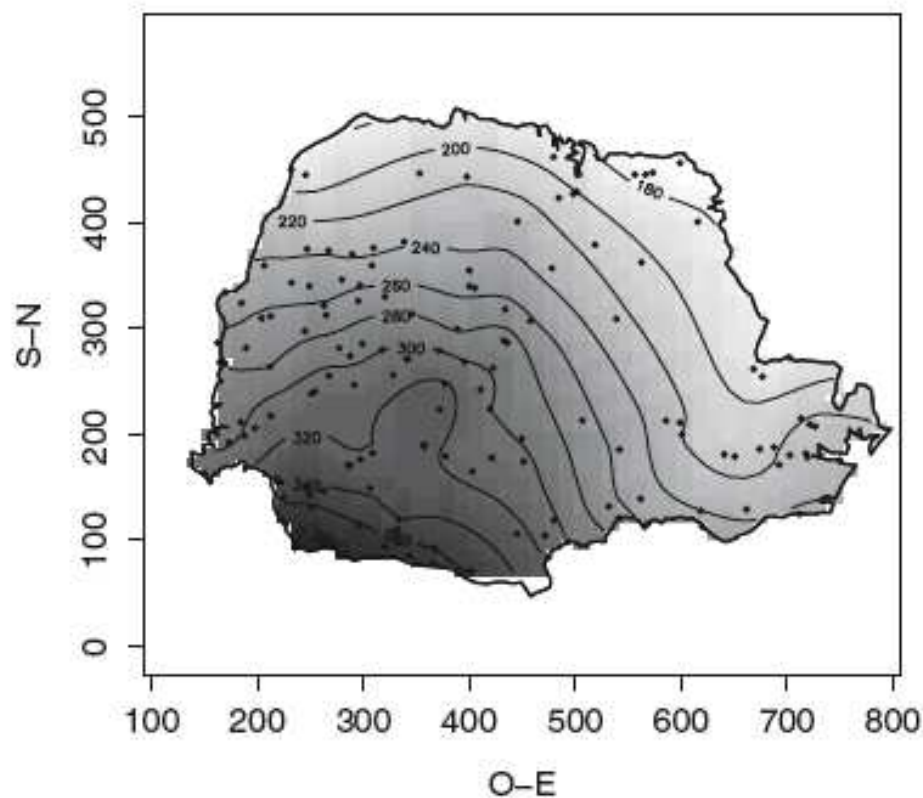
**Données de pluies Parana** : codage en (taille/gris) associé aux quartiles empiriques de la distribution observée.



# Krigeage des pluies Parana

Modèle : régression affine et résidu Gauss + Pépité

- à gauche carte des *hauteurs* de pluie (données en ●)
- à droite carte des *écarts types* des prédictions



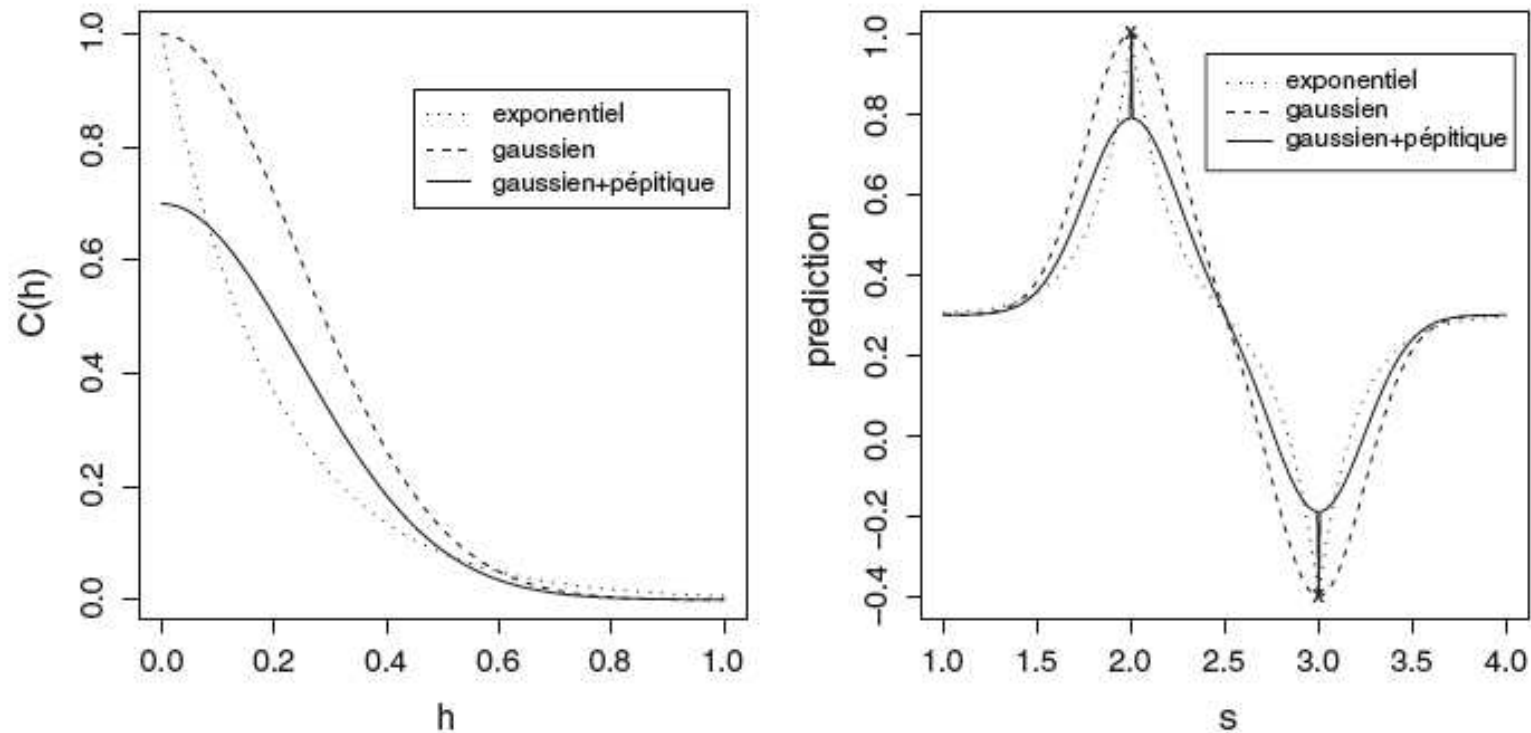
# Propriété de la surface de Krigeage

1. ***Interpolateur universel*** : krigeage  $\equiv$  observation en un point d'observation
2. ***Régularité de la surface de krigeage*** fonction de la ***régularité du variogramme*** (covariance) en 0 :
  - ***Pépitique*** : krigeage constant partout  $\equiv$  la moyenne arithmétique des observations, discontinuité aux points d'observation.
  - ***Linéaire pointu en 0*** : surface continue mais non dérivable aux points d'observations.
  - ***Parabolique en 0*** : continue et dérivable partout.



# Régularité du Krigeage et régularité à l'origine de la covariance spatiale

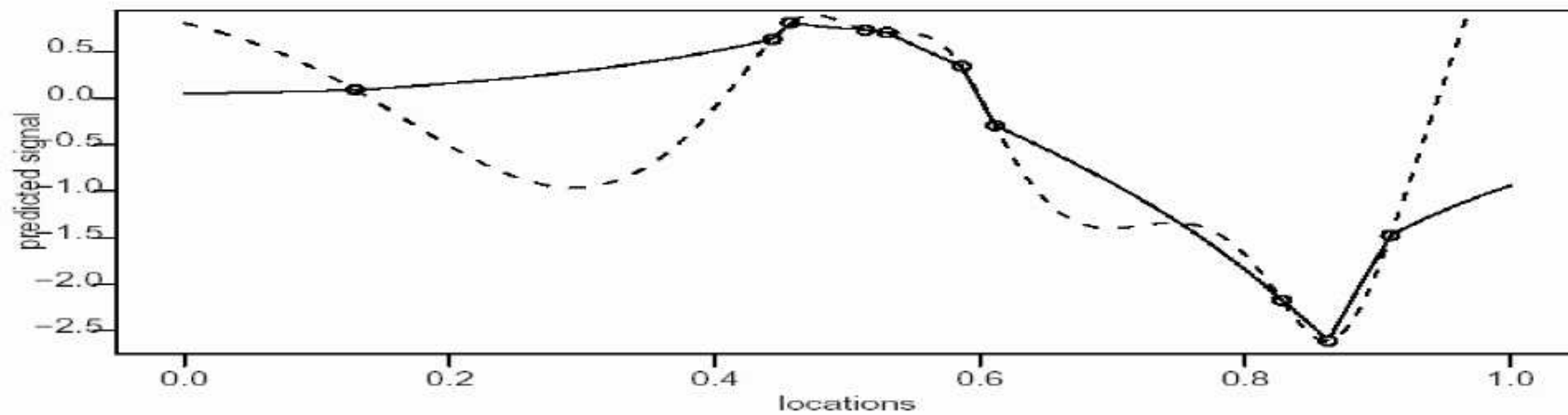
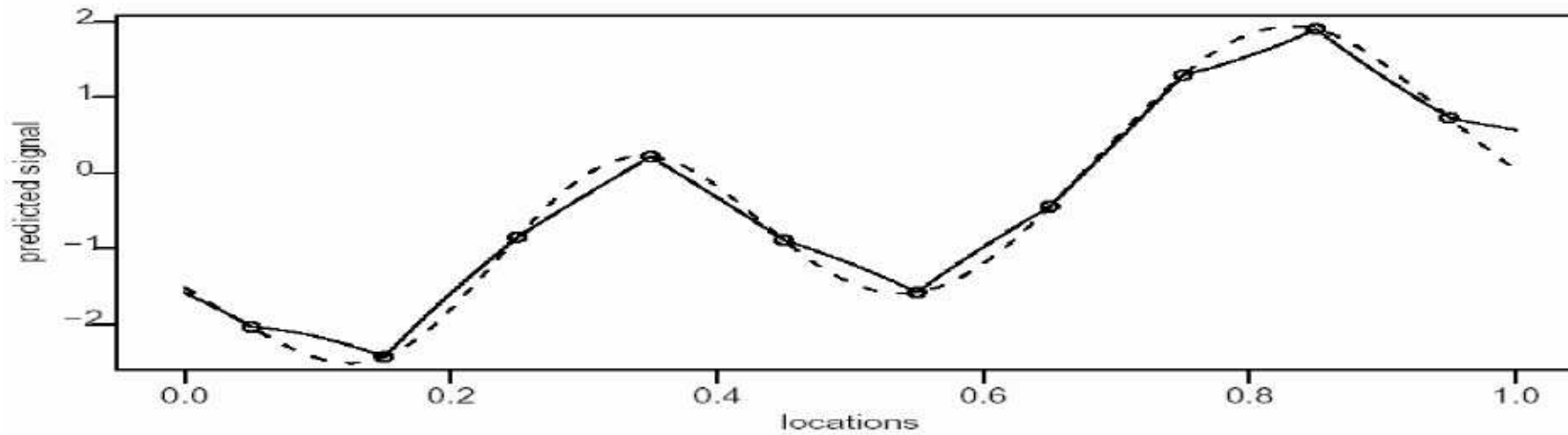
Exemple d'école : 2 observations en (x) et 3 reconstructions



## Prédictions sur un intervalle à partir de 10 points

Deux variogrammes : exponentiel (continu) et Matérn (pointillé);

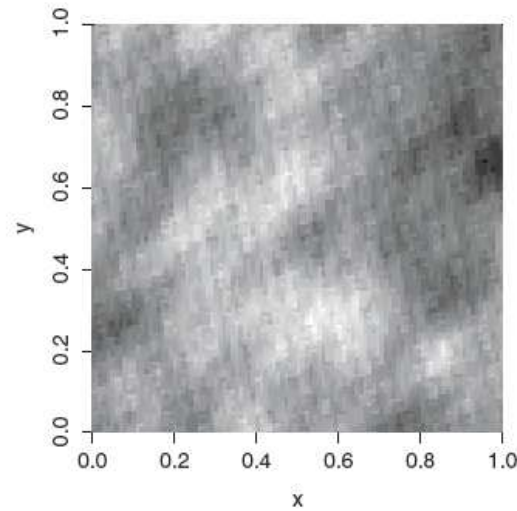
Deux dispositifs : en haut, points régulièrement espacés; bas, placés au hasard



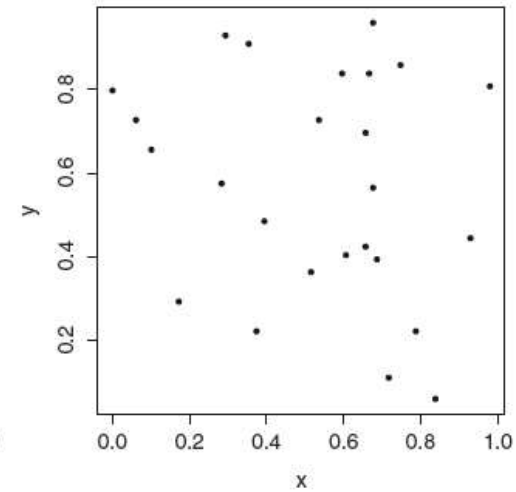
# Krigeage ou simulation conditionnelle

(a)  $X$  simulé, gaussien, cov. exponentielle  
et

(b) 25 points échantillonnés



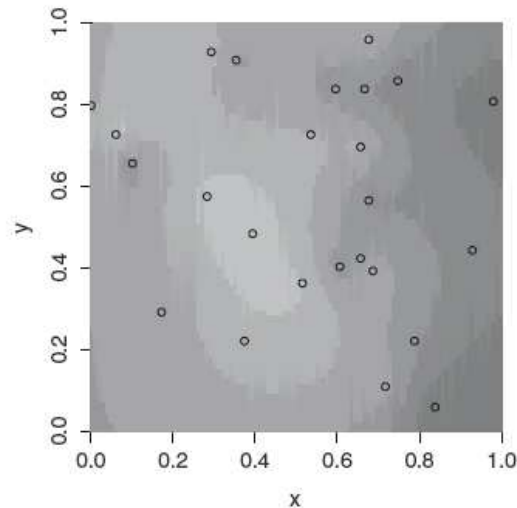
(a)



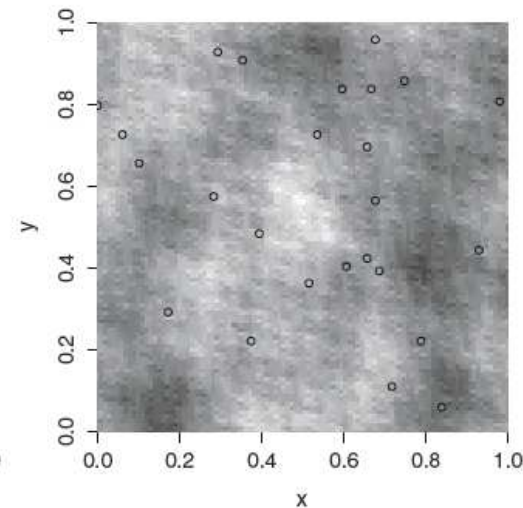
(b)

(c) reconstruction de  $X$  par krigeage

(d) reconstruction de  $X$  par simulation conditionnelle aux 25 observations

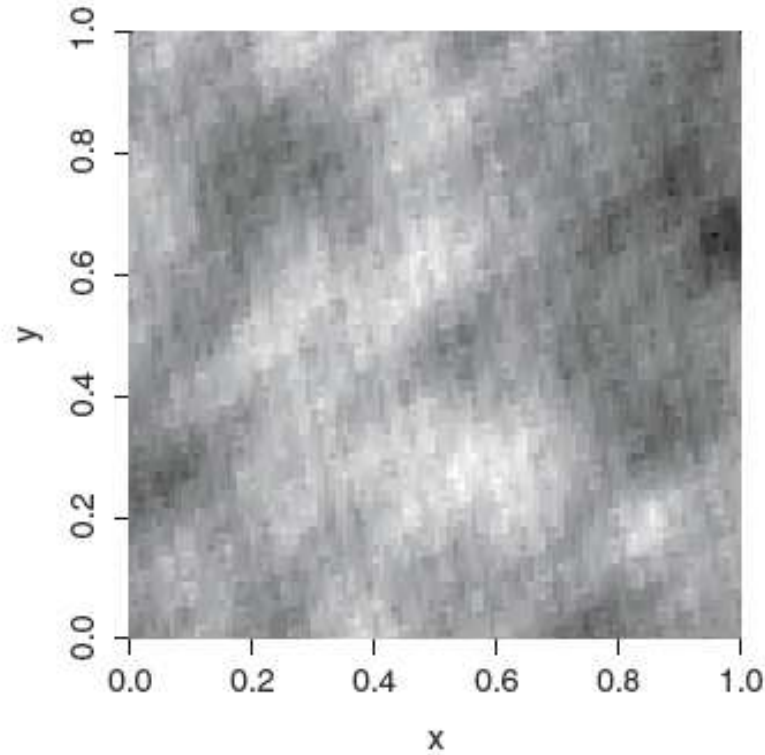


(c)

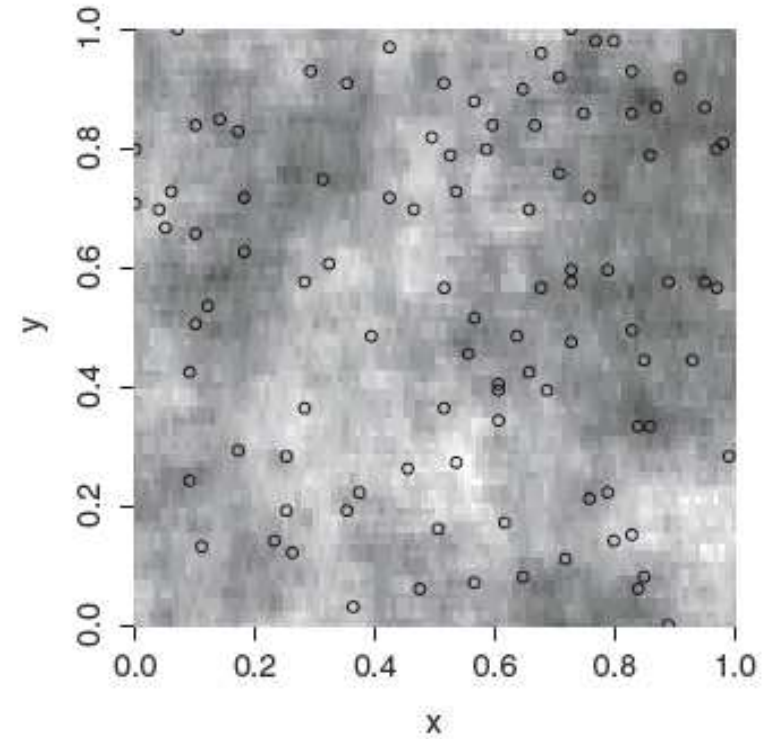


(d)

# Simulation conditionnelle à 100 points



X initial



simulation cond.

---

# Statistique des modèles de géostat

- Deux niveaux : au **premier ordre** (moyenne, tendance) et au **deuxième ordre** (covariance, variogramme).
- Statistique sans modèle : estimations empiriques.
- Statistique avec modèle : MCO, MCP, MCG ou MV.
- Validation et choix de modèle : validation croisée, Bootstrap paramétrique ou méthode de Monte Carlo.

# Nuée variographique (cas isotropique)

Observation de  $X$  en  $n$  points  $\mathcal{O} = \{s_1, s_2, \dots, s_n\}$

Nuée variographique = nuage des  $\frac{n(n-1)}{2}$  points :

$$\mathcal{N} = \{(\|s_i - s_j\|, (X_{s_i} - X_{s_j})^2, i \neq j)\}$$

Comme  $E((X_{s_i} - X_{s_j})^2) = \gamma(s_i - s_j)$ ,

$\implies \mathcal{N}$  "estime" sans biais  $h \mapsto \gamma(h)$

- Plus lissage avec bon noyau de convolution
- si  $X$  non isotropique, nuées dans les 4 directions cardinales  $\{E, NE, N, NW\}$  avec une tolérance angulaire de  $\pm 22.5^\circ$

# Estimation empirique du variogramme

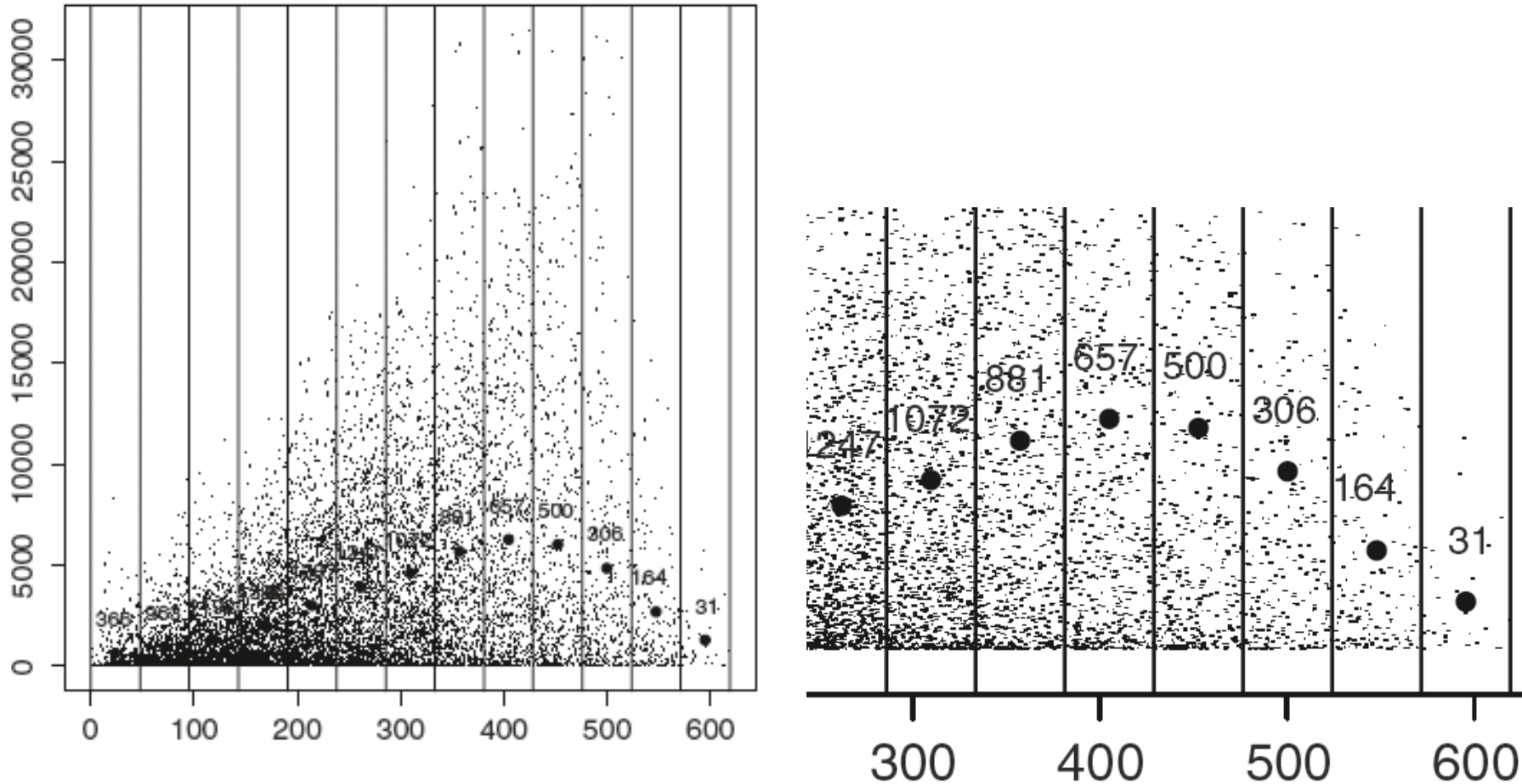
- 1 -  $N(h)$  : ensembles des couples  $r$ -voisins à  $\Delta$ -près (au moins 30 points dans chaque  $N(h)$ )
- 2 - *Avantage* : ne nécessite pas d'estimation préalable de la moyenne (nécessaire pour estimer une covariance)
- 3 - *Version robuste* aux grandes valeurs  $(X(s(i))-X(s(j)))^{**2}$  (Cressie et Hawkins)

$$\hat{\gamma}_n(h) = \frac{1}{\#N(h)} \sum_{(s_i, s_j) \in N(h)} (X_{s_i} - X_{s_j})^2, \quad h \in \mathbb{R}^d.$$

$$N(h) = \{(s_i, s_j) : r - \Delta \leq \|s_i - s_j\| \leq r + \Delta; i, j = 1, \dots, n\}.$$

# Estimation empirique du variogramme :

## Choix de classes de distances et effectifs par classes

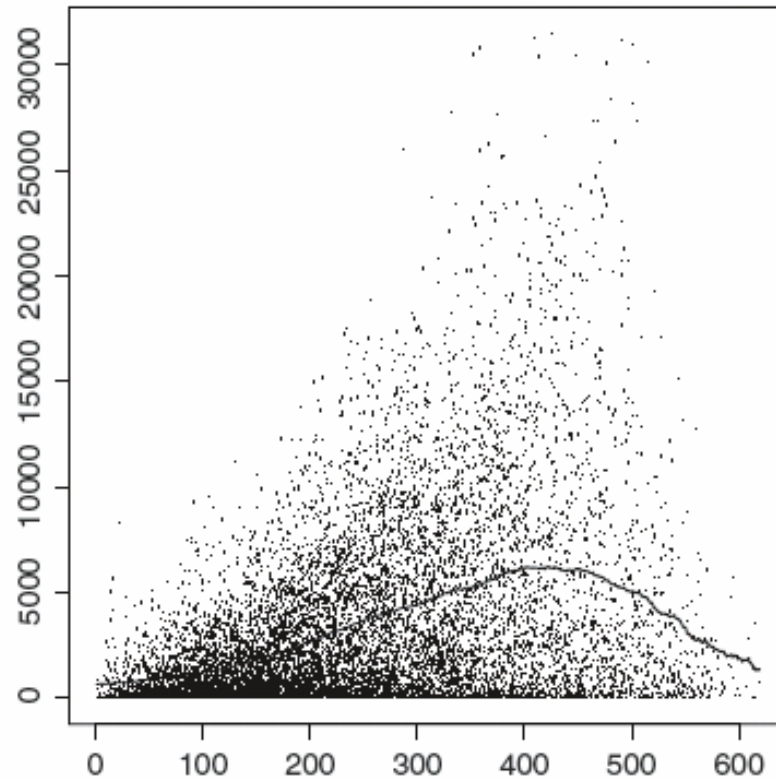




# Nuée variographique et variogrammes lissés

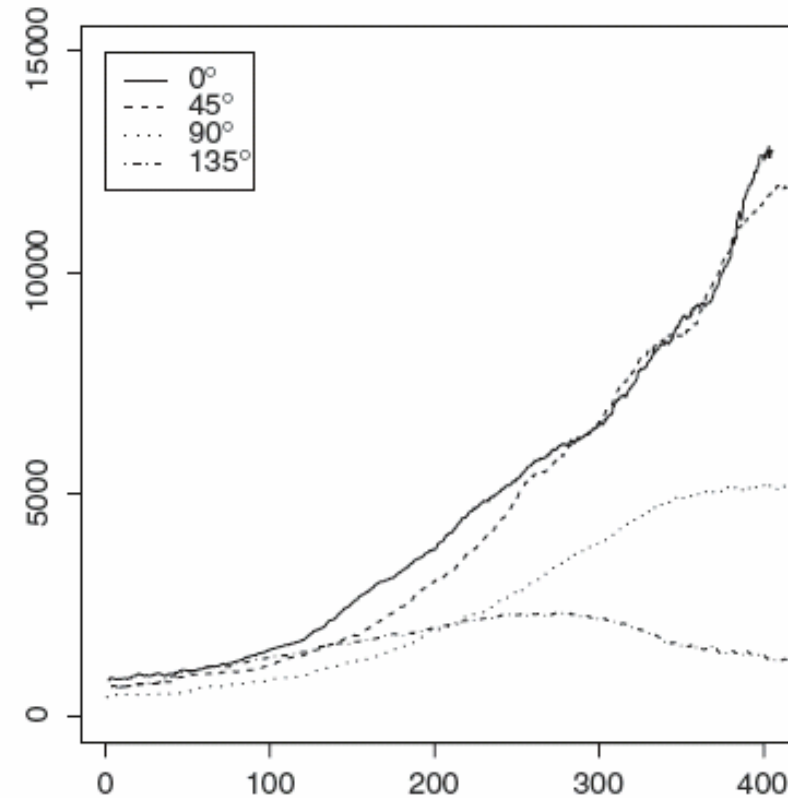
(données de pluies au Parana)

(a) modèle isotropique



(a)

(b) dans 4 directions



(b)

## `variog` : calcul du variogramme empirique

- isotrope ou non (`variog4`),
- Fixation ou non du nombre de classes (bin)
- estimation classique ou robuste

Retourne une estimation par classe (bin), le nuage variographique, le variogramme lissé ...

```
> variog(ca20)
```

```
> print(x)
```

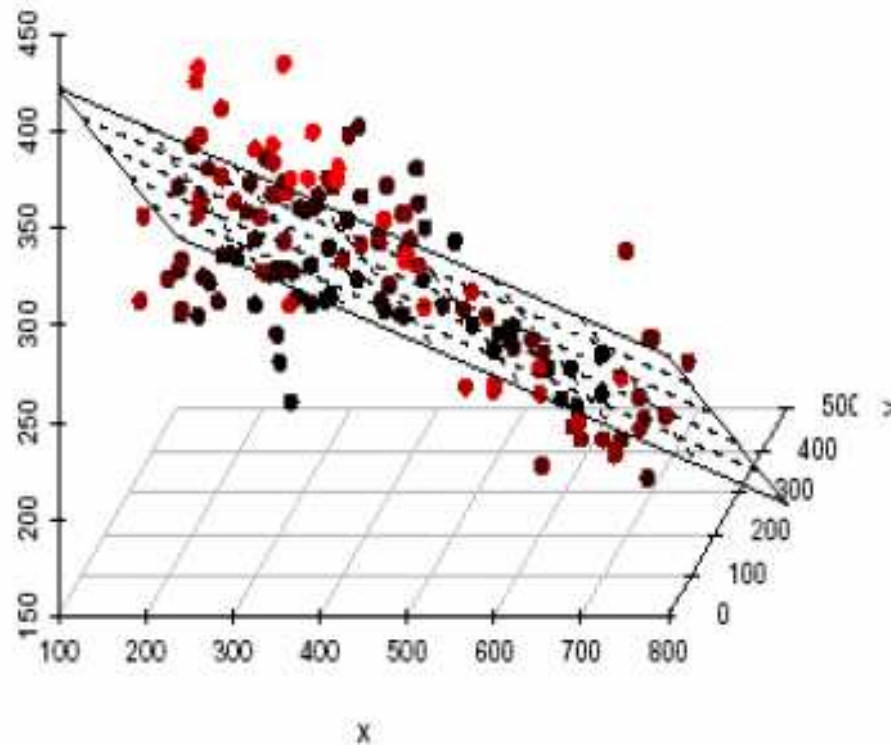
```
> plot(x)
```

# Données Parana : ajustement affine

**Rouges** : au dessus

**Noirs** : au dessous

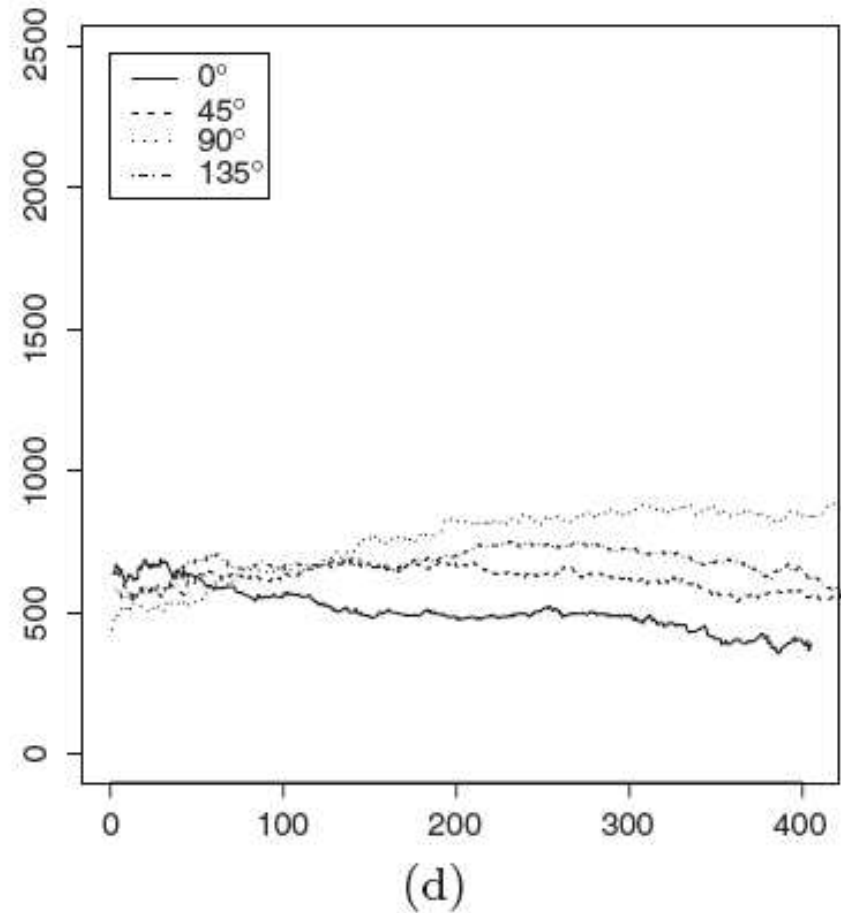
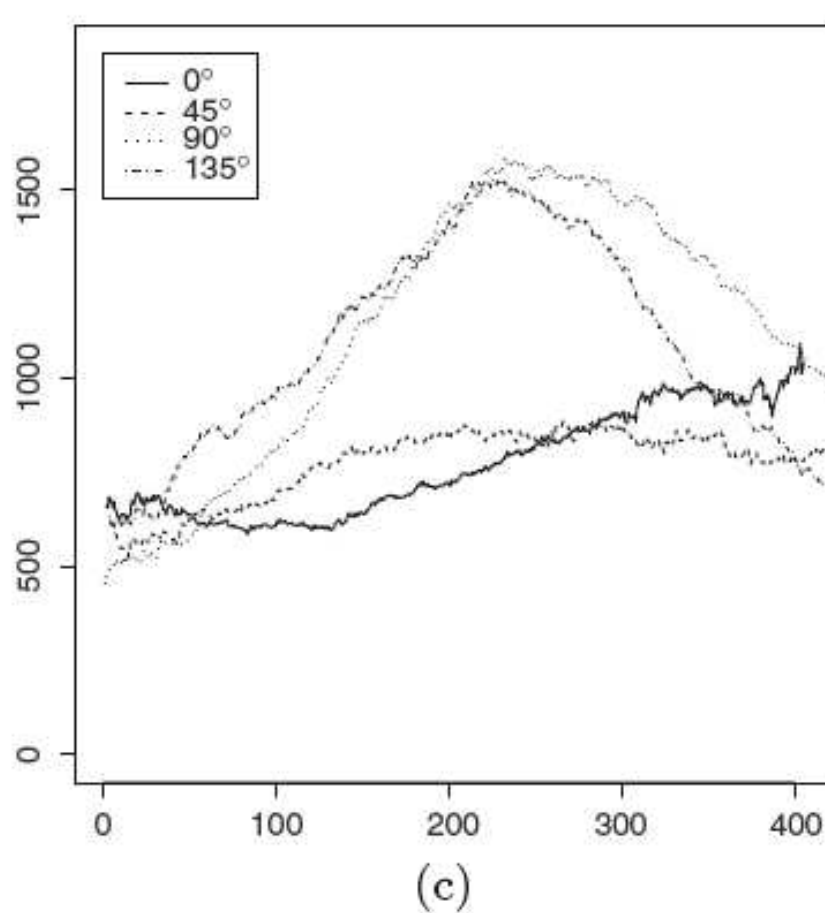
→ calcul des résidus puis variogramme des résidus



# Variogrammes des résidus dans 4 directions pour:

(c) le modèle moyen affine (pluies au Parana);

(d) le modèle moyen quadratique .....



# Estimation d'un modèle paramétrique :

Choisir  $k$  *classes* de distances *rendant identifiable le modèle* de variogramme :

variogramme  $\gamma(\cdot; \theta)$

$k$  classes  $\mathcal{H} = \{h_1, h_2, \dots, h_k\}$

# Estimations d'un modèle de variogramme

- 1 – Moindres carrés ordinaires (MCO)
- 2 – Moindres carrés pondérés (MCP)
- 3 – Moindres carrés généralisés (MCG) et MCQG
- 4 – Maximum de Vraisemblance (MV)

$$\hat{\theta}_{MCO} = \operatorname{argmin}_{\alpha \in \Theta} \sum_{i=1}^k (\hat{\gamma}_n(h_i) - \gamma(h_i; \alpha))^2$$

$$\hat{\theta}_{MCP} = \operatorname{argmin}_{\alpha \in \Theta} \sum_{i=1}^k \frac{\#N(h_i)}{\gamma^2(h_i; \alpha)} (\hat{\gamma}_n(h_i) - \gamma(h_i; \alpha))^2$$

$$\hat{\theta}_{MCG} = \operatorname{argmin}_{\alpha \in \Theta} {}^t(\hat{\gamma}_n - \gamma(\alpha)) \{Cov_{\alpha}(\hat{\gamma}_n)\}^{-1} (\hat{\gamma}_n - \gamma(\alpha))$$

# Estimation du variogramme en présence d'une tendance linéaire

$$X_s = {}^t z_s \delta + \varepsilon_s, \quad \delta \in \mathbb{R}^p$$

résidu champ intrinsèque centré,

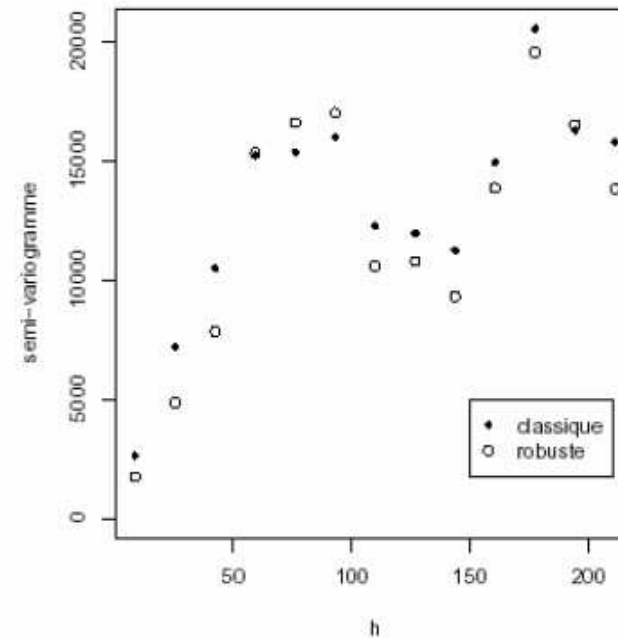
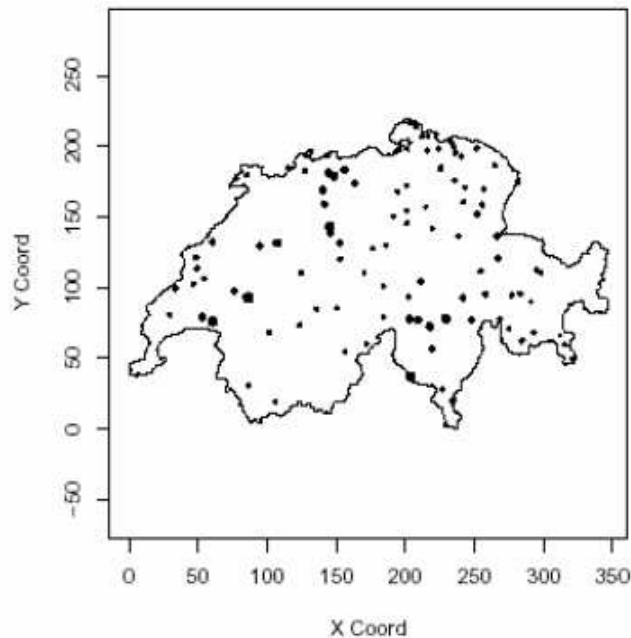
$$E(\varepsilon_{s+h} - \varepsilon_s)^2 = 2\gamma(h, \theta), \quad \theta \in \mathbb{R}^q.$$

1. Estimer par *MCO*
2. En déduire les résidus des *MCO*
3. Variogramme de ces résidus *MCO*

# Estimations modèles de pluies en Suisse

A gauche : données (stations et intensité);

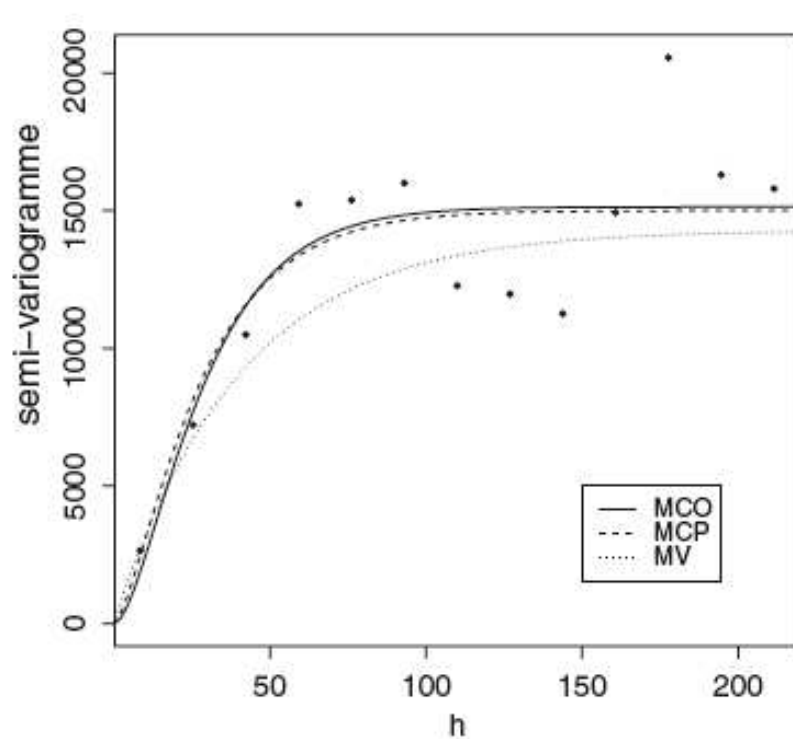
A droite : estimations empiriques classique et robuste sur 13 classe de distances.



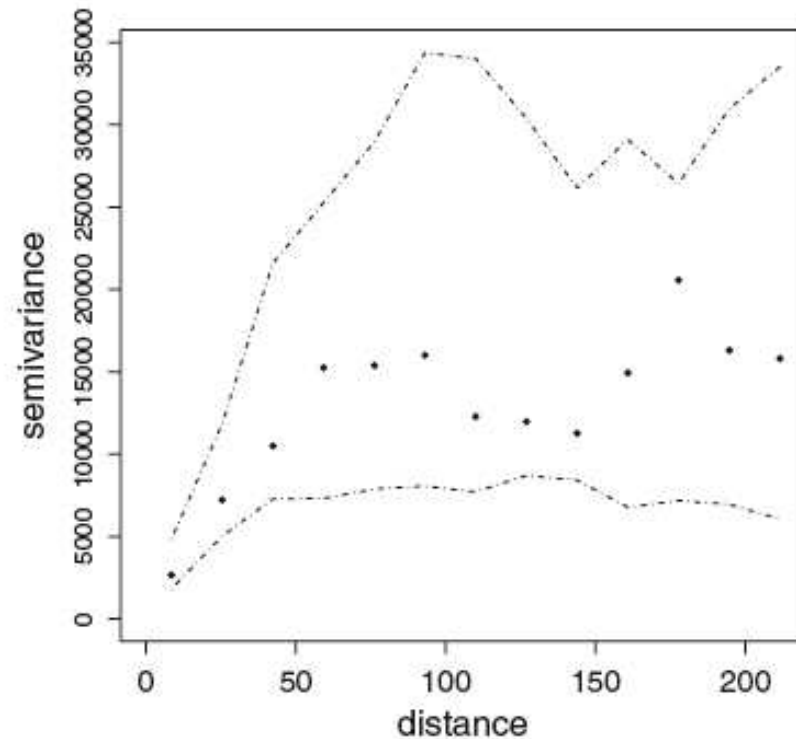


# Estimation du modèle de Matern

- (a) MCO, MCP et MV du variogramme de Matern  
(b) Estimations empiriques + *enveloppes sup* et *inf* à partir de  $m=40$  simulations du modèle de Matern estimé par MCP (intervalle de confiance à  $1 - 2/(m+1) = 95\%$ )



(a)



(b)

# `variofit` et `likfit` : estimation d'un variogramme paramétrique

- `variofit`

estime les paramètres d'un modèle de covariance (variogramme) par MCO ou par MCP, ceci à partir du variogramme empirique (variog)

- `likfit`

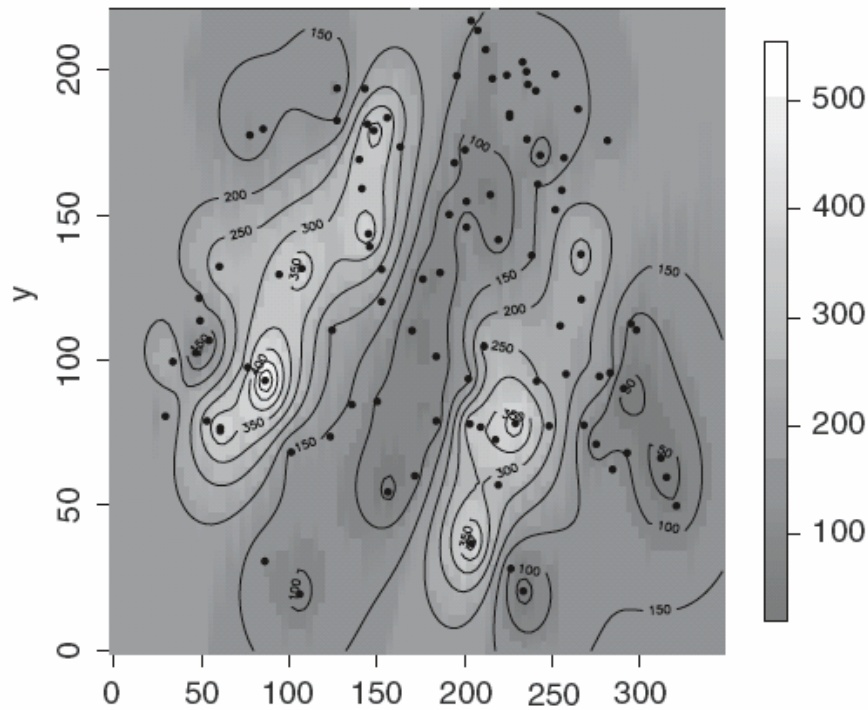
réalise l'estimation par maximum de vraisemblance

- **Exercice** : *sur un jeu de données (personnelles), estimer un modèle paramétrique par MCO et par MV. Représentations simultanées des estimations empiriques et paramétriques.*

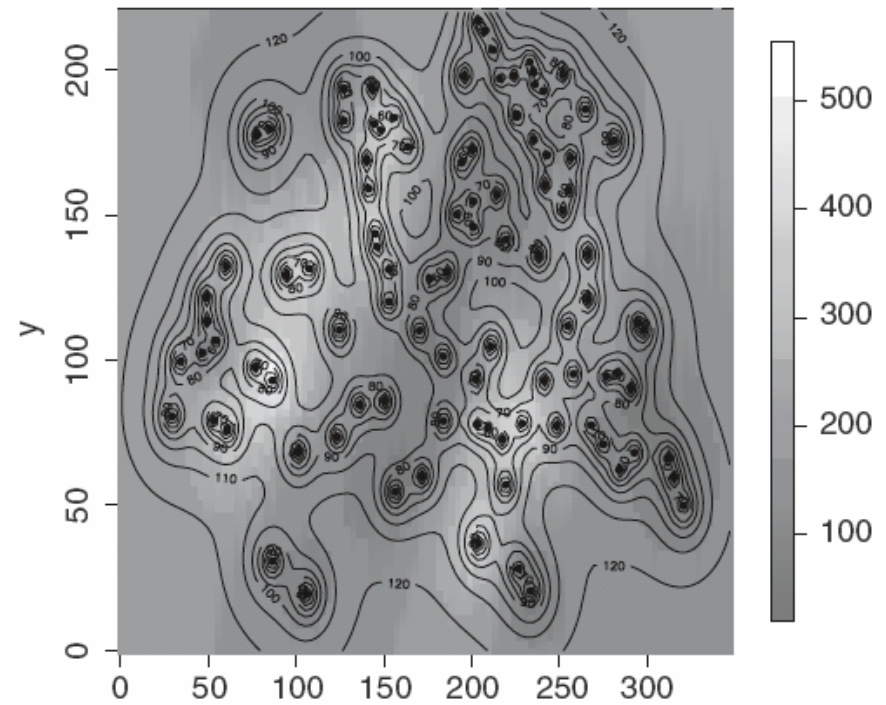
# Krigeage pour le modèle de Matern

*(données suisses)*

(a) carte des pluies et (b) de leurs écarts types

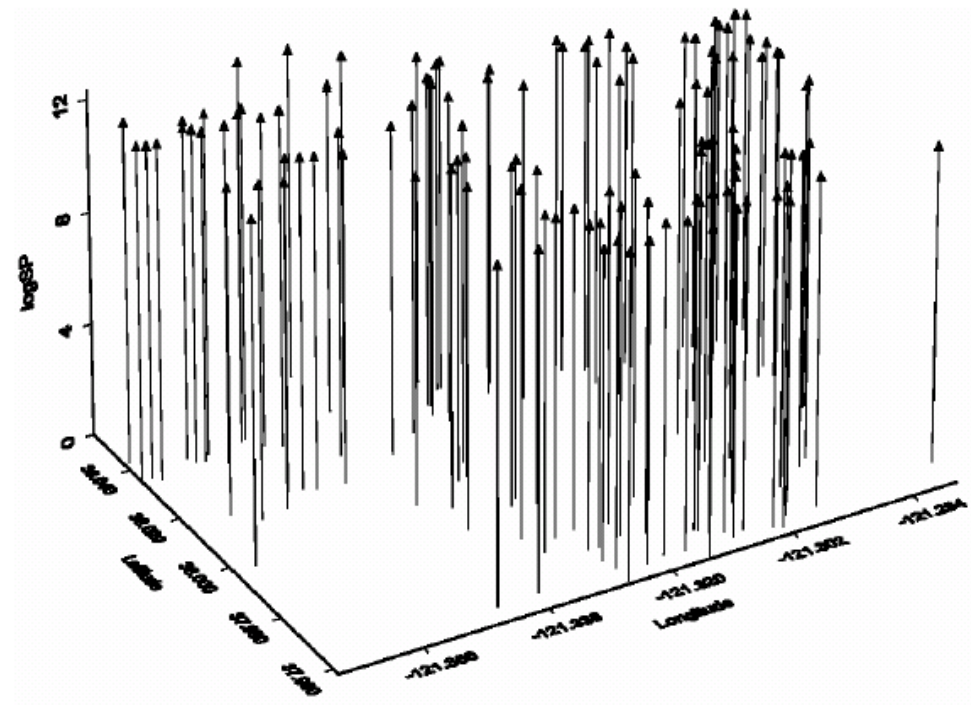


(a)



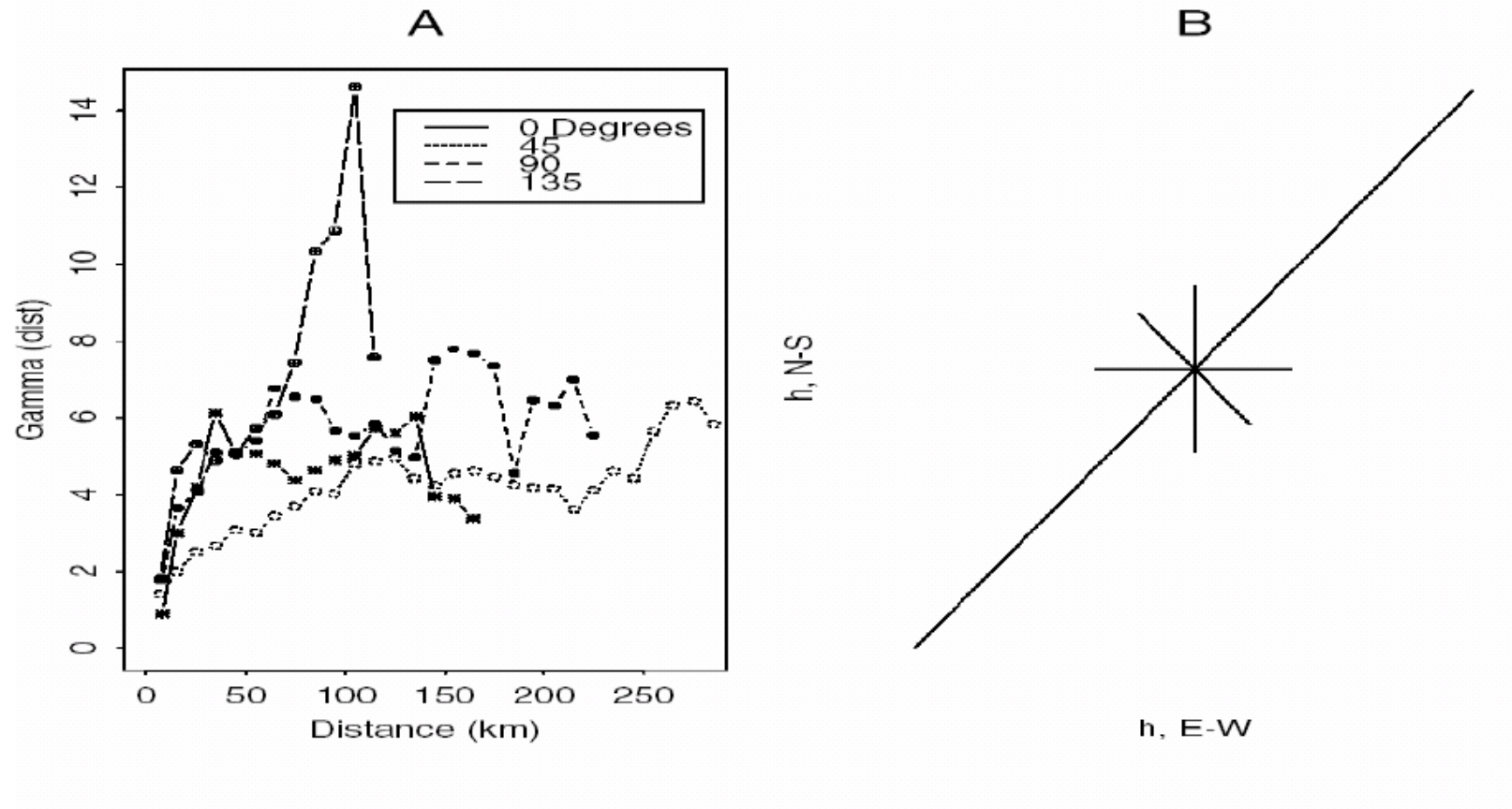
(b)

# Exemple : densité de pêche de la coquille Saint Jacques dans l'atlantique nord



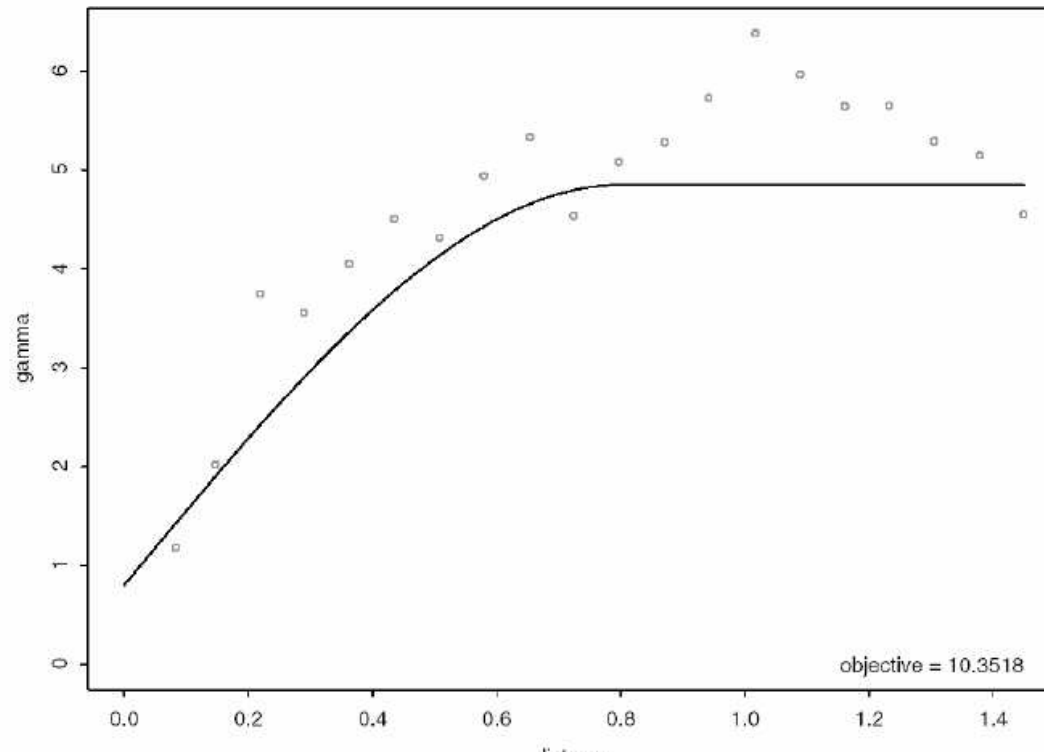
# Données de « Pêche de coquilles saint Jacques »

## Variogrammes directionnels et rose d'anisotropie.

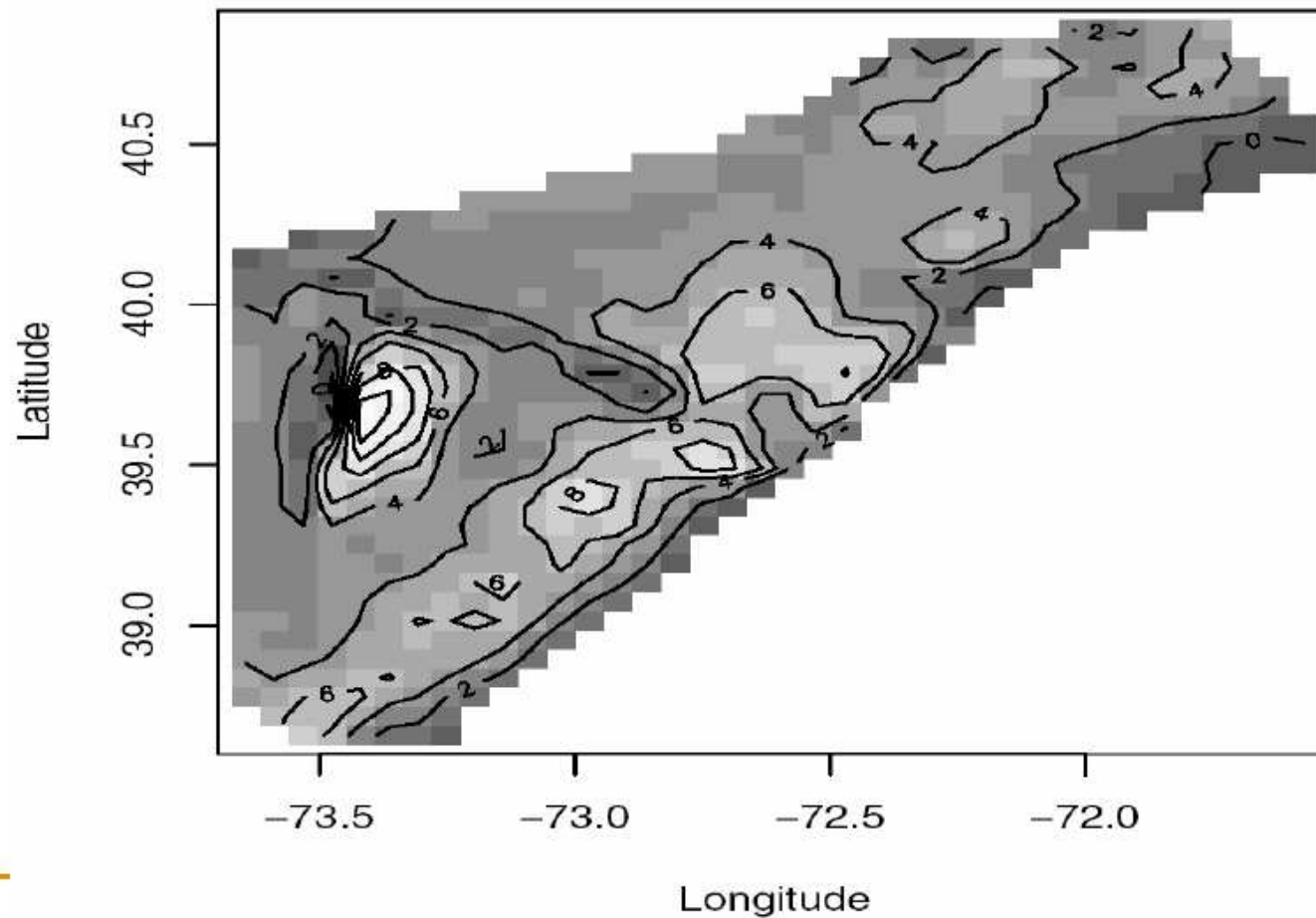


# Estimation du modèle « sphérique + pépitique » ( après transformation isotropiante )

Estimation empiriques robustes sur 20 classes de distances;  
Estimation du modèle « sphérique + pépitique » : **portée** = 0.8 ; **var**= 4.8 et bruit  
additif de variance 0.7.



# Krigeage : carte de prédiction de la densité en coquille Saint Jacques pour le modèle sphérique estimé.



# Convergence et normalité asymptotique des estimateurs

- Choix de  $k$  classes de distances identifiant le modèle paramétrique
- Régularité suffisante de la fonctionnelle d'estimation
- Faible dépendance (mélange) du champ géostatistique
- Extension des résultats au cas où il existe une tendance paramétrique (paramètres d'ordre 1, moyenne) et 2 (variogramme des résidus).

→ Voir les résultats (Lahiri, Cressie, C. Gaetan – X. G.)



# Validation d'un modèle paramétrique

- Validation croisée
- Validation par Bootstrap paramétrique

# Validation croisée d'un variogramme

- Estimer le modèle paramétrique
- Éliminer à tour de rôle une observation  $x(i)$  et la prédire par *krigeage* avec le modèle estimé
- Calculer l'Ecart Quadratique Normalisé (par la variance de prédiction) Moyen sur tous les  $x(i)$  :

$$EQNM = \frac{1}{n} \sum_{i=1}^n \frac{(X_{s_i} - \hat{X}_{s_i})^2}{\tilde{\sigma}_{s_i}^2}$$

- Si l'**EQM** proche de 1, valider le modèle (à 95% de sécurité) :

$$|EQNM - 1| \leq 1.96 \sqrt{\frac{2}{n}}$$

# Exemple : validation du modèle de Matern (données suisses)

Modèle de Matérn et trois méthodes d'estimation (MCO, MCG et MV) pour données « pluies suisses ».

	$\hat{a}$	$\hat{\sigma}^2$	$\hat{\nu}$	EQNM
MCO	17.20	15135.53	1.21	1.37
MCP	18.19	15000.57	1.00	1.01
MV	13.40	13664.45	1.31	1.09

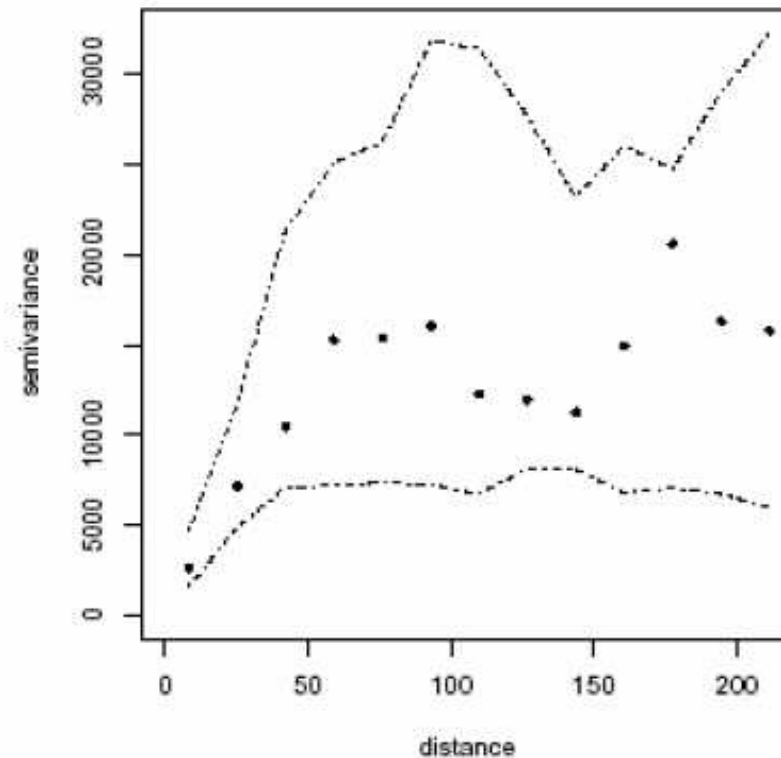
# Validation par bootstrap paramétrique

- Modèle paramétrique → estimation du paramètre
- $m$  simulations des données sous le modèle estimé
- pour chaque simulation → variogramme empirique
- enveloppes inférieure + supérieure des variogrammes empiriques  
→ *bande de confiance* au niveau  $\alpha = (1 - 2/(m+1))$
- Si le variogramme empirique des données initiales dans la bande de confiance, le modèle est validé.

# Validation du modèle de Matern

(données Parana)

- $m = 40$  simulations sous le modèle de Matern estimé
- intervalle de confiance à 95% pour le variogramme aux 13 distances
- contient les variogrammes empiriques  $\rightarrow$  *modèle de Matern valide*



# Bande de confiance, validation de modèle

- `variog.model.env`

Répétition de simulations d'un modèle (donné ou estimé) aux sites d'observations (modèle gaussien)

→ variogrammes empiriques pour chaque simulation

→ bande de confiance pour le variogramme

→ si variogramme théorique dans la bande → modèle valide

- `xvalid`

validation croisée via le krigeage : chaque observation est comparée à sa prédiction à partir des autres)

# Une étude épidémiologique : prévalence du paludisme chez l'enfant

(donnée *gambia* de *geoR*, Diggle et altri)

- **Expliquer** la *prévalence du palud* (% d'enfants malades) dans un village **s**

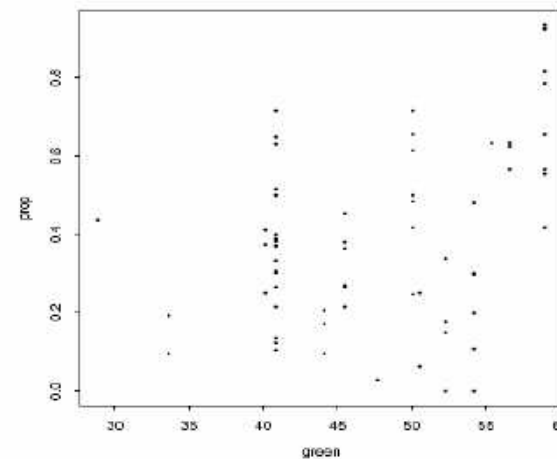
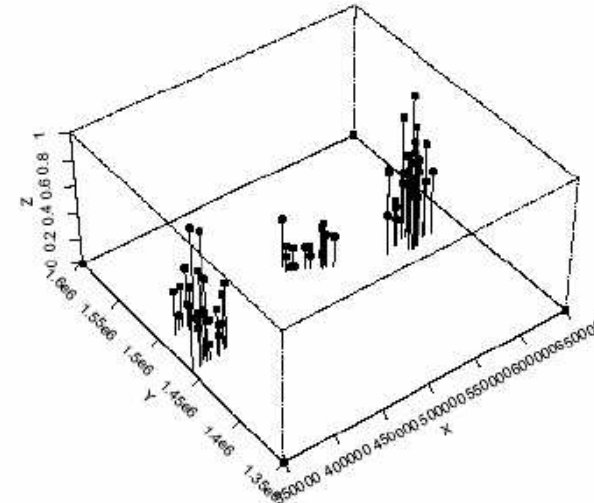
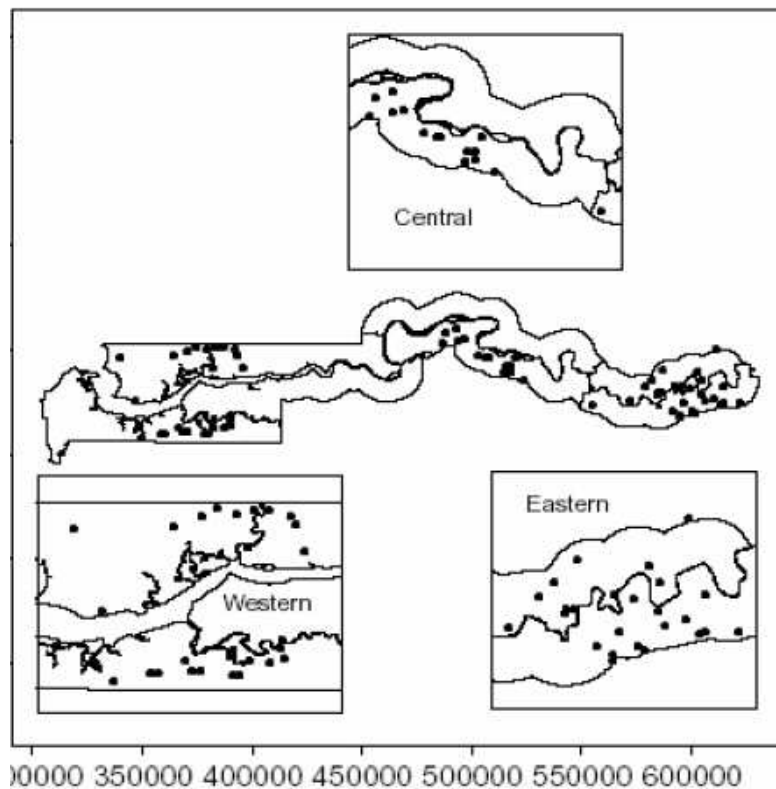
## **Covariables :**

- Existence ou non d'un centre de santé primaire
- Un indice de végétation (données satellitaires)
- Moustiquaire ou non
- Si oui, traitée ou non

- **Question** : existe-t-il un facteur de risque spatial ?

# Données malaria en Gambie

- 1 – localisations **s** des villages échantillonnés
- 2 – prévalence **prop = y/n** (**y** = nb de malades, **n** nb d'enfants)
- 3 – nuage (**green.prop**) pour indice de végétation **green**





# Le modèle log-linéaire d'étude

- $Y_{i,s}$  = présence / absence du parasite dans le sang enfant  $i$ , village  $s$
- $p$  covariables ( $z_{i,s,k}; k = 1, p$ ) : age, sexe, moustiquaire, traitée ou non, indice végétation ....

Modèle d'étude proposé : **régression Logistique**

$$\text{Logit } P(Y_{i,s} = 1) = \sum_{k=1}^p z_{i,s,k} \beta_k + U_s + S(x(s))$$

où  $U_s \sim N(0, \nu^2)$  un résidu BBG habituel

$\{S(x) : x \in \text{Gambie}\}$  est un champ gaussien stationnaire

**Analyse :**

- présence de  $S(x)$  affecte fortement l'estimation de  $\beta$
- la carte de  $\hat{S}(x)$  montre un facteur de risque spatial



---

## Géostatistique : bilan

- Inventaires des variables influentes disponibles
- Analyse graphiques préalables pour réduire leur nombre
- Propositions de modèles à l'ordre 1 (la tendance) et 2 (la covariance ou le variogramme).
- En particulier, choix a priori de régularité du variogramme
- Nuée(s) variographiques : isotropie ou non ?
- Estimation empirique puis paramétrique
- Validation de modèle (validation croisée, bootstrap paramétrique)
- Krigeage simple ou universel
- Cartes de krigeage : prédiction et précision....

## Géostatistique avec *R* : *geoR, gstat, geoRlm et RandomFields*

### *geoR*

- Nombreuses données (*gambia, parana*, etc ...)
- Quelques exemples de programmes :
  - geoR2RF*: simulation d'un GMF (Gaussian Markov Field).
  - likefit*: MV pour un GRF.
  - krige.conv*: krigeage conventionnel.
  - plotvariogram*: variogramme empirique.
  - variofit*: MV/MC d'un modèle à partir du variogramme empirique.
  - xvalid*: validation croisée d'un modèle de variogramme;
  - etc ...

## Géostatistique avec *R* (suite)

*geoRglm* : general linear spatial modèle.

*gstat* : Modélisation, estimation, prédiction en géostatistique uni-multidimensionnelle.

*Random Fields* : simulation et analyse des champs aléatoire (RF) gaussiens. Exemple de « package » :

*CondSimul* : simulation conditionnelle.

*fitvario* et *mleRF* : MC et MV pour modèles de RF.

# Modèle sur un réseau discret

## **SAR** ou **Champ de Markov**

**Deux différences entre ces 2 familles :**

(a) Spécificité des états  $E$

(b) linéarité ou non du modèle

- Pour **AR spatial** :  $E = R$  ou  $R^{**d}$  + modèle « linéaire » d'équations **Simultanées** (**SAR** gaussien ou non)
- Pour **Champ de Markov** :  $E$  général (i.e.  $\{0, 1\}$ , fini,  $N$ ,  $R$ ,  $R+$ ,  $R^{**d}$  etc...) et modèle conditionnel (en général) non linéaire

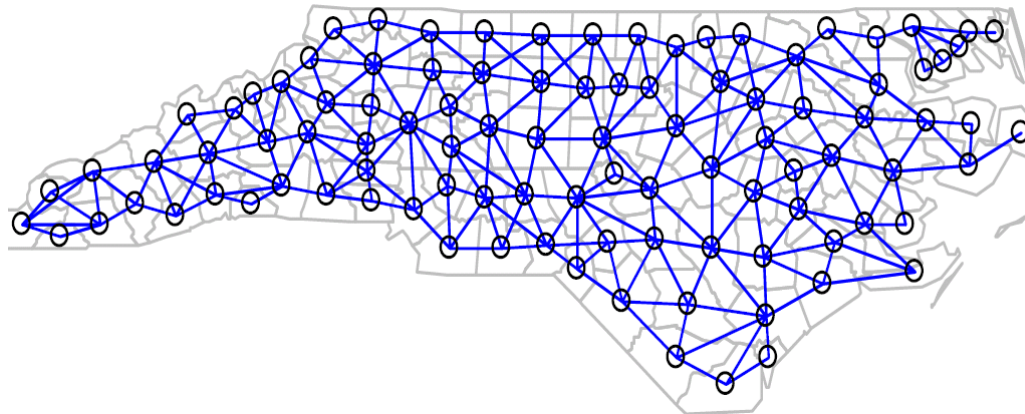
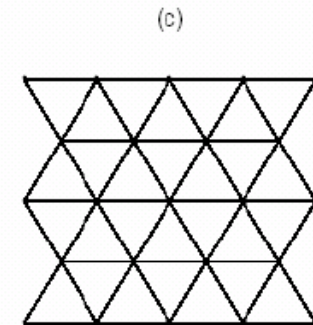
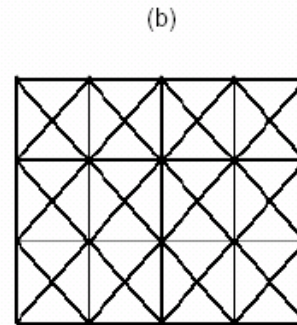
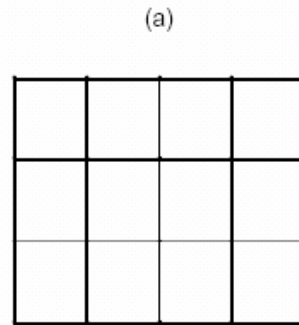
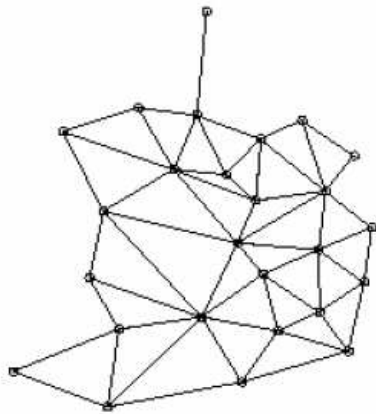
Pour  $E = R$ , les **CAR** sont aussi Markov

# **(I) Champ de Markov**

# Champ de Markov sur réseau

- Valables pour un espace d'états général (pas limité à  $\mathbf{R}$ ).
- Lattice régulier (imagerie, agronomie) ou non (épidémiologie, , environnement, etc...)
- Modèle définit par ses spécifications conditionnelles locales.
- Outils spécifiques : simulation MCMC, estimation PVC.

# Exemple de réseau : *eire*, 3 réseaux réguliers, et *sids*





# Champ de Markov sur **S** fini

- **S** réseau fini (régulier ou non)
- Loi de **X** caractérisée par ses conditionnelles
- Loi conditionnelle est « locale »
- Espace d'état **E** général : binaire, fini, N, R+, R, R\*\*p, mixte (gris x variété), etc

$$\pi_A(x_A | x^A) = \pi_A(x_A | x_{\partial A}).$$

# Champ de Gibbs

1. **Potentiels** réels  $\Phi = \{\Phi(x(A))\}$  définis sur une **famille de parties  $A$**  de  $S$
2.  $\Phi \rightarrow$  **énergie**  $U(\Phi; \Lambda)$  sur  $\Lambda$  conditionnelle  $\partial\Lambda$
3. S'assurer que  $\exp\{U(\Phi; \Lambda)\}$  intégrable (potentiel admissible)
4. Le champ de Gibbs est de **log-densité conditionnelle** à  $\partial\Lambda$  proportionnelle à  $\exp\{U(\Phi; \Lambda)\}$

# Energie et loi conditionnelles d'un champ de Gibbs

Soit  $\Lambda$  une partie de  $S$ ,  $\partial\Lambda$  son voisinage

$$U_{\Phi}(x(\Lambda) \mid x(\partial\Lambda)) = \sum_{A:A \cap \Lambda \neq \emptyset} \Phi_A(x(A))$$

et loi conditionnelle

$$\pi(x(\Lambda) \mid x(\partial\Lambda)) = Z^{-1}(x(\partial\Lambda)) \times \exp U_{\Phi}(x(\Lambda) \mid x(\partial\Lambda))$$

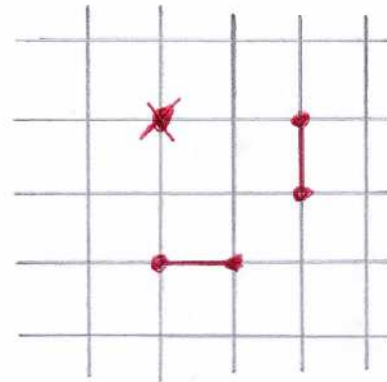
# Spécification de Gibbs

- Famille des parties  $A$  sur lesquelles sont définis les potentiels
- Les potentiels  $\phi(A)$
- Vérifier l'admissibilité de  $\exp U(\Phi)$
- **Exemple** : famille exponentielle de potentiels de paramètre  $\theta$

$$\Phi_A(y) = \langle \theta, \phi_A(y) \rangle \text{ où } \phi_A \in \mathbb{R}^p \text{ connues,}$$

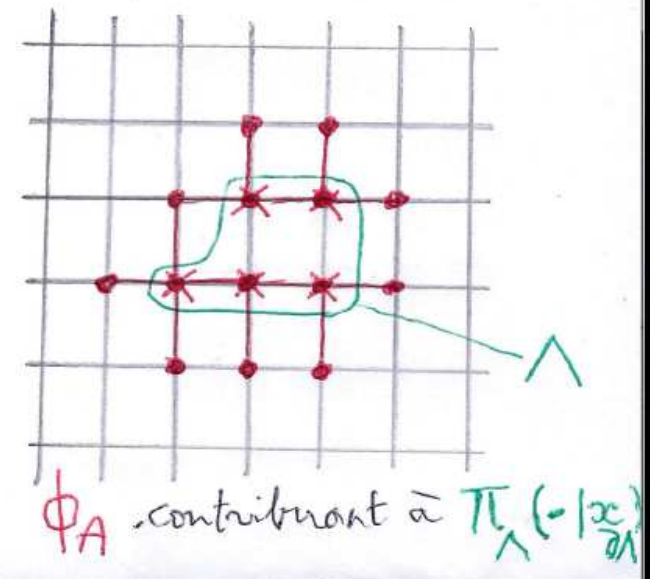
# Exemple de potentiels contribuant à une loi conditionnelle (loi aux 4 ppv)

- Cliques à un point  $\{x\}$
- À deux points  $\bullet \text{-----} \bullet$



Cliques à 1 point  $\times$   
 et  $\bullet$  à 2 points (H-Voisins  
 V-Voisins)

- Potentiels contribuant à  $\pi(\Lambda / \partial\Lambda)$



$\phi_A$  contribuant à  $\pi_{\Lambda}(-|x)$   
 $\frac{\partial \Lambda}{\partial \Lambda}$

# Modèle d'Ising : état $E = \{-1,+1\}$ et sites $S = \{1,2,\dots,n\}^{**2}$

- **Cliques** : singletons et paires de ppv (plus proche voisin)
- **Potentiels** :  $\Phi(x(i))=\alpha.x(i)$  et  $\Phi(x(i),x(j))=\beta.x(i).x(j)$
- **Energie** :  $U_{\Lambda}(x_{\Lambda}|x^{\Lambda}) = \alpha \sum_{i \in \Lambda} x_i + \beta \sum_{i \in \Lambda, j \in S : \langle i,j \rangle} x_i x_j,$
- **loi conditionnelle en  $i$**  :

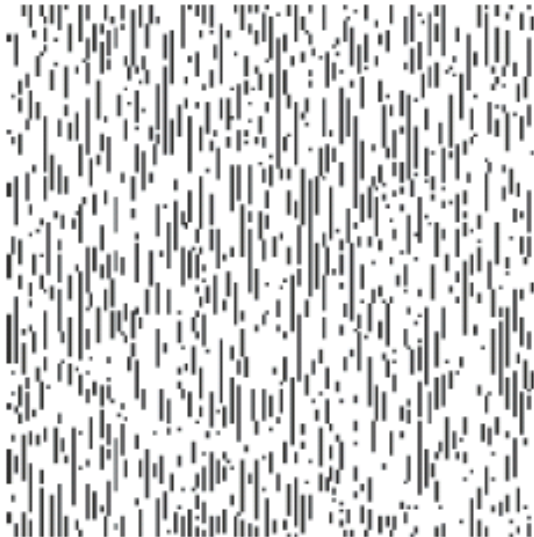
$$\pi_i(x_i|x^i) = \pi(x_i|x^i) = \frac{\exp x_i(\alpha + \beta v_i(x))}{2ch(\alpha + \beta v_i(x))}$$

# Généralisations du modèle d'Ising

- **Etats  $\{0, 1\}$**  : présence - absence (écologie); sain - malade (épidémiologie)
- **Plus d'états** : niveaux de gris,  $E$  fini ( $K$  variétés)
- **Anisotropie** :  $\beta(H)$  pour horizontal,  $\beta(V)$  pour vertical
- **Non stationnaire** :  $\alpha(i)$  et  $\beta(i,j)$  suivant les sites  $i, j$
- **Élargissement du voisinage  $\partial i$**  : i.e. aux 8 ppv
- **Potentiels de triplets, quadruplets ....**
- **Simulations par MCMC** (échantillonneur de Gibbs)

## Exemples : 3 textures binaires aux 8 – ppv

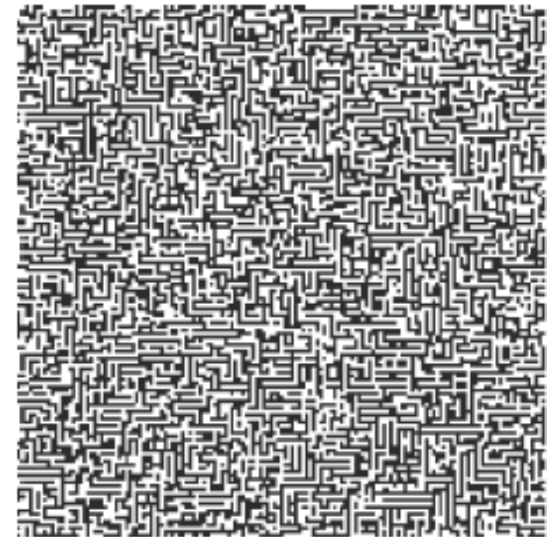
- Potentiels au plus de paires  $\rightarrow$  5 paramètres
- Simulation par *échantillonneur de Gibbs* (*AntsInfields*)
- Arrêt après 3000 itérations



0.0	1.0	0.0
-1.0	<b>2.0</b>	-1.0
0.0	1.0	0.0



3.0	1.0	0.0
-1.0	<b>0.0</b>	-1.0
0.0	1.0	3.0



-0.4	0.4	-0.4
0.4	<b>0.0</b>	0.4
-0.4	0.4	-0.4



# Modèle de Gibbs à nombre d'états fini (modèle de Potts)

- **Etats** :  $E = \{a(1), a(2), \dots, a(K)\}$ ,  $K$  états
- **Potentiels** : singletons et paires de sites voisins

$$\begin{aligned}\Phi_{\{i\}}(x) &= \alpha_k, & \text{si } x_i = a_k, \\ \Phi_{\{i,j\}}(x) &= \beta_{k,l} = \beta_{l,k}, & \text{si } \{x_i, x_j\} = \{a_k, a_l\}\end{aligned}$$

- **Energie** :  $n(k)$  = nb sites modalité  $k$ ;  $n(k,l)$  = nb de sites voisins de modalités  $(k,l)$ .

$$U(x) = \sum_k \alpha_k n_k + \sum_{k < l} \beta_{kl} n_{kl}$$

# Modèle de Potts échangeable

*Très utile en traitement d'image*

- Tous les états ont un comportement analogue c'est-à-dire :  $\alpha(k) = \alpha$  et  $\beta(k,l) = \beta$  pour tout  $k,l$
- Dans ce cas il y a un seul paramètre de dépendance spatiale  $\beta$  et la loi jointe vaut :

$$\pi(x) = Z^{-1} \exp\{-\beta n(x)\}$$

# Modèle échangeable à 3 états

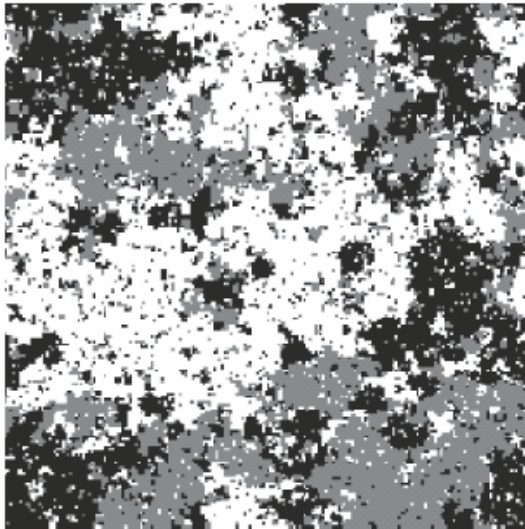
$\beta \uparrow$  augmente la régularité géométrique des plages constantes

(a)  $\beta = 0.5$

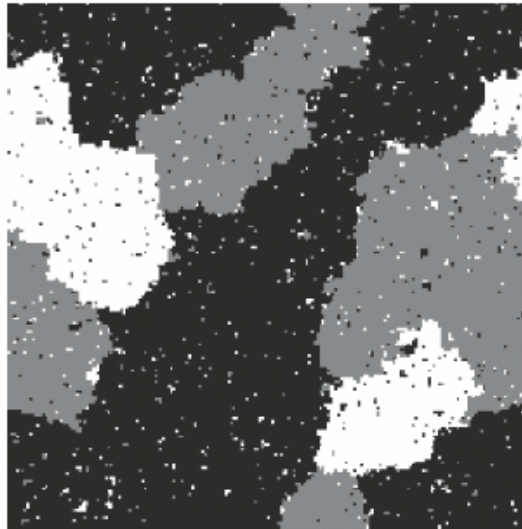
(b)  $\beta = 0.6$

(c)  $\beta = 0.7$

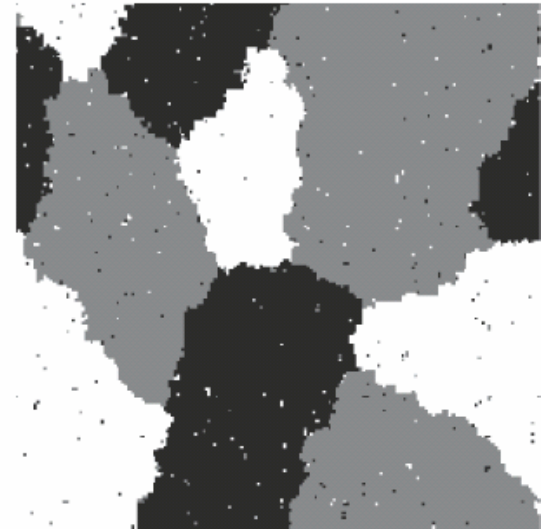
(Echantillonneur de Gibbs à 5000 itérations avec *AntsInFields*)



(a)



(b)



(c)

# Un champ gaussien est un champ de Gibbs

- Soit  $X = N(\mu, \Sigma)$  de moyenne  $\mu$  et covariance  $\Sigma$  ( $Q$  inverse  $\Sigma$ )
- Alors  $X$  est de Gibbs avec :
  - potentiels de *singletons* et de *paires*
  - énergie  $-U$
- Obtention des lois conditionnelles comme champ de Gibbs

$$U(x) = \frac{1}{2} {}^t(x - \mu)Q(x - \mu),$$

de potentiels de singletons  $\Phi_{\{i\}}$  et de paires  $\Phi_{\{i,j\}}$  :

$$\Phi_{\{i\}}(x) = x_i \sum_{j:j \neq i} q_{ij} \mu_j - \frac{1}{2} q_{ii} x_i^2 \quad \text{et} \quad \Phi_{\{i,j\}}(x) = -q_{ij} x_i x_j \quad \text{si } i \neq j.$$

# **Simulation d'un champ de Markov**

# Simulation par dynamique de chaîne de Markov

(MCMC pour *Monte Carlo Markov Chain*)

- **Objectif** : simuler une loi  $\Pi$  sur  $E$
- **Principe MCMC** : construire une chaîne de Markov  $(X(n), n > 0)$  sur  $E$  de transition  $P(x, \bullet)$  t.q. :
  - 1 –  $P$  est **irréductible** (tous les états communiquent),
  - 2 –  $P$  est  $\Pi$  - invariante ( $\Pi P = \Pi$ )
  - 3 –  $P$  est **apériodique**

**Propriété** : Si (1-2-3), la loi de  $X(n)$  tend vers  $\Pi$

# Irréductible, invariance et apériodicité

- **Irréductible** : la chaîne  $P$  fait communiquer tous les états de  $E$  entre eux
- **$\Pi$  - invariante** ( $\Pi P = \Pi$ ) : si la loi de  $X(n)$  est  $\Pi$ , celle de  $X(n+1)$  est encore  $\Pi$  ( $\Pi$  est la loi invariante de la chaîne)
- **Périodique** : si il existe une partition  $\{E(1), E(2), \dots, E(k)\}$  de  $E$  t.q. la chaîne circule successivement dans  $E(1) \rightarrow E(2) \rightarrow \dots \rightarrow E(k) \rightarrow E(1)$  etc
- **Apériodique** : si non périodique

# Comment construire une telle chaîne ?

- **Apériodicité et irréductibilité** se vérifient au cas par cas
- Pas facile de trouver  $P$  qui soit  $\Pi$  - invariante :  $\Pi$  est vecteur propre de  $P$  associée à la valeur propre  $1$  !
- Condition suffisante assurant la  $\Pi$  – invariance de  $P$  :  
la  $\Pi$  – réversibilité :  
pour tout  $x, y$  :  $\Pi(x)P(x,y) = \Pi(y)P(y,x)$



# Deux algorithmes markoviens classiques

- **Echantillonneur de Gibbs** sur un espace produit  $E^{**S}$  :  
 $S=\{1,2,\dots,n\}$  ensemble des sites, état  $E$  en chaque site
- **L'algorithme de Metropolis** (espace  $E$  général)

Pour l'un et l'autre, la transition  $P$ , est, par construction,  $\pi$  – réversible, donc  $\pi$  – invariante

→ **simulateur MCMC** de  $\pi$  si on vérifie de plus que  $P$  est apériodique et irréductible

# Simulation par échantillonneur de Gibbs d'une loi $\pi$ sur $E^{**}S$

- Loi  $\pi$  sur espace produit  $E^{**}S$
- Connaître les lois conditionnelles en tout  $i$  :  $\pi_i(y_i|x^i)$
- Relaxation « site par site » sur  $S$  suivant la conditionnelle :

$$P_i(x, y) = \pi_i(y_i|x^i)1(x^i = y^i)$$

- Un « balayage » de  $S$  : itérer de  $i = 1, n \rightarrow$  la transition  $P$
- Itérer les balayages
- **Propriété** :  $P$  est  $\pi$  – réversible.

*Donc l'échantillonneur de Gibbs simule approximativement  $\pi$   
Après un grand nombre de balayages de  $S$ .*

# Transition pour un balayage de **S**

- Visite séquentielle de **S** :  $1 \rightarrow 2 \rightarrow 3 \rightarrow \dots \rightarrow n$
- Au *i-ème* pas, relaxation au site *i*
- Enchaînement sur un balayage donne la transition :

$$P_S(x, y) = \prod_{i=1}^n \pi_i(y_i | y_1, \dots, y_{i-1}, x_{i+1}, x_{i+2}, \dots, x_n)$$

# La transition $P$ de l'algorithme de Métropolis

- $\Pi$  loi sur espace d'état  $E$  général,  $x, y$  deux états

**Construction de  $P$  en deux étapes :**

- 1 – *Proposition de changement*  $x \rightarrow y$  suivant une transition  $Q(x,y)$  symétrique ( $Q$  est la proposition de changement)
- 2 – *Acceptation du changement* avec une probabilité  $a(x,y)$

**Propriété :** si  $a(x,y) = \min \{1, \Pi(y)/\Pi(x)\}$ , alors  $P$  est  $\Pi$  – réversible (donc  $\Pi$  - invariante)

Si de plus  $P$  est irréductibilité et apériodique, on a un autre algorithme de simulation de  $\Pi$

# Algorithme de Métropolis

- 1 -  $x$  état initial. Changement  $x \rightarrow y$  suivant  $Q(x,y)$
- 2 - Si  $\Pi(y) \geq \Pi(x)$ , garder  $y$ .
- 3 – Sinon, tirer  $U$  uniforme sur  $[0, 1]$  :
  - (a) si  $U > p = \Pi(y)/\Pi(x)$ , garder  $x$ .
  - (b) Sinon, garder  $y$ .
- 4 – Revenir en (1)

L'algorithme de *Métropolis – Hastings* correspond à une proposition de changement  $Q$  non symétrique.

**Remarque importante** : il suffit de **connaître  $\Pi$  à un facteur près** pour construire cet algorithme (le cas pour  $\Pi$  de Gibbs).

# Exemple : simulation d'un Ising isotropique aux 4-ppv

## (I) - Par échantillonneur de Gibbs

Les lois conditionnelles en chaque site  $i$  sont explicites en terme de  $v(i)$ , la somme aux 4-ppv

$$\pi(x) = Z^{-1} \exp\left\{\alpha \sum_i x_i + \beta \sum_{\langle i,j \rangle} x_i x_j\right\},$$
$$\pi_i(x_i | x^i) = \frac{\exp x_i(\alpha + \beta v_i)}{2ch(\alpha + \beta v_i)}.$$

# AntsInFields

« **Boite noire** » illustrant le livre de G. Winkler (*Springer, 2002*) :  
« *Image Analysis, random fields and dynamic MC Methods* »

*Largement buggée et non ouverte à la programmation.*

*Outil de démonstration, sur des thèmes d'analyse d'image et de statistique de champ de Gibbs.*

## **Exemple : § 3 – 2**

→ simulation **Ising** isotropique aux 4-ppv par échantillonneur de Gibbs  
 *$h=0$ ,  $b=0$ , 0.2 et 0.4*. Voir petit à petit se former des plages

→ simulation d'un modèle de **Potts** à 4 niveaux de gris, paramètres :  
nombre de classes,  $h$  et  $b$  et distance retenue entre les configurations voisines

# Simulation d'un Ising (suite)

## (II) - Métropolis par échange de spins

- 1 - on tire au hasard 2 sites  $i$  et  $j$  et on permute les spins  $x(i)$  et  $x(j)$  : on passe ainsi de  $x \rightarrow y$  avec une probabilité de transition  $Q(x,y)$
- 2 – le quotient  $\Pi(x)/\Pi(y)$  s'explique facilement en fonction de  $x(i)$ ,  $x(j)$ ,  $v(i)$  et  $v(j)$  (cf. poly)
- 3 – Mettre en œuvre Métropolis.

$$Q(x, y) = \begin{cases} \frac{2}{n^2(n^2-1)} & \text{pour un tel échange,} \\ 0 & \text{sinon} \end{cases}$$



# Champ de Markov et champ de Gibbs

- $S = \{1, 2, 3, \dots, n\}$  et  $G$  graphe symétrique sur  $S$
- $\langle i, j \rangle$  :  $i$  et  $j$  voisins pour  $G$ .
- $\partial A$  = voisinage de  $A$  pour  $G$

**Clique** de  $G$  : singletons + parties  $A$  t.q. les points de  $A$  sont tous voisins

- $C(G)$  = toutes les cliques de  $G$

**Champ de Markov :**

$$\pi_A(x_A | x^A) = \pi_A(x_A | x_{\partial A}).$$

# Le théorème de Hammersley – Clifford

## Gibbs $\equiv$ Markov

- $\pi$  un champ de Markov pour graphe  $G$
- *Positivité* : pour tout  $x$ ,  $\pi(x) > 0$
- **Propriété (H-C)** : alors  $\pi$  est- un champ de Gibbs dont les potentiels sont limités aux cliques de  $G$
- **Réciproque** : tout champ de Gibbs est un champ de Markov pour le graphe engendré par les potentiels de Gibbs.

# Recollement de lois conditionnelles

- **Objectif** : définir un modèle à partir de ses spécifications locales
- **En général**, des spécifications locales « ne se recollent pas »
- $E$  sous ensemble réel
- Les spécifications de la « famille exponentielle » ci-dessous se recollent.

**Résultat** : si pour tout  $i$

**Recollement des  $\pi(x_i / x_{\partial i})$**

$$\log \pi(x_i / x_{\partial i}) = A_i(x_{\partial i})B(x_i) + C_i(x_i) + D_i(x_{\partial i})$$

$$\text{où } B_i(0) = C_i(0) = 0 \text{ et } A_i(x_{\partial i}) = \alpha_i + \sum_{j \in \partial i} \beta_{ij} x_j, \text{ et } \beta_{ij} = \beta_{ji},$$

Alors  $\implies$   $\pi$  est un champ de Gibbs de potentiels :

$$\Phi_i(x_i) = \alpha_i B_i(x_i) + C_i(x_i) \text{ et } \Phi_{i,j}(x_i, x_j) = \beta_{ij} B_i(x_i) B_j(x_j).$$

# Auto – modèle de Besag (1974)

- $E$  sous ensemble de  $R$
  - Famille exponentielle du type précédent
- Alors les lois conditionnelles se recollent en la loi de Gibbs  $\pi$  :

$$\pi(x) = Z^{-1} \exp\left\{ \sum_{i \in S} \Phi_i(x_i) + \sum_{\langle i; j \rangle} \beta_{ij} x_i x_j \right\}$$

# Auto-modèle de Markov : auto - régression pour espace $E$ général

- $E = \{0, 1\}$  : **auto-logistique** (états binaires)
- $E = \{0, 1, 2, \dots, K\}$  : **auto-binomial**
- $E = N$  : **auto-poisson** (comptage en épidémiologie)
- $E = R_+$  : **auto-exponentiel** (Gamma), pluviométrie
- $E = R$  ( $R^{**d}$ ) : **auto-gaussien**
- Possibilité d'ajouter des *covariables explicatives*

# Modèle Auto-Logistique : $E = \{0, 1\}$

- paramètre  $\theta(x(i))$  et loi **Logit conditionnel**

$$\theta_i(x_i) = \{\alpha_i + \sum_{j:\langle i,j \rangle} \beta_{ij}x_j\},$$
$$\pi_i(x_i|x^i) = \frac{\exp x_i \{\alpha_i + \sum_{j:\langle i,j \rangle} \beta_{ij}x_j\}}{1 + \exp \{\alpha_i + \sum_{j:\langle i,j \rangle} \beta_{ij}x_j\}}.$$

- **L'énergie jointe** du champ de Gibbs est

$$U(x) = \sum_i \alpha_i x_i + \sum_{\langle i,j \rangle} \beta_{ij} x_i x_j.$$

- Généralisation à l'**auto-binomial**

# Auto – modèle de Poisson : $E = N$

(comptage en épidémiologie, etc)

- la loi conditionnelle Poisson suit un MLG
- admissibilité :  $\beta < 0$  (compétition)
- coopération possible en bornant  $E$  à  $K < \infty$

Si  $\forall i \in S : \pi(x_i | x_{\partial i}) = \mathcal{P}(\lambda_i(x_{\partial i}))$ , un MLG :

$$\log(\lambda_i(x_{\partial i})) = \alpha_i + \sum_{j \in \partial i} \beta_{ij} x_j \text{ où } \alpha_i + \sum_{j \in \partial i} \beta_{ij} x_j = \beta_{ji}.$$

*Admissibilité* :  $\forall i, j, \beta_{ij} < 0$  (compétition).

Alors  $\pi$  est d'énergie jointe

$$U(x) = \sum_{i \in S} \alpha_i x_i + \log(x_i!) + \sum_{\langle i, j \rangle} \beta_{ij} x_i x_j$$

# Auto – exponentiel : $E = R_+$

(variable  $>0$  : pluviométrie, etc)

- loi conditionnelle exponentielle suit un MLG
- admissibilité :  $\beta < 0$  (compétition)
- coopération possible en bornant  $E$  à  $K < \infty$

Auto-Exponentiel : si  $\forall i \in S : \pi(x_i | x_{\partial i}) = \text{Exp}(\mu_i(x_{\partial i}))$ , un MLG :

$$\log(\mu_i(x_{\partial i})) = \alpha_i + \sum_{j \in \partial i} \beta_{ij} x_j \text{ où } \alpha_i + \sum_{j \in \partial i} \beta_{ij} x_j = \beta_{ji}.$$

Admissibilité :  $\forall i, j, \alpha_i > 0$  et  $\beta_{ij} \geq 0$  (compétition).

Alors  $\pi$  est d'énergie jointe

$$U(x) = -\left\{ \sum_{i \in S} \alpha_i x_i + \sum_{\langle i, j \rangle} \beta_{ij} x_i x_j \right\}$$



# Auto – modèle avec covariables $z$

- Il y a trop de paramètres  $\alpha$  et  $\beta$  si le modèle est non stationnaire
- Modéliser les  $\alpha, \beta$  à partir de covariables  $z$  :

*Exemple :*

Poids connus :  $(a_i), (w_{ij})$  symétrique

Covariable :  $z = (z_i), z_i \in \mathbb{R}^p$  observable sur  $S$

$$\beta_{ij} = \delta w_{ij} \text{ et } \alpha_i = a_i \times {}^t \gamma z_i$$

$\implies$  modèle à  $(p + 1)$  paramètres.

# Estimation d'un champ de Markov

3 procédures

- **Max de Vraisemblance** : efficace, difficile (cste de normalisation) → méthodes numériques MCMC
- **Pseudo Vraisemblance Conditionnelle (PVC)** : facile à mettre en place, bonnes propriétés, proche MV si peu de dépendance spatiale.
- **Codage** : facile, moins efficace, test de chi2 direct

# Maximum de vraisemblance sur $D(n)$

Vraisemblance conditionnelle à  $x(\partial D_n)$  :

$$\pi_n(x(D_n) / x(\partial D_n); \theta) = Z^{-1}(\partial D_n; \theta) \exp U(x(D_n) | x(\partial D_n); \theta)$$

où  $Z$ , la constante de normalisation, vaut (cas  $E$  fini) :

$$Z(\partial D_n; \theta) = \sum_{y(D_n)} \exp U(y(D_n) | x(\partial D_n); \theta).$$

- L'estimation du MV est convergente si  $\pi$  appartient à famille exponentielle invariante par translation ( $S = Z^{**2}$ )
- Difficulté de calcul de la constante de normalisation  $Z$   
→ calcul de  $Z$  par MCMC ou par algorithme de score

# Pseudo – Vraisemblance Conditionnelle (PVC – Besag : 1974)

$$PVC_n(X) = \prod_{i \in D_n} \pi_i(x_i | x_{\partial i}; \theta)$$

*Exemple auto -logistique isotropique aux 4 -ppv :*

$$PVC_n(x) = \prod_{i \in D_n} \frac{\exp x_i (\alpha + \beta \sum_{j: \langle i, j \rangle} x_j)}{1 + \exp(\alpha + \beta \sum_{j: \langle i, j \rangle} x_j)}$$

- PVC = produit en chaque site *i* des probabilités conditionnelles
- *Bonne fonctionnelle* d'estimation (convergence, normalité sous des hypothèses raisonnables)
- Pour un CAR gaussien  $\equiv$  MCO sur les résidus
- Obtention estimation via logiciel dédié aux MLG .....
- **Attention** : *calcul spécifique* de la variance d'estimation

# Influence du taux de nitrate des eaux sur la mortalité par cancer, Valence – Espagne

(Ferrandiz et al. *Biometrics* – 1995)

**Auto modèle poissonnien** ( $Y(i)$  variable de comptage);

**Covariables** :  $x(1)$  = % population > 40 ans; et  $x(2)$  = % nitrate dans l'eau.

**Matrice de poids** : fonction des populations  $u(i)$  et des distances  $d(i,j)$ ;

$$\langle i, j \rangle \text{ si } a_{ij} = \frac{\sqrt{u_i \times u_j}}{d_{ij}} > a \text{ et } \gamma_{ij} = \gamma \times a_{ij}$$

$$\log(\lambda_i) = \alpha + \log(u_i) + \beta_1 x_{i1} + \beta_2 x_{i2} + \gamma \sum_{j : \langle i, j \rangle} a_{i,j} y_j$$

# Influence du taux de nitrate dans les eaux sur la mortalité par cancer (suite)

$$\log(\lambda_i) = \alpha + \log(u_i) + \beta_1 x_{i1} + \beta_2 x_{i2} + \gamma \sum_{j : (i,j)} a_{i,j} y_j$$

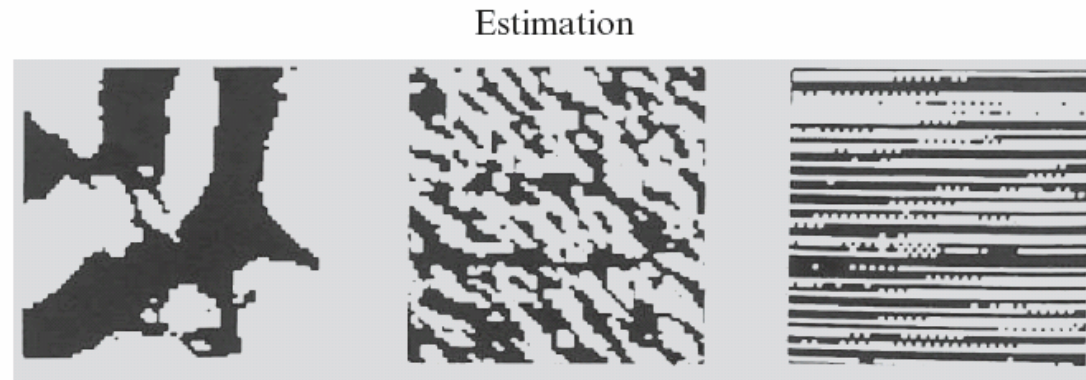
Modèle	$\alpha$	$\gamma$	$\beta_1$	$\beta_2$	$\chi^2$
Constant	-7.3	*****	*****	*****	483.1
AR-Poisson : PVC	-7.76	$-6.28 \cdot 10^{-9}$	*****	*****	384.2
MV	-7.76	$-6.52 \cdot 10^{-9}$	*****	*****	363.2
Régression	-8.91	*****	2.96	$-1.96 \cdot 10^{-3}$	323.2
Complet : PVC	-8.776	$-2.83 \cdot 10^{-9}$	2.69	$-2.15 \cdot 10^{-3}$	333.0
MV	-8771	$-2.80 \cdot 10^{-3}$	2.67	$-2.17 \cdot 10^{-3}$	309.1

# Estimation et simulation de 3 textures binaires:

(a) cailloux; (b) liège, (c) rideau (Cross et Jain)

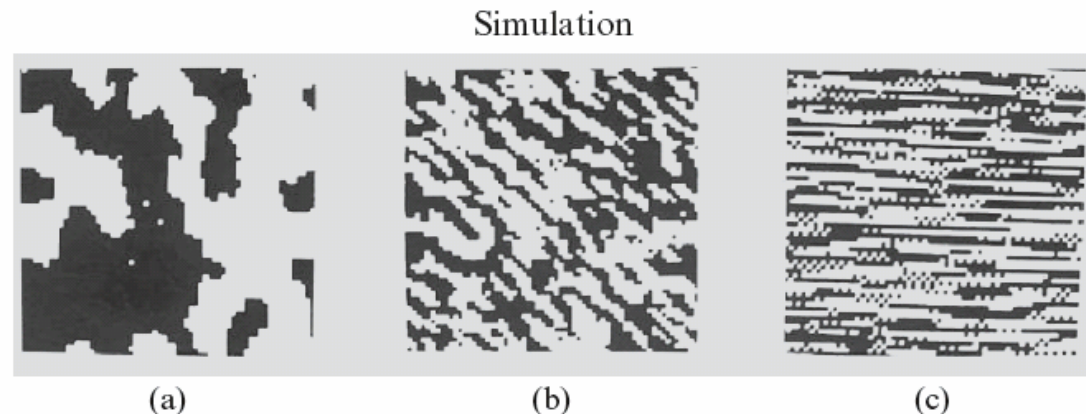
- 3 textures binaires réelles de taille 64 x 64 →

*Estimation* par PVC de 3 modèle auto-logistique



- *Simulation* des textures → estimées (éch. de Gibbs)

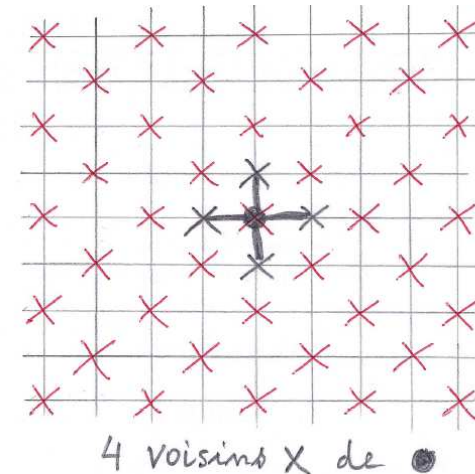
- **Bilan** : bonne adaptation de:  
(1) la modélisation Markov,  
(2) l'estimation par PVC et  
(3) la simulation par éch. de Gibbs.



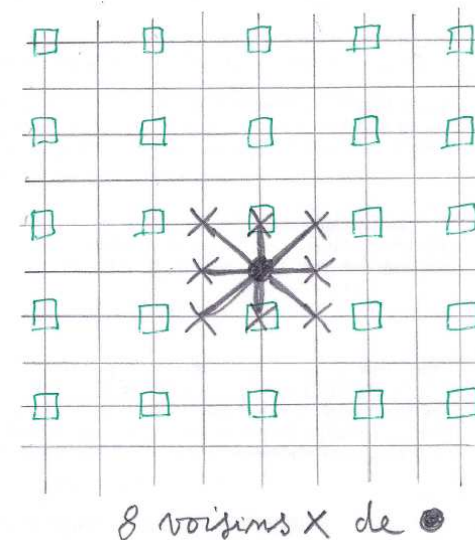
# C « Codage » de S

**Définition :** C est un **codage** de S si 2 sites  $s \neq t$  de C ne sont jamais voisins

- **Ex 1 :** les x rouges pour la relation aux 4 ppv →



- **Ex 2 :** les □ verts pour la relation aux 8 ppv →





# Estimation par **C** - codage

- Soit **C** un ensemble de codage de **S**
- **La propriété fondamentale** : indépendance des  $X(s)$ ,  $s$  dans **C**, conditionnellement aux  $x(S \setminus C)$  extérieurs
- La vraisemblance sur **C** conditionnelle aux  $x(S \setminus C)$  est (exactement) *le produit des lois conditionnelles*

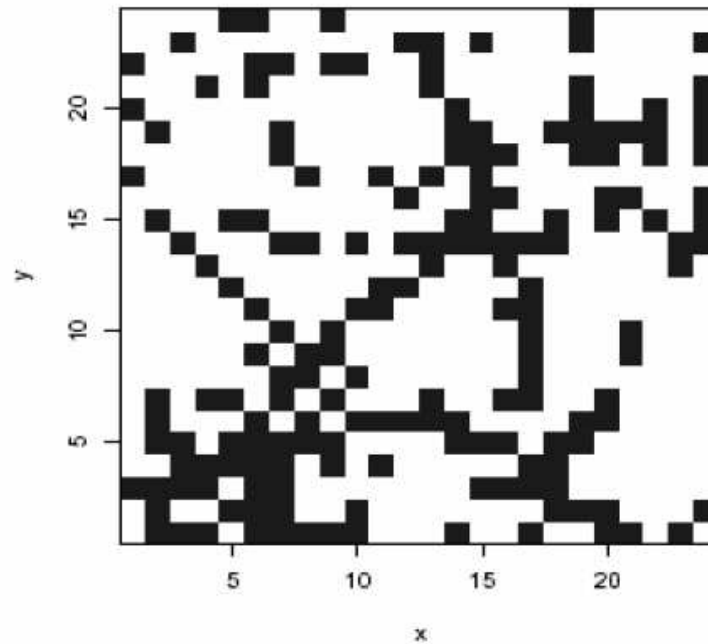
$$\begin{aligned} L_C(x(C) \mid x(S \setminus C)) &= \prod_{i \in C} \pi_i(x_i \mid x(S \setminus C)) \\ &= \prod_{i \in C} \pi_i(x_i \mid x_{\partial i}) \end{aligned}$$

# Conséquences

- Même propriétés que pour un estimateur du MV de variables indépendantes (i.n.i.d.)
- Normalité, test du chi2 d'une sous hypothèse
- Calcul de l'estimation et de la variance d'estimation avec un logiciel dédié aux MLG
- Plusieurs choix d'ensemble de codage possible, mais les estimateurs associés sont dépendants

# Modèle de répartition spatiale d'une espèce végétale (présence / absence)

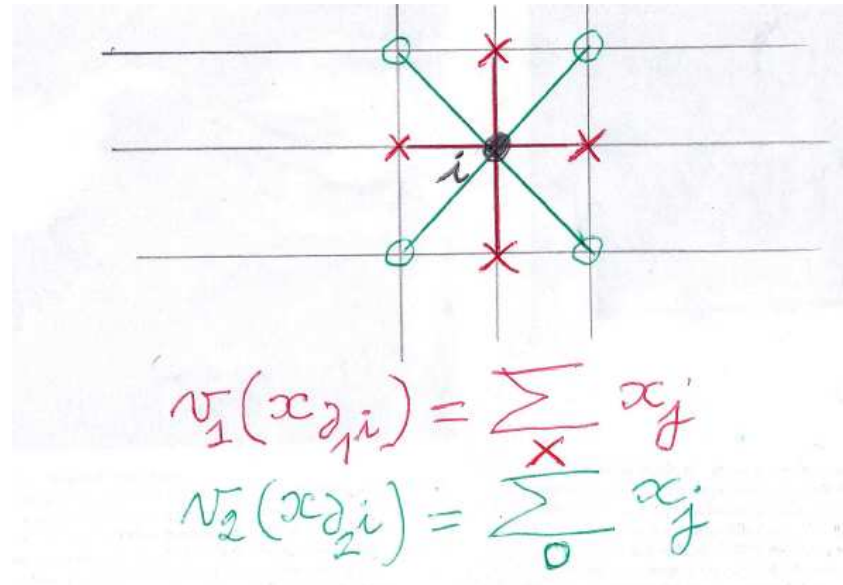
*Variété : la grande laîche*



(a) Présence (■) ou absence (□) de grande laîche.

## 2 modèles auto - logistiques :

(I) aux 4 – ppv ( $v(1)$ ) - (II) aux 8 – ppv ( $v(2)$ )



$$P(X_i = 1 \mid x_j, j \neq i) = \frac{\exp V_i(x)}{1 + \exp V_i(x)} \text{ où}$$

(I) :  $V_i(x) = \beta_0 + \beta_1 v_1(x_{d_1i})$  , où  $v_1(x_{d_1i})$  est la somme des 4 p.p.v.

(II) :  $V_i(x) = \beta_0 + \beta_1 v_1(x_{d_1i}) + \beta_2 v_2(x_{d_2i})$  ,

$v_2(x_{d_2i})$  est la somme des 4 voisins à distance  $\sqrt{2}$  (voisins diagonaux)

# Résultats : estimations MV, PMV et codage

	$\beta_0$	$\beta_1$	$\beta_2$
Codage			
I	-1.486 (0.235)	0.497 (0.133)	-
II	-2.093 (0.442)	0.477 (0.223)	0.391 (0.213)
Pseudo - Vraisemblance			
I	-1.552 (0.172)	0.531 (0.098)	-
II	-1.884 (0.206)	0.433 (0.102)	0.350 (0.106)
Maximum de Vraisemblance			
I	-1.645 (0.196)	0.612 (0.129)	-
II	-1.888 (0.229)	0.441 (0.137)	0.360 (0.152)

---

# Logiciel pour les champs de Markov

- $R$  ne fait pas grand-chose en dehors des champs gaussiens.
- Simulation de textures générales : *AntsInFields* (Winkler), logiciel gratuit téléchargeable.
- Pour l'estimation par PVC ou par codage, on peut utiliser un logiciel dédié aux **modèles linéaires généralisés** : ils donneront la bonne estimation.

**Attention** : si la variance donnée est correcte pour l'estimation par codage, elle ne l'est plus pour l'estimation par PVC.

La variance de l'estimation par PVC doit se calculer analytiquement ou par Monte Carlo à partir de la formule adéquate.

# **(II) – Modèle AR spatial**

## (II) - Modèle AR spatial

- L'espace d'état est  $E = R$  ou  $R^{**p}$
- Le modèle est défini par un ensemble *d'équations* « *spatiales* » *simultanées* (comme en économétrie) en référence à un graphe d'influence
- Le bruit de modèle est blanc (*SAR*) ou coloré (*CAR*)
- Souvent le modèle est supposé gaussien



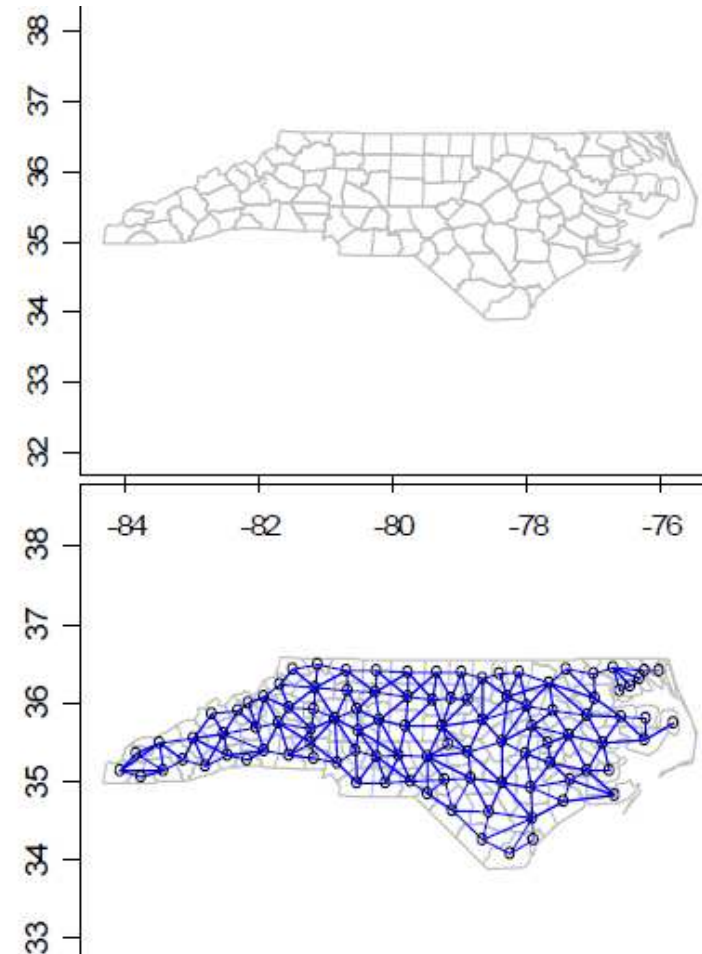
# Ex 1 : la mort subite du nourrisson

(données *sids* de *spdep*, Cressie et al)

- $X(s)$  = taux de *sids* pour le canton  $s$  (*sudden infant deaths in north carolina 74-78*)
- Régions et graphe de voisinage

## Covariables :

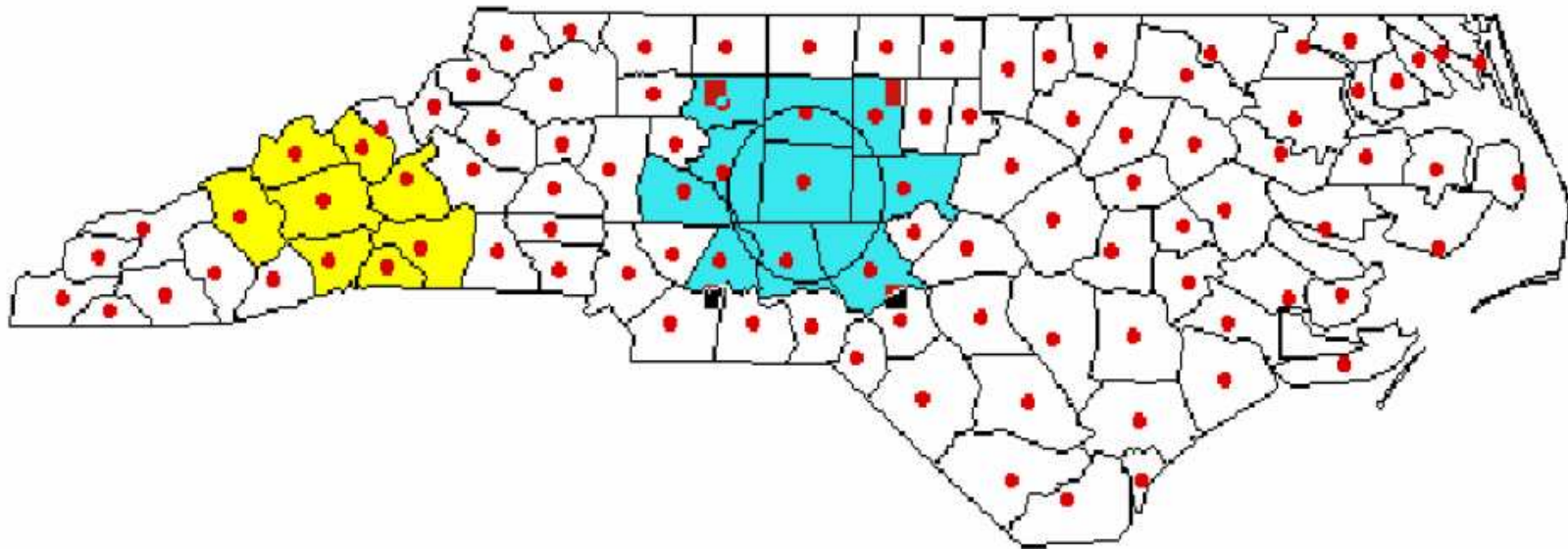
- Nombre de naissances,
- Pourcentage de bébés de chaque communauté,
- Variables socio-économiques, etc ...



## Ex. de système de voisinage

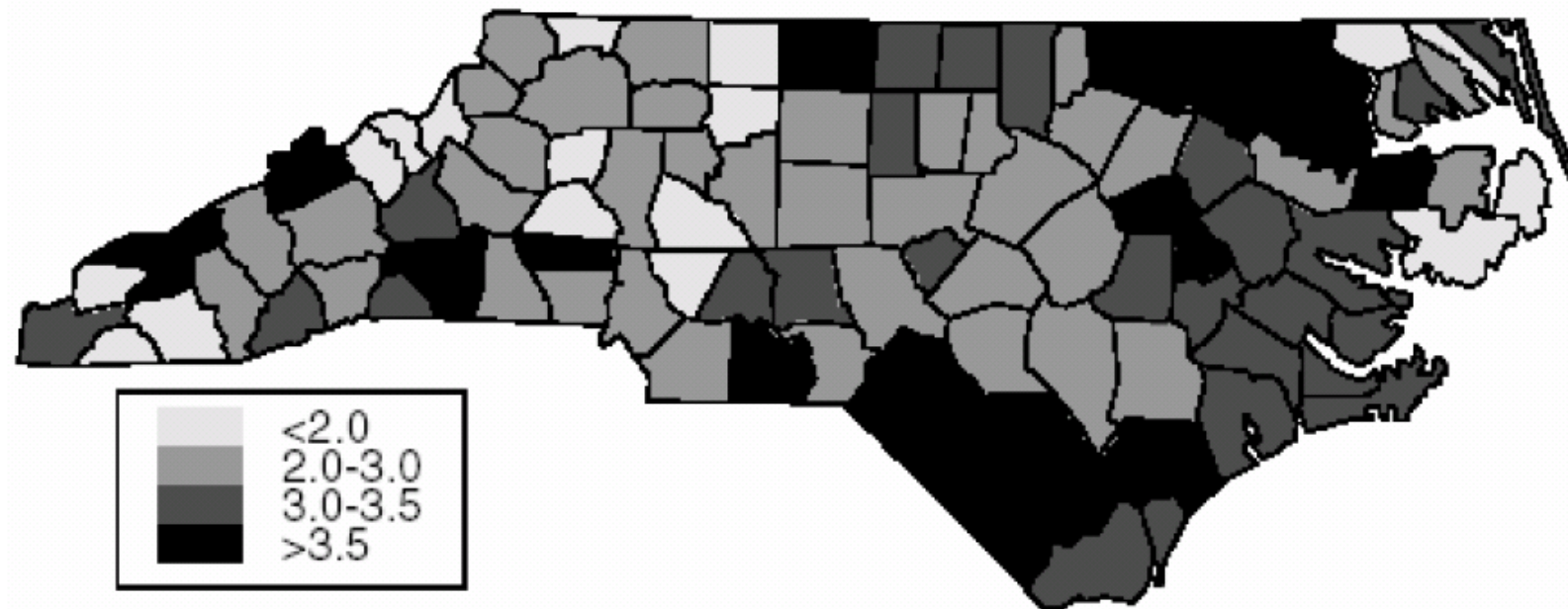
à gauche : cantons avec frontière commune

à droite : centres sont à moins de 30 miles



## Données latticielles : « la mort subite du nourrisson »

Nombre de cas dans 100 comtés de Caroline du nord entre 1974-1978 (données *sids* de *R*; Cressie,1993)



## `nc.sids` dans le package `spdep`

- Voir le descriptif de `nc.sids`
- Représentation des 100 comtés et des 21 variables
- Graphes de voisinage

# Ex. 2 : Taux groupe sanguin A

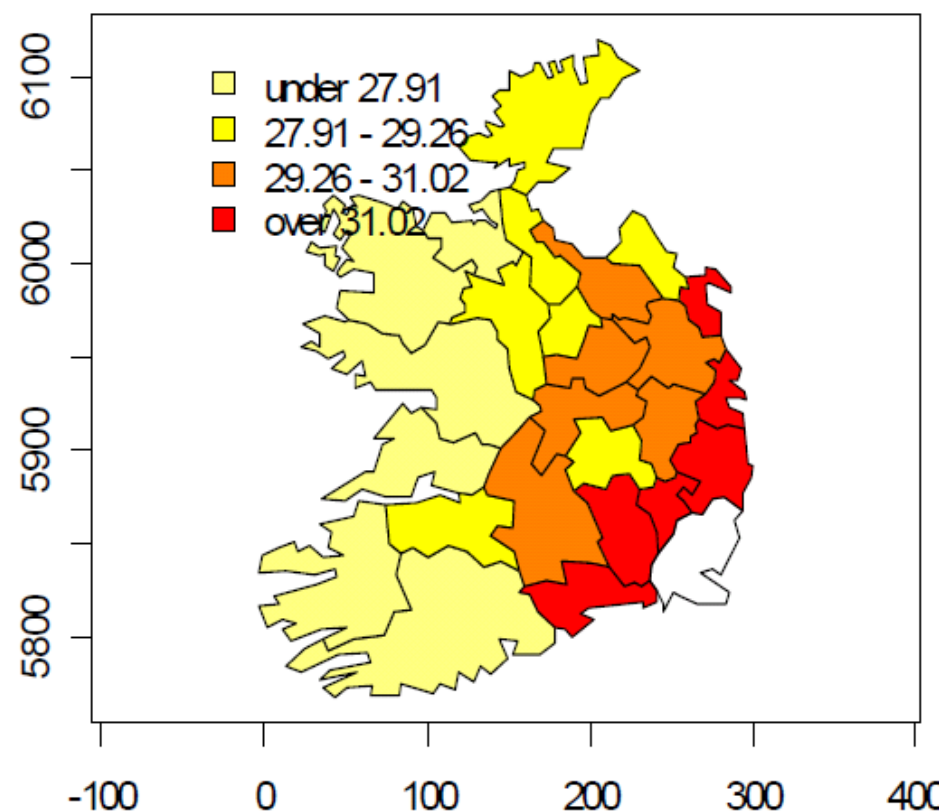
36 comtés de l'Irlande (données *eire*)

- $X(s)$  = % du groupe A dans le comté  $s$

## Covariables :

- Taux d'urbanisation (*towns*);
- Anciennement sous contrôle anglais (*pale*);
- Et d'autres ,.....

Percentage with blood group A in Eire



# Graphe de voisinage, données *eire* « *avoir de la frontière commune* »



# Réseau régulier (agronomie, télédétection) **ou non**

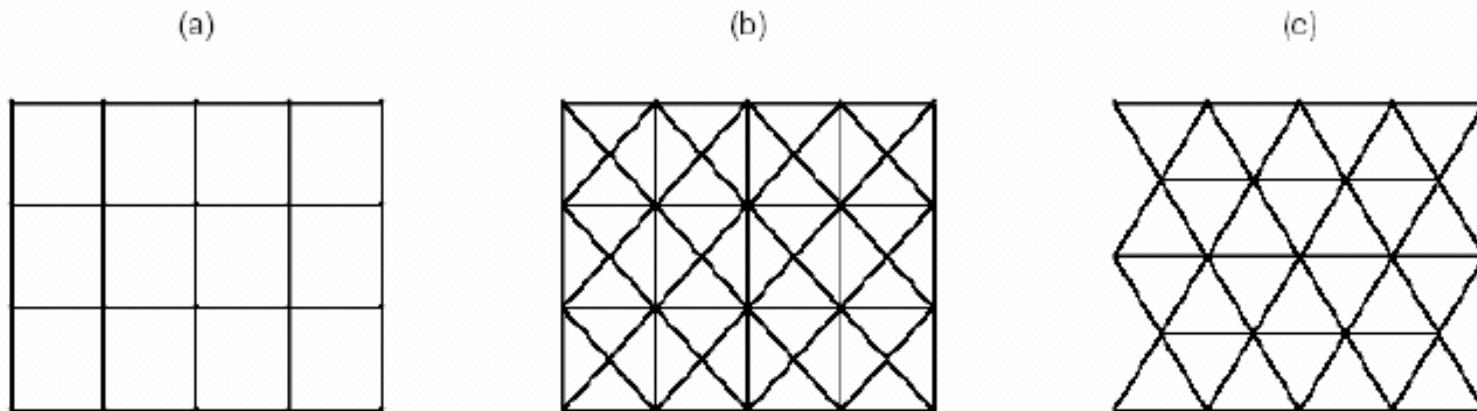
Choix du graphe de voisinage reste à faire

**Exemples de graphes réguliers symétriques :**

(a) Lattice carré : voisinage aux *4 plus proches voisins* (p.p.v.)

(b) Lattice carré : voisinage aux **8 - ppv**

(c) Lattice triangulaire : voisinage aux **6 - ppv**



# SAR général sur $S = \{1, 2, \dots, n\}$

- recentrage de  $X$
- **Graphe** et **poids** d'influence  $W$  + **Bruit**
- $n$  équations simultanées de paramètres  $A$  avec  $\varepsilon$   $BB$
- $X$  existe si  $A$  inversible
- Covariance  $\Sigma$  en termes des paramètres  $A$

$$X_t - \mu_t = \sum_{s \in S: s \neq t} a_{t,s} (X_s - \mu_s) + \varepsilon_t \text{ ou}$$

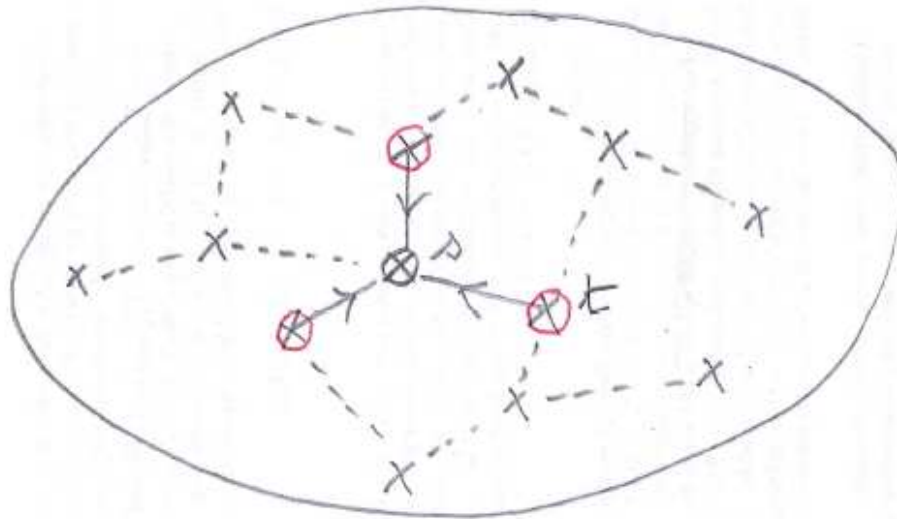
$$A(X - \mu) = \varepsilon, \text{ définie si } A^{-1} \text{ existe.}$$

$$\Sigma^{-1} = \sigma_\varepsilon^{-2} \{ {}^t A A \}$$



# SAR général : graphe et poids

$S =$  ensemble des sites



Modèle SAR

$A =$  matrice  
d'influence sur  
tout  $S$

$\otimes$  influence  $\otimes = \rightarrow$  ,  $a(s,t) =$  poids influence  $t \rightarrow s$

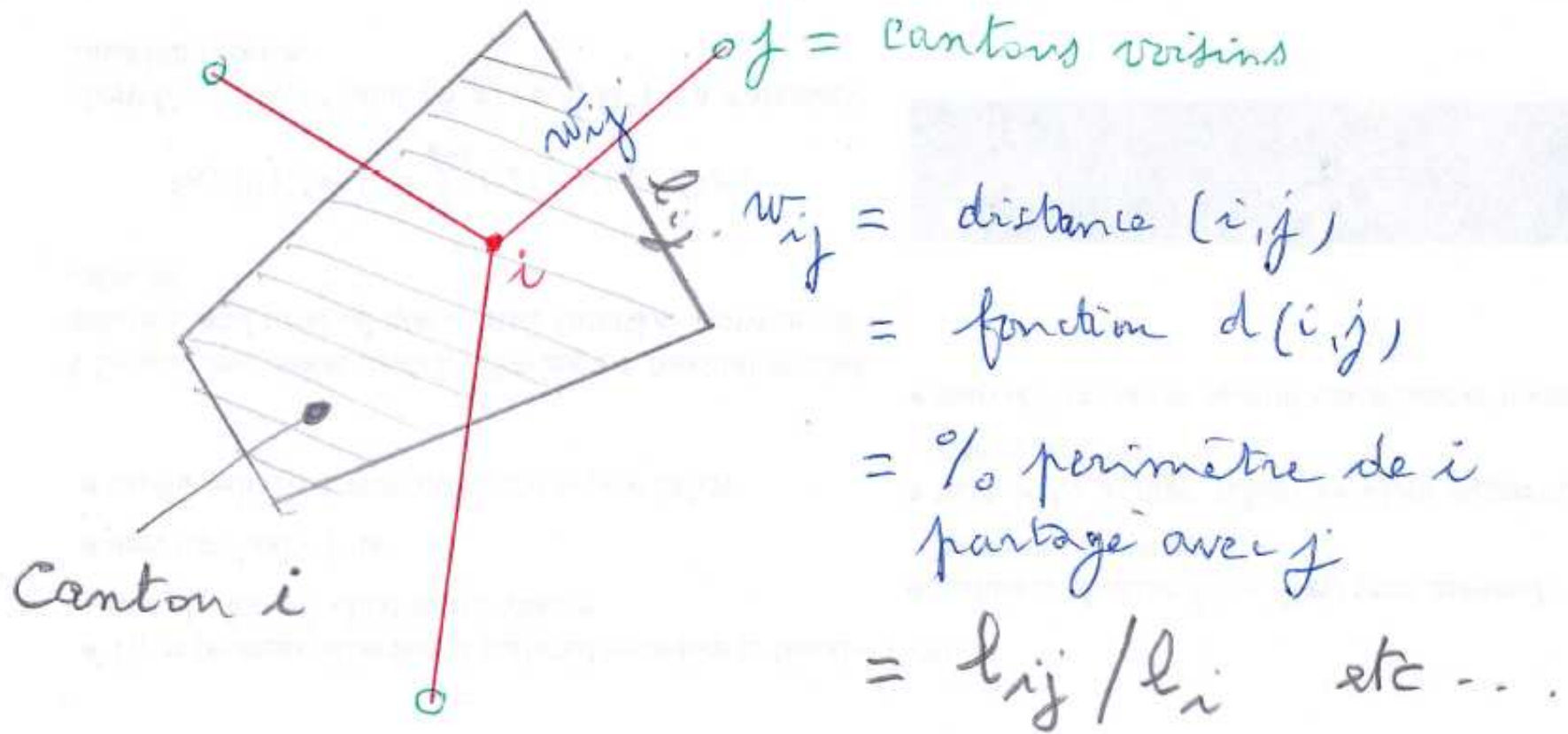
## Spécifier une **SAR**

- $\partial s$  = voisins de  $s$  (symétrique ou non)
- Dépendance « locale » :  $X(s) = F ( X(\partial s), \theta) + \varepsilon(s)$
- $F$  linéaire via  $\theta$  inconnu et  $W$  matrice de poids connus  
 $W = \{W(t,s), t \text{ voisins de } s\}$
- $\varepsilon$  un BB (éventuellement gaussien)

**Exemple** :  $\rho$  = corrélation spatiale,  $(I - \rho W)$  inversible

$$X_t = \rho \sum_{s:s \neq t} w_{t,s} X_s + \varepsilon_t, \text{ ou } X = \rho W X + \varepsilon.$$

## Choix de la matrice de poids de voisinage $W$



# Choix ad hoc de W

- Fonction des distances inter-centres, des (portions) de frontières communes, des réseaux de communications entre deux cellules, etc...
- Paramètres  $\gamma$  et  $\tau$  préalablement calibrés

- $w_{ij} = \|\mathbf{s}_i - \mathbf{s}_j\|^{-\gamma}, \gamma \geq 0$

- $w_{ij} = \exp\{\|\mathbf{s}_i - \mathbf{s}_j\|^{-\gamma}\}$

- $w_{ij} = (l_{ij}/l_i)^\gamma$  where  $l_{ij}$  is the length perimeter of the border of area  $i$

- $w_{ij} = (l_{ij}/l_i)^\tau / \|\mathbf{s}_i - \mathbf{s}_j\|^{-\gamma}.$

# **SAR** stationnaire sur $Z^{**2}$

- bruit blanc :  $\eta_t$  (gaussien ou non)
- Équations avec variables «spatialement retardées»

$$X_t = \sum_{s \in R} a_s X_{t-s} + \eta_t.$$

- $X$  existe si  $P \neq 0$  sur le tore :

$$P(e^{i\lambda}) = 1 - \sum_{s \in R} a_s e^{i^t \lambda s}.$$

- Graphe  $R$  de voisinage orienté (ou non)

# Exemples de SAR

Semi causal Espace x Temps

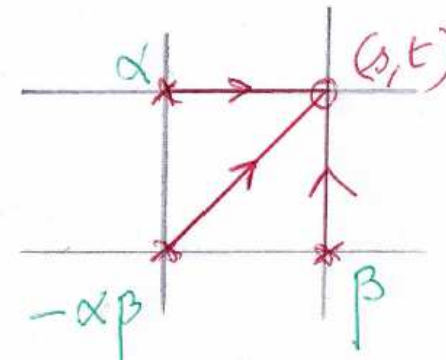
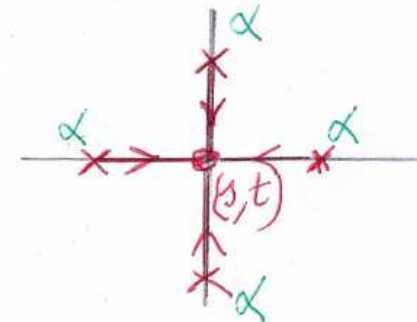
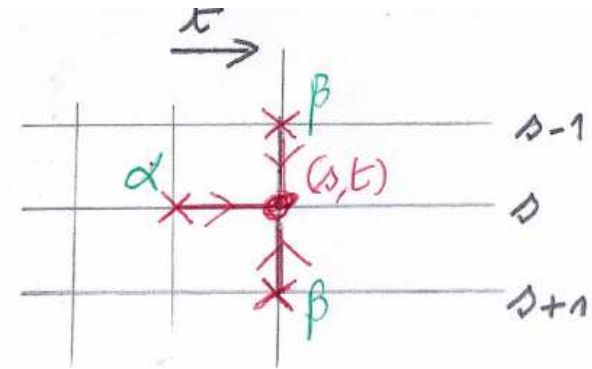
$$X_{st} = \alpha X_{s,t-1} + \beta (X_{s-1,t} + X_{s+1,t}) + \eta_{st}$$

Isotropique aux 4 ppv

$$X_{st} = \alpha (\text{somme 4 ppv}) + \eta_{st}$$

Factorisation (causal) aux 3-ppv

$$X_{st} = \alpha X_{s-1,t} + \beta X_{s,t-1} - \alpha\beta X_{s-1,t-1} + \eta_{st}$$



# Précautions sur un SAR

- Un *SAR bilatéral*  $\neq$  *AR causal* (pour l'ordre lexicographique; cf. exemple dans le polycopié)
- *Sans contrainte*, un *SAR non identifiable*
- L'estimation des *MCO* est *non convergente*
- *Avantage* : *SAR* est *parcimonieux* en paramètres

## **AR** conditionnel général (**CAR**) sur **S**

- Écrire l'espérance conditionnelle de  $X(t)$  sur autres  $X$  :  
Les résidus  $e$  sont corrélés entre eux, décorrés des  $X$  :

$$X_t = \sum_{s \in S: s \neq t} c_{t,s} X_s + e_t, \quad \forall t \in S$$

$$\text{Var}(e_t) = \sigma_t^2 > 0, \quad \text{Cov}(X_t, e_s) = 0 \text{ si } t \neq s.$$

- *Notations* :  $D$  diagonale des résidus,  $\Sigma = \text{Cov}(X)$   
 $C$  paramètres **CAR**. On a l'identité :

$$\Sigma^{-1} = D^{-1}(I - C)$$

→ *contraintes sur les paramètres du CAR*



## Contraintes sur les paramètres d'un **CAR**

$$X_t = \sum_{s \in S: s \neq t} c_{t,s} X_s + e_t,$$

$$(I - C)\Sigma = D.$$

$$c_{t,s}\sigma_s^2 = c_{s,t}\sigma_t^2, \quad \forall t \neq s \in S.$$

# AR Conditionnelle (**CAR**) stationnaire

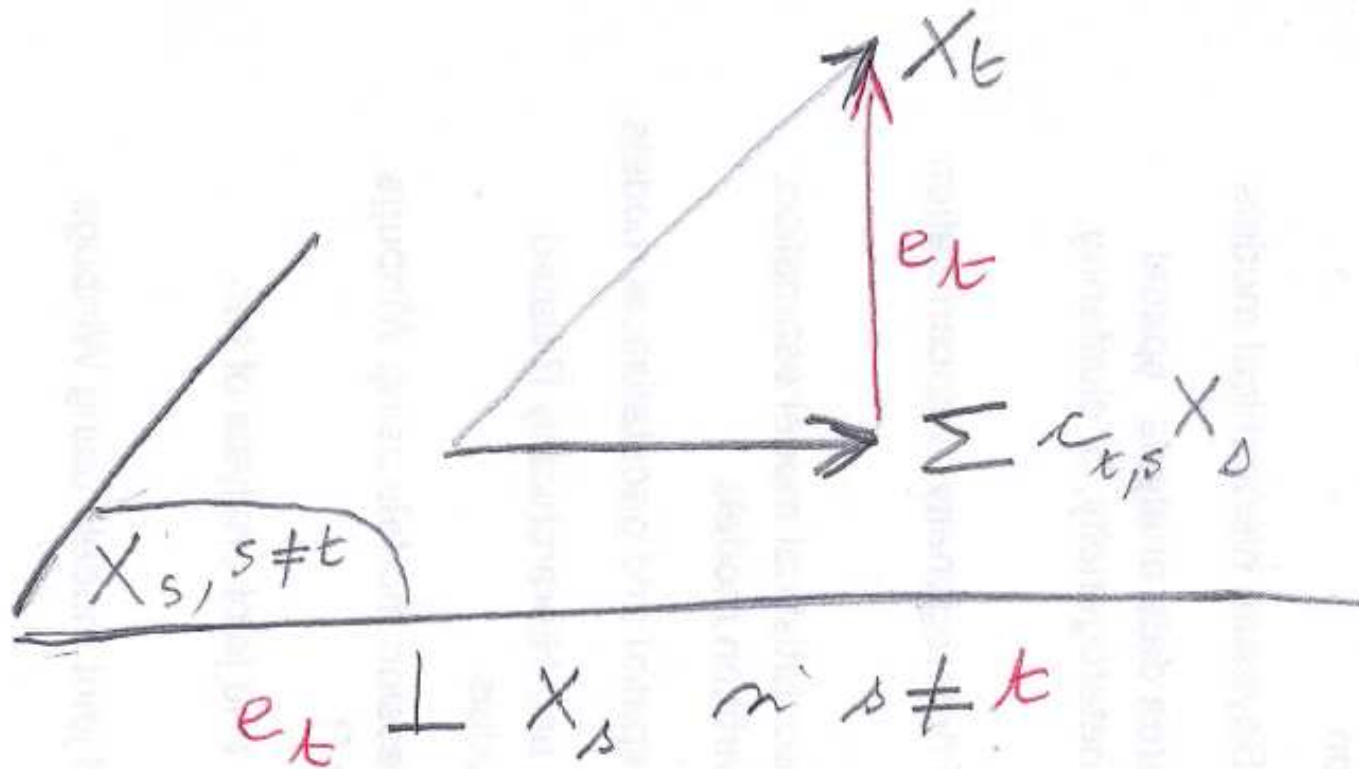
$$X_t = \sum_{s \in L} c_s X_{t-s} + e_t \text{ avec, si } s \in L^+ : c_s = c_{-s}$$

$$\forall s \neq t : Cov(e_t, X_s) = 0 \text{ et } E(e_t) = 0$$

- L'espérance conditionnelle linéaire de  $X(s)$  sur les autres  $X(t)$
- $e(t)$  est décorrélé des  $X(s)$  pour  $s \neq t$
- Le graphe d'un **CAR** symétrique ainsi que les  $c$ .
- Le résidu conditionnel  $e$  est un *bruit coloré* (c-à-d corrélé) :

$$Cov(e_t, e_{t+s}) = \begin{cases} \sigma_e^2 & \text{si } s = 0, \\ -\sigma_e^2 c_s & \text{si } s \in L \end{cases} \text{ et } Cov(e_t, e_{t+s}) = 0 \text{ sinon}$$

# CAR : espérance conditionnelle linéaire et résidu



## ***SAR* ou *CAR* : pour résumer**

- Un ***SAR*** est spécifié par  $n$  équations simultanées à résidus ***BB***
- Un ***CAR*** est spécifié par ses « espérances conditionnelles linéaires ». Le bruit résiduel est coloré

# CAR ou SAR ?

- Tout **SAR** est un **CAR**
- Si **S fini**, **CAR**  $\equiv$  **SAR** ( $\neq$  sinon)
- Écriture **CAR** *intrinsèque*, celle d'un **SAR** *non*
- Estimation **MCO**
  - d'un CAR convergente*
  - d'un SAR non*
- **SAR** : plus *parcimonieux* en nombre de paramètres
- **CAR** : *contraintes sur les paramètres*

# Correspondances des graphes :

**$R$**  d'un **SAR**  $\leftarrow$  et  $\rightarrow$   **$G$**  d'un **CAR**

- **$R$**  graphe du **SAR** : *orienté*
- **$G$**  du **CAR** : *non orienté*, le « double » de  **$R$**  :

Le graphe  $\mathcal{G}$  de la représentation markovienne *CAR* de  $X$  est

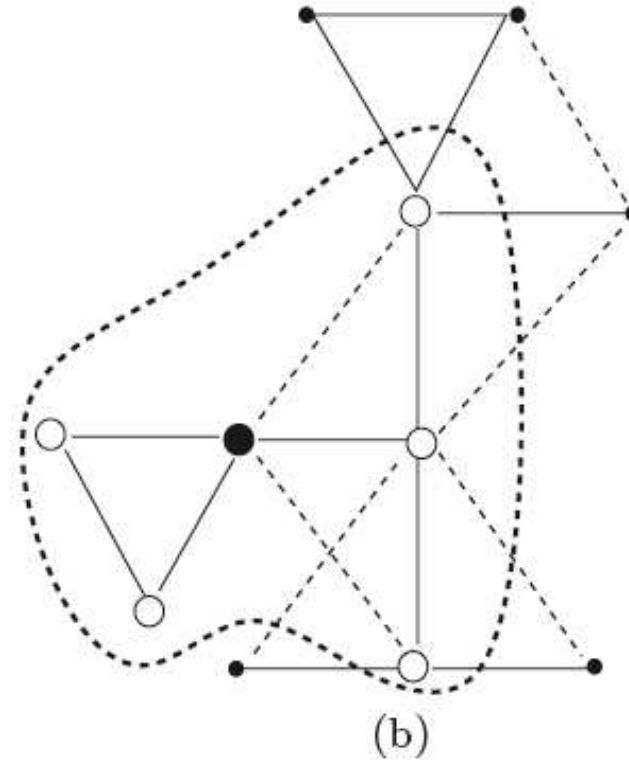
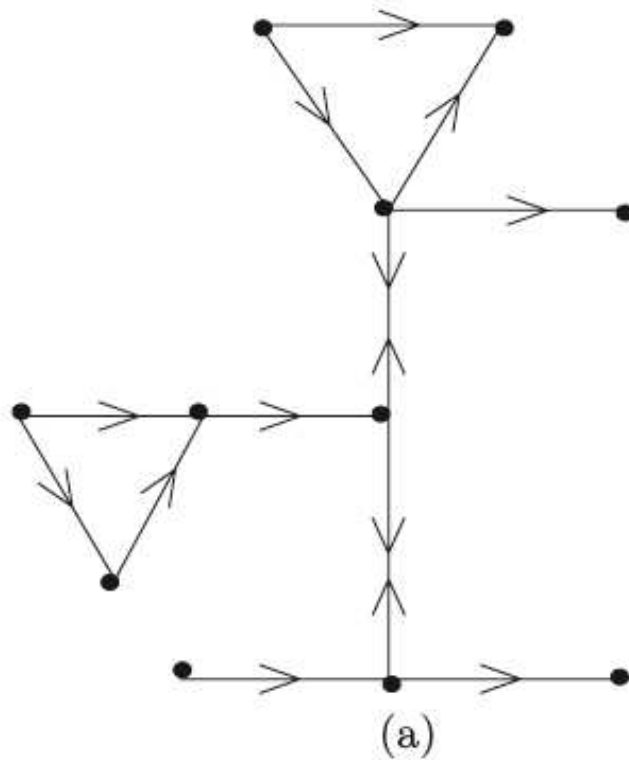
$$\langle t, s \rangle_{\mathcal{G}} \iff \begin{cases} \text{soit } \langle t, s \rangle_{\mathcal{R}}, \\ \text{soit } \langle s, t \rangle_{\mathcal{R}} \\ \text{soit } \exists l \in S \text{ t.q. } \langle l, t \rangle_{\mathcal{R}} \text{ et } \langle l, s \rangle_{\mathcal{R}} \end{cases}$$

# Exemple de correspondance

(a) SAR et  $R$

(b) CAR associé et  $G$

en pointillé : voisinage CAR de ●

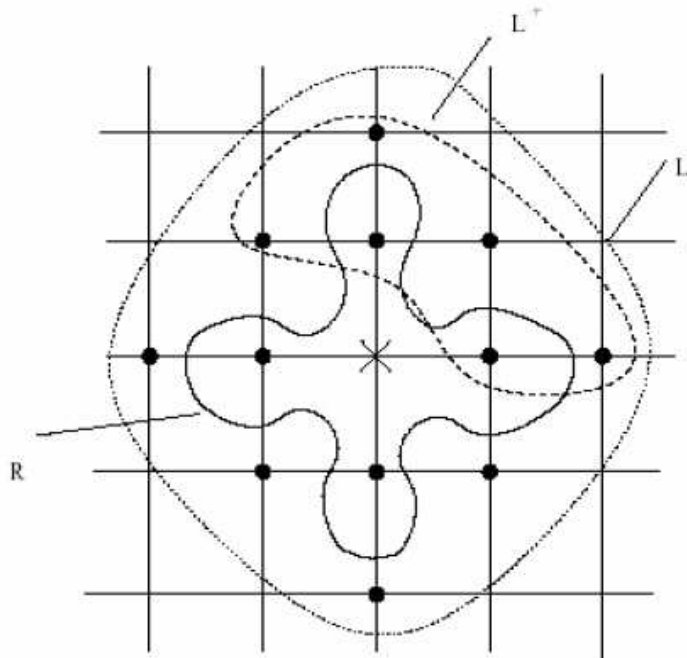


# **SAR** aux 4 ppv et **CAR** associé

- **SAR** aux 4-ppv avec  $R = \{(1,0),(-1,0),(0,1),(0,-1)\}$

$$X_{s,t} = a(X_{s-1,t} + X_{s+1,t}) + b(X_{s,t-1} + X_{s,t+1}) + \varepsilon_{s,t}$$

- **CAR** aux 12-ppv (cf. poly. pour les coeff.  $c(s)$ ) avec  
 $L+ = \{(1,0),(2,0),(1,1),(0,1),(0,2),(1,-1)\}$



avec un **gain de prédiction**

$$(1 + 2a^2 + 2b^2)^{-1}$$

$$K^{**2} =$$

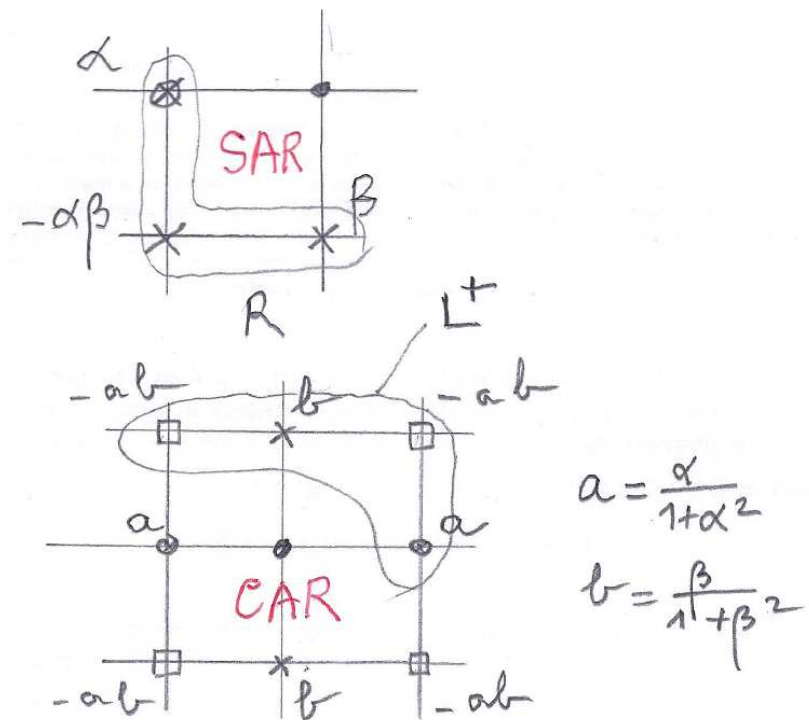


# SAR factorisant aux 3 – ppv et CAR aux 8-ppv associé

$R$  = support du SAR et  $L$  = support du CAR (8 voisins)

$$X_{s,t} = \alpha X_{s-1,t} + \beta X_{s,t-1} - \alpha\beta X_{s-1,t-1} + \varepsilon_{s,t}$$

$$\kappa^2 = (1 + \alpha^2)^{-1} (1 + \beta^2)^{-1}$$



# SARX avec exogènes

- $X$  endogène,  $Z$  matrice des exogènes
- Matrice de retard  $W$  sur endogène et exogène
- 3 types de variables expliquent  $X(t)$  :
  - (1) endogène retardée  $WX$ ,
  - (2) exogène  $Z$  et
  - (3) exogène retardée  $WZ$ .
- On obtient facilement  $E(X)$  et  $Cov(X)$

$$X = \rho W X + Z \beta + W Z \gamma + \varepsilon, \quad \rho \in \mathbb{R}, \quad \beta \text{ et } \gamma \in \mathbb{R}^p$$

# Deux modèles avec exogènes

- Modèle de *Durbin spatial* ( $X-Z\beta \sim SAR(\rho, W)$ )

$$(I - \rho W)X = (I - \rho W)Z\beta + \varepsilon$$

- Modèle à *décalage spatial* ( $\gamma = 0$ )

$$X = \rho W X + Z\beta + \varepsilon$$

# Auto - corrélation de Moran

- $X$  sur  $S=\{1,2,\dots,n\}$  centré (adaptation si modèle de régression sur  $E(X)$ )
- $W(i,j)$  matrice de poids  $i \rightarrow j, i \neq j$  ( $W(i,i)=0$ ) donnée
- $W$  – **auto-corrélation** de Moran :

$$I_M = \frac{n \sum_{i,j} w_{i,j} (X_i - \bar{X})(X_j - \bar{X})}{s_0 \times \sum_i (X_i - \bar{X})^2}$$

$$s_0 = \sum_{i,j} w_{i,j} = \text{somme des poids } W$$

# Test de non corrélation spatiale ( $H(0)$ )

- $I(M)$  petit de variance identifiée

$$\text{Sous } (H_0) \quad : \quad E(I_M) = o(1) \text{ et } \text{Var}(I_M) \simeq \frac{s_1}{s_0^2}$$

$$\text{où } s_1 = \sum_{i,j} (w_{i,j}^2 + w_{i,j}w_{j,i}).$$

- Si  $X$  gaussien, plus précisions sur  $E()$  et  $\text{Var}()$ .

- En général, sous ( $H_0$ ), normalité :  $\frac{s_0 n}{\sqrt{s_1 n}} I_n^M \xrightarrow{\text{loi}} \mathcal{N}(0, 1)$ .

# Indice de Geary

Mesure la dépendance spatiale comme le fait un variogramme :

$I(G)$  est petit si les valeurs voisines sont proches

$$I_n^G = \frac{(n-1) \sum_{i,j \in D_n} w_{ij} (X_i - X_j)^2}{2s_{0n} \sum_{i \in D_n} (X_i - \bar{X})^2}.$$

# ***Loi permutacionnelle*** d'une statistique $I(X)$

- $X=\{X(i),i=1,n\}$  et  $I(X)$  une statistique réelle
- Distribution empirique des  $\{I(X(\sigma)), \sigma \text{ permutation}\}$
- Intervalle de confiance associé à la statistique d'ordre
- Mais  $n!$  est trop grand  $\rightarrow$  le faire pour  $m$  permutations choisies au hasard

$$\mathcal{I} = [I_{(\alpha\sigma)}, I_{((1-\alpha)\sigma)}]$$

# Application : test de permutation de ( $H_0$ ) : indépendance des $\{X(i), i=1, n\}$

- Choix de  $I(X)$ , l'indice de Moran : sous ( $H_0$ ),

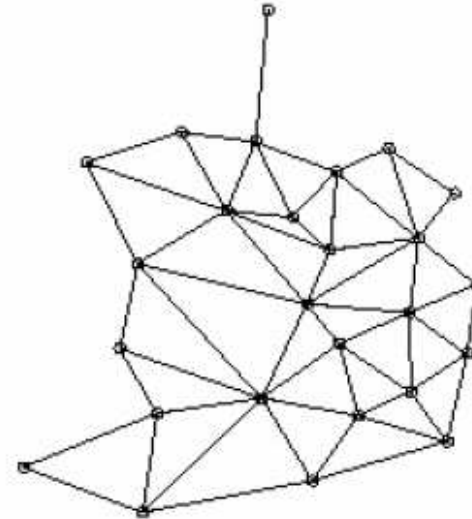
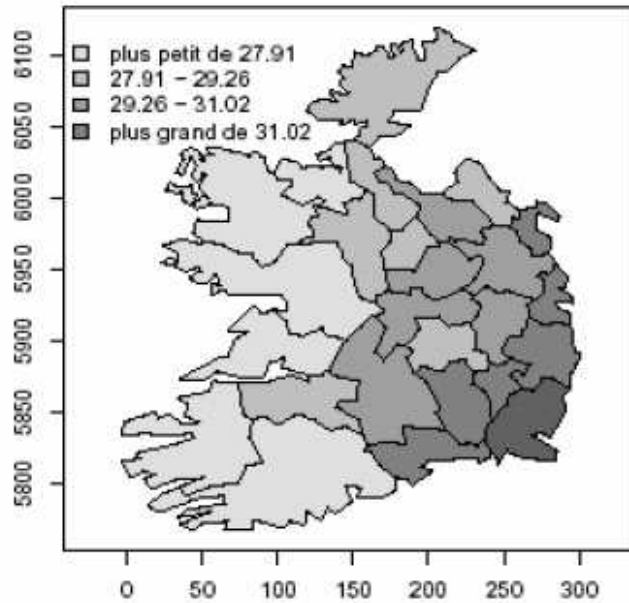
$$(X_i, i = 1, \dots, n) \sim (X_{\sigma(i)}, i = 1, \dots, n).$$

- Calcul des  $I(\sigma, X)$  pour  $m$  permutations au hasard (i.e.  $m=1000$ ) et de l'intervalle de confiance empirique  $IC(1-\alpha)$
- Calcul de  $I(x)$  pour l'observation  $x$
- Si  $I(x)$  n'est pas dans  $IC(1-\alpha)$ , rejet de ( $H_0$ )
  
- **Avantage** : non – asymptotique, libre du modèle sur  $X$
- **Inconvénient** : le niveau est approximatif



# Données *eire* : groupe sanguin A

- 1 -  $G$  = graphe de voisinage de contiguïté des 26 contés
- 2 -  $w(i,j) = 1/(\text{nb voisins de } i)$  si  $j$  est voisin de  $i$



# Indice de Moran (Geary) + indices réduits $t(a)$ probabilités $p(a)$ de dépassement

- 1 –  $I$  = index,  $t(a)$  index réduit, asymptotique gaussienne
- 2 –  $p(a)$  = proba de dépassement de  $t(a)$
- 3 –  $pmc(a)$  pour le test de permutation MC ( $m = 1000$  permutations)

	Index	$t^a$	$p^a$	$p_{MC}^a$
Moran	0.554	4.663	0	0.001
Geary	0.380	-4.547	0	0.001

# Le package `spdep` :

## étude sur l'exemple des données `eire`

- Données i.e. : `eire` groupe sanguin en Irlande
- Tester bloc par bloc l'exemple de traitement de ces données
  - 1 - représentation des données (valeur, graphe de voisinage)
  - 2 - indice de Moran et test de non corrélation spatiale
  - 3 - régression sur `town` et `pale`, analyse des résidus (indice de Moran, SAR sur les résidus)
- Autres données sur la consommation intérieure

## Quelques programmes de `spdep`

- `moran`, `moran.test` et `moran.mc`
  - `lagsarlm` : estimation du MV d'un SAR avec covariables
$$y = \rho W y + X \beta + e$$
  - `knearneigh` : matrice des k-ppv pour un choix de distance
  - `lm.morantest` : test de Moran pour l'auto-corrélation spatiale des résidus d'un modèle linéaire
  - `lm.morantest.exact` : test exact de non corrélation
  - `sp.correlogram` : corrélogramme spatial pour l'indice de moran
- et d'autres programmes ....

# Estimation d'une régression spatiale

$$X = Z\delta + \varepsilon \text{ où } \text{Cov}(\varepsilon) = \Sigma$$

- Estimation MCO de  $\delta$  :  $\tilde{\delta} = ({}^t Z Z)^{-1} {}^t Z X$ .
  - Sous bonnes conditions, consistance des MCO
- MCO bonne estimation initiale dans une procédure itérative type MCQG

# Moindres Carrés Généralisés (MCG)

- Si  $\Sigma = \text{cov}(\varepsilon)$  est connue, le BLUE vaut

$$\hat{\delta}_{MCG} = ({}^t Z \Sigma^{-1} Z)^{-1} {}^t Z \Sigma^{-1} X;$$

$$\text{Var}(\hat{\delta}_{MCG}) = ({}^t Z \Sigma^{-1} Z)^{-1}.$$

- Si  $X$  gaussien, c'est l'EMV, efficace
- En général  $\Sigma$  inconnue  $\rightarrow$  MC Quasi G (MCQG)

## MCQG : $\Sigma = \Sigma(\theta)$ , $\theta$ inconnu

1. estimer  $\delta$  par MCO :  $\tilde{\delta} = ({}^tZZ)^{-1}{}^tZX$ .
2. calculer les résidus des MCO :  $\tilde{\varepsilon} = X - Z\tilde{\delta}$
3. sur la base de  $\tilde{\varepsilon}$ , estimer  $\tilde{\theta}$  (pour  $\Sigma(\theta)$  ou  $2\gamma(\theta)$ ) par MC.
4. estimer  $\Sigma$  par  $\tilde{\Sigma} = \Sigma(\tilde{\theta})$  puis  $\delta$  par  $MCG(\tilde{\Sigma})$ . Itérer.

# Régression Gaussienne : MV

- Régression à covariance non sphérique :

$$X \sim \mathcal{N}_n(Z\delta, \Sigma(\theta))$$

- Mardia-Marshall donnent le comportement limite de l'EMV de  $(\theta, \vartheta)$  (cf. poly)
- Log-vraisemblance est explicite :

$$2l(\delta, \theta) = \log |\Sigma(\theta)| + {}^t(X - Z\delta)\Sigma^{-1}(\theta)(X - Z\delta)$$



# Données *eire* : 2 modèles de régression avec 2 covariables

- *towns* (densité urbaine) et
- *pale* (binaire, 1 si colonisation anglaise, 0 sinon)

(R1) : *cste*, *towns* et *pale* + résidus i.i.d.

(R2) : *cste*, *pale* + résidus SAR aux ppv :  $r = (\rho W) r + e$

	Modèles	
Coefficient	(a)	(b)
Intercept	27.573 (0.545)	28.232 (1.066)
towns	-0.360 (2.967)	-
pale	4.342 (1.085)	2.434 (0.764)
$\rho$	-	0.684 (0.148)

## *Spdep*, le package de *R* pour les SAR, SARX, Indice de Moran, test Monte Carlo, ...

- *spdep* : spatial dependence, weighting schemes, statistics, models.

*anova.sarlm* : compare des SAR

*deviance.sarlm*

*errorsarlm* : MV du modèle de Durbin spatial,  $Y = Xb + u$  ou  $u = rWu + e$

Données : *sids*, *eire*, *getisord* (télédétection)

*moran*, *moran.mc* (test de permutation), *moran.test* (gaussien)

*plot.spcor* : corrélogramme spatial, etc.....

- **Autres packages :**

- *nlme* : linear and non linear (mixed effect) models

*gls* (et *gnls*) : GLS pour modèle linéaire (non linéaire), avec ou sans effet aléatoire.

*logLik.g(n)ls* : MV pour ces modèles

- ***RandomFields*** : simulation et analyse des RF.

---

# Processus Ponctuels Spatiaux

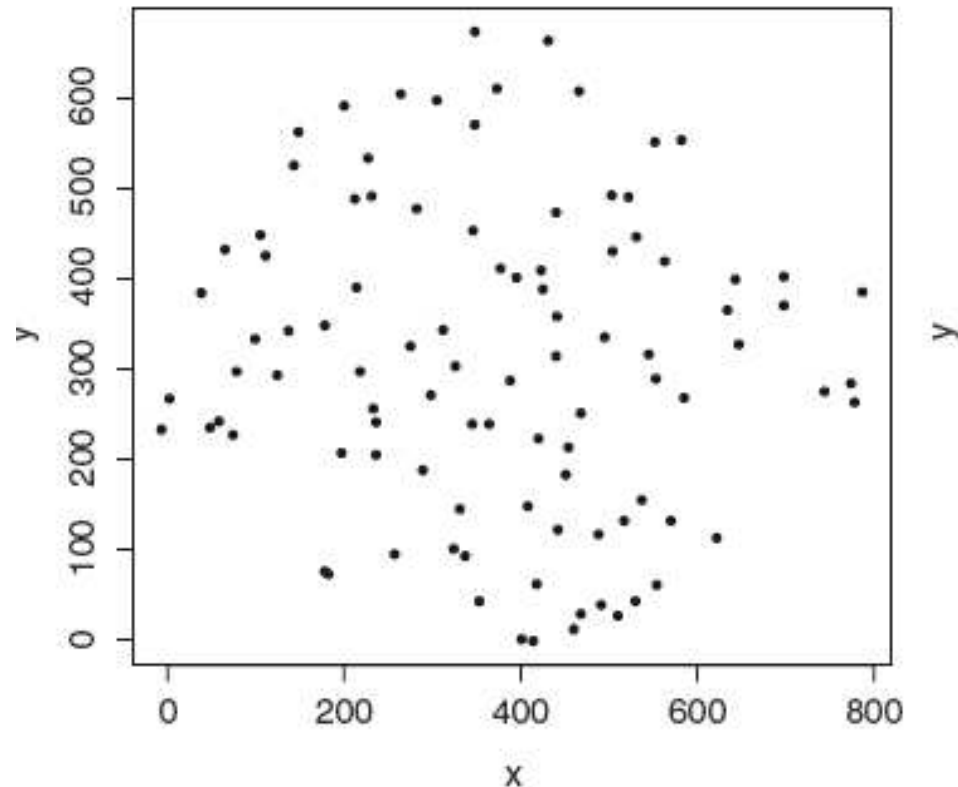
---

- Ici, c'est la **répartition spatiale des points** où ont lieu les observations qui **est aléatoire**.
- Hypothèse de base a tester : la répartition est homogène et au hasard (CSR pour *Complete Spatial Randomness*), celle d'un *PP de Poisson*.
- D'autres répartitions sont plus régulières (noyau dur), d'autres moins (agrégats).

# Répartition de 97 fourmilières

(données *ants* du package *spatstat*)

**Question :** la répartition s'est elle faite au hasard ?

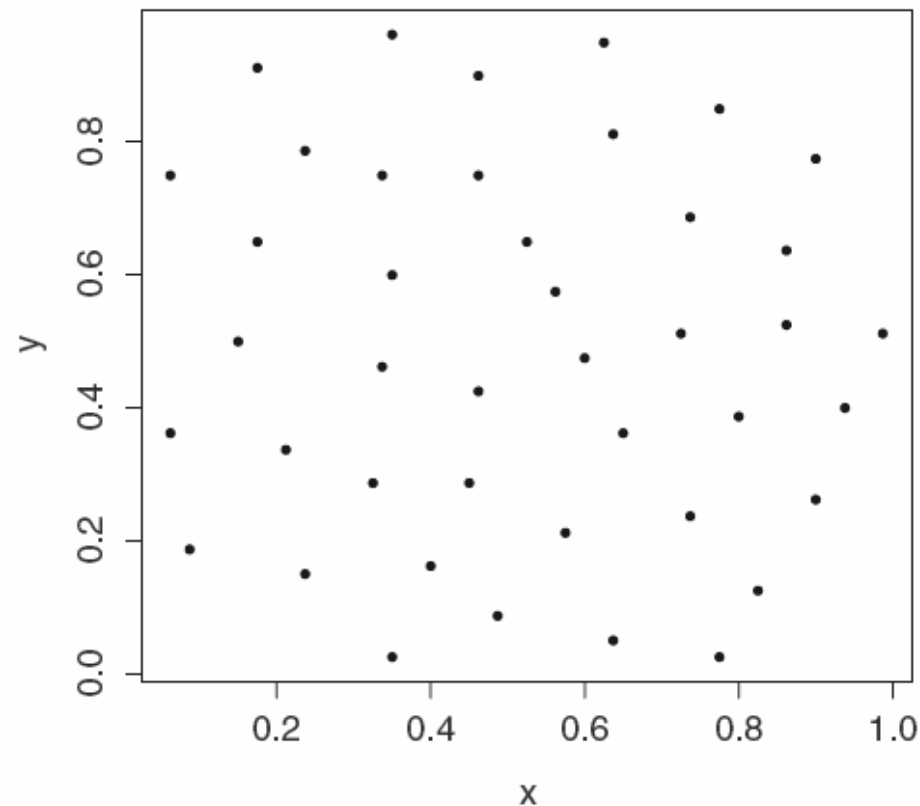


(a)

# 42 centres de cellules d'une coupe histologique

(données *cells* de *spatstat*)

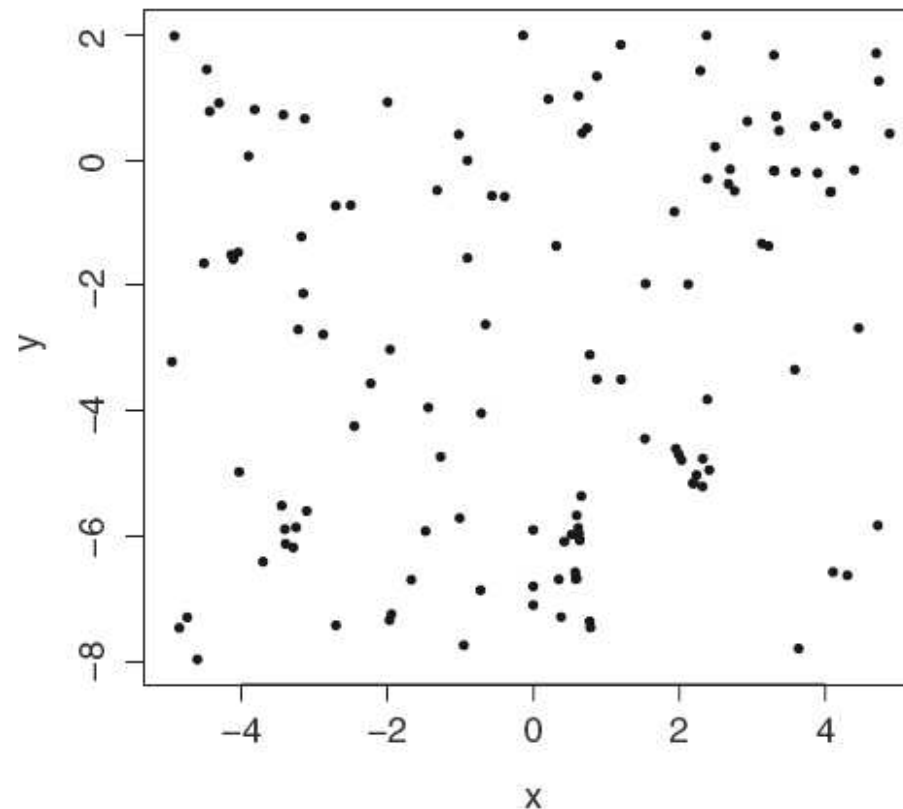
- 1 - La répartition est – elle au hasard ?
- 2 - Sinon (+ de régularité), quel modèle proposer ?



# 126 pins d'une forêt finlandaise

(données *finpines* de *spatstat*)

- 1 - La répartition est elle au hasard ?
- 2 - Sinon (des agglomérats ?), quel modèle proposer ?



# Modèle de Processus Ponctuel $X$ (PP)

- *Configuration*  $x$  : ensemble fini de points de la fenêtre d'observation  $S$
- *Configuration à  $n$  points* :  $x = \{x(1), x(2), \dots, x(n)\}$
- $E(n)$  = espace des configurations à  $n$  points
- $E = \bigcup E(n)$  : l'espace exponentiel de toutes les configurations, réunion des  $E(n)$
- $N(A)$  : le nombre de points de  $X$  dans  $A$
- *Loi de  $X$*  : loi jointe de toutes les variables de comptage  $N(A)$ ,  $A$  partie de  $S$

# PP de Poisson homogène d'intensité $\lambda$

***PPP( $\lambda$ )*** : répartition spatiale *homogène et au hasard*

(1)  $N(A)$  suit une loi de Poisson de paramètre  $\lambda |A|$

(2) La répartition sur  $A$  est uniforme

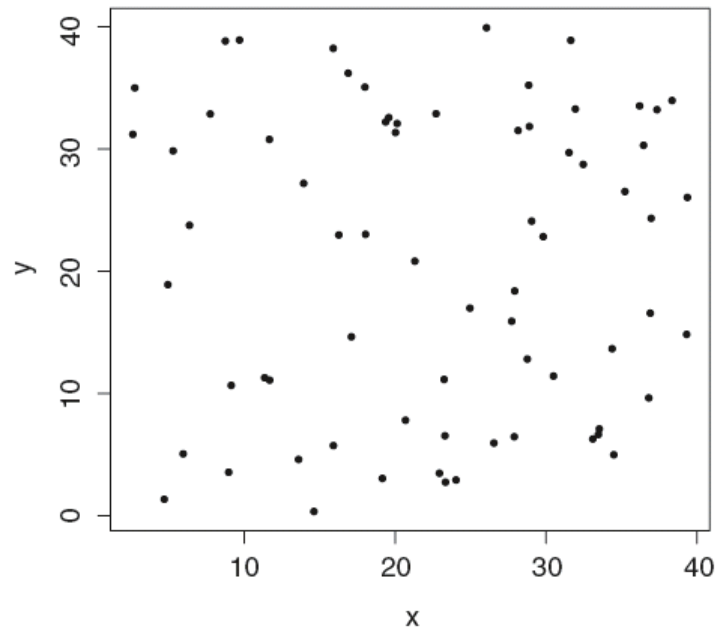
(1-2)  $\equiv$  (1-2\*) où

(2\*) : si  $A$  et  $B$  sont disjoints,  $N(A)$  et  $N(B)$  sont indépendants

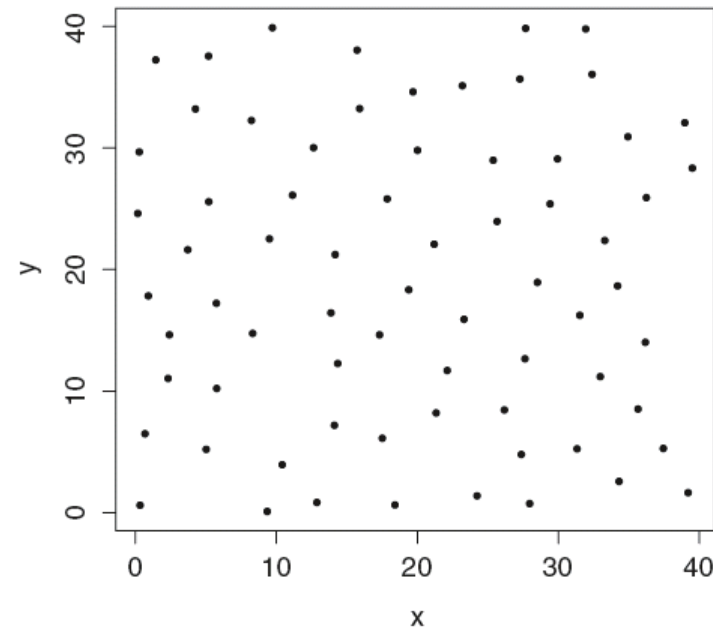


# Deux répartitions spatiales homogènes à 70 points

- (a) une répartition de *Poisson* homogène
- (b) une répartition « à *r* - noyau dur» homogène (couples de points à distance  $< r = 3.5$  sont interdits)



(a)



(b)

# PPP inhomogène d'intensité $\lambda(\bullet)$

Soit  $\lambda(\bullet)$  une *mesure* sur la fenêtre d'observation  $S$

$X$  est un  $PPP(\lambda(\bullet))$  si :

- (1)  $N(A)$  suit une loi de Poisson de paramètre  $\lambda(A)$
- (2) si  $A$  et  $B$  sont disjoints,  $N(A)$  et  $N(B)$  sont indépendants

# Simulation d'un $PPP(\lambda(\bullet))$

Supposons que pour tout  $x$  :  $\lambda(x) \leq c < \infty$

La méthode par **effacement de points** est la méthode de simulation par rejet suivante :

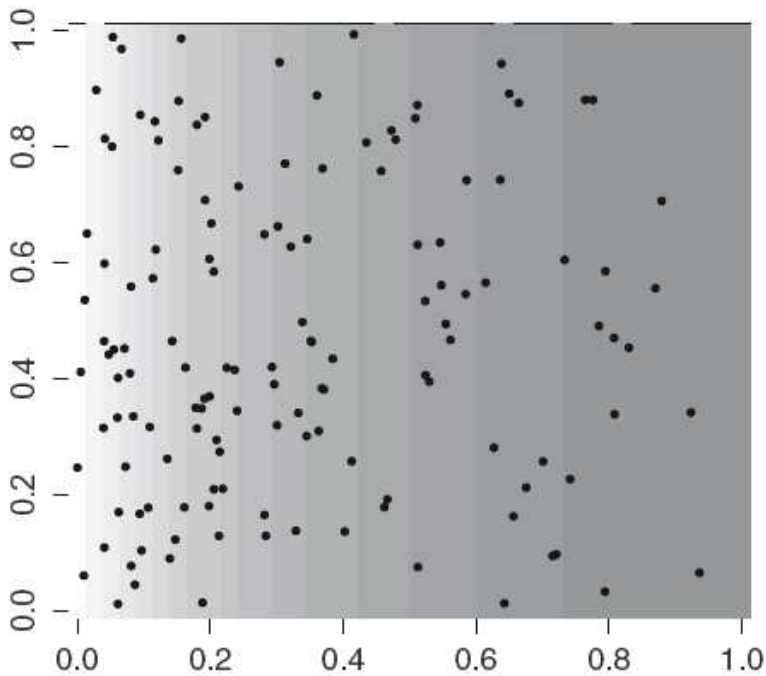
1. Simuler  $x^*$ , un PPP homogène d'intensité  $c$ ;
2. Effacer indépendamment un  $x(i)$  de  $x^*$  avec la probabilité  $p(x(i)) = \{1 - \lambda(x(i))/c\}$ .

# Simulation de 2 PPP inhomogènes

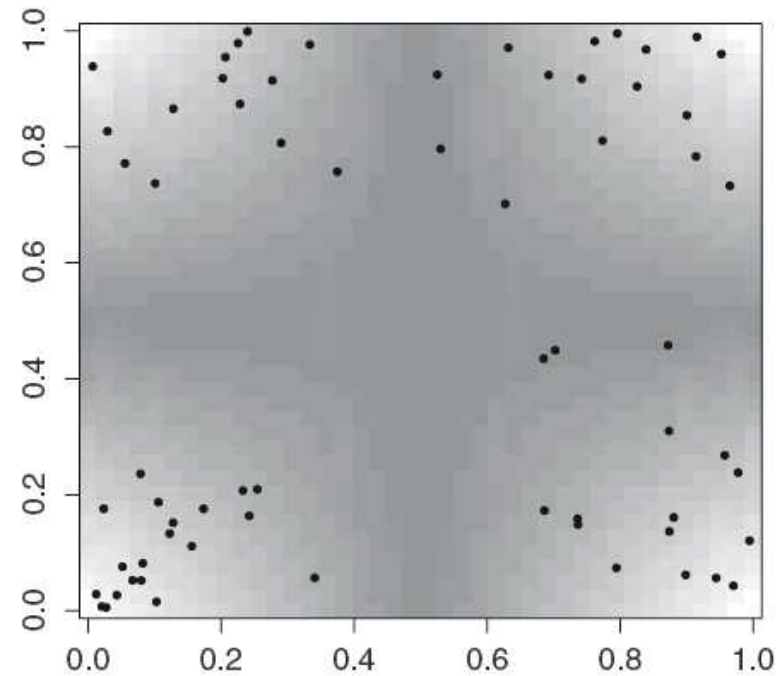
( $\lambda$  plus grand  $\rightarrow$  fond plus clair)

Deux intensités inhomogènes

(a) :  $\lambda(x, y) = 400 e^{-3x}$  et (b)  $\lambda(x, y) = 800 \times |0.5 - x| \times |0.5 - y|$



(a)



(b)

# Répartition plus régulière : modèle à noyau dur (ou hardcore)

La règle : *interdire les points trop proches*

- **Exemples :**
  - répartition spatiale d'animaux (compétition)
  - boulangeries dans une ville
  - centres de cellules
  - arbres dans une forêt (??)
  - en physique, centres d'«atomes impénétrables»
- Ces modèles vont être défini par leur *densité de Gibbs*

# Répartition moins régulière : formation d'agrégats (clusters)

Exemple : le PP de Neymann – Scott

1. Un processus  $P$  « parent » :  $PPP$  homogène  $\lambda$
2. Chaque parents  $P(i)$  engendre des enfants en *nombre*  $N$  et en *positions*  $D$  centrées autour de  $P(i)$ , *aléatoires*,  $N$  et  $D$  indépendantes

*Paramètres* :  $\lambda$ , les lois  $N$  et  $D$

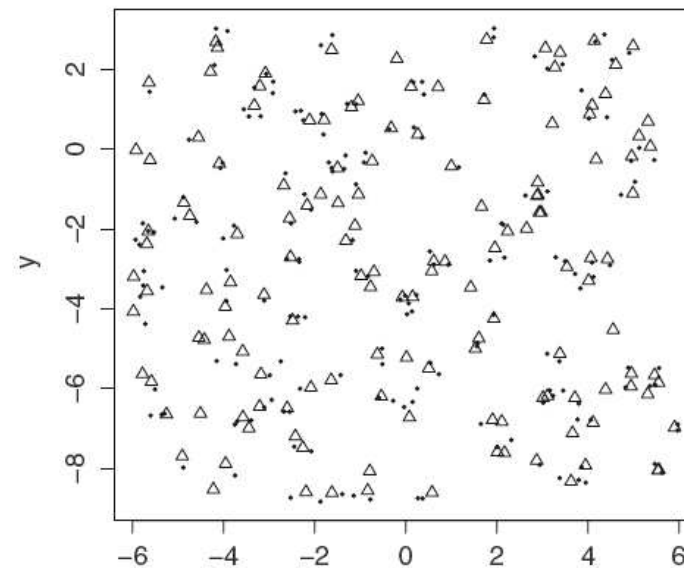
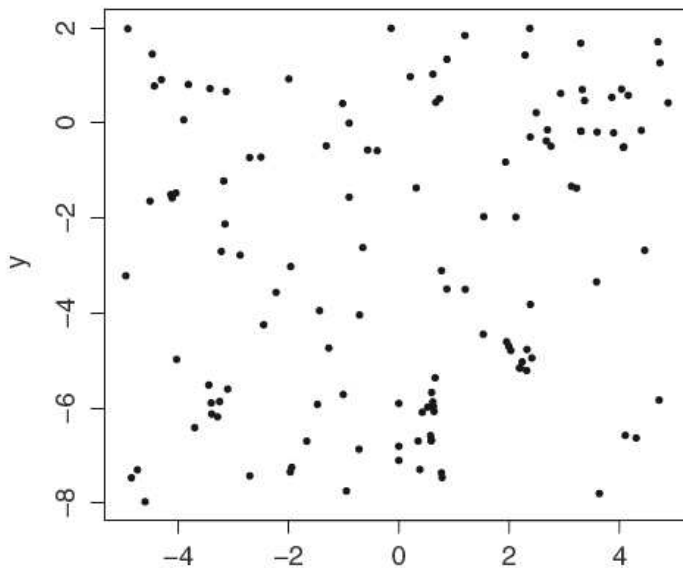
# Simulation de PP spatiaux avec `statspat`

- `owin` : crée la fenêtre d'observation (si nécessaire)
- `runifpoint` : n points uniformes.
- `runifpoint3` : idem mais dans  $R^{*3}$  (installer le package `scatterplot3d` pour la représentation 3d).

```
> X = runifpoint3(5000)
> plot(X)
```
- `rpoispp` : simulation d'un PPP (homogène ou non)
- `rNeymanScott` : PP de N-S avec agrégat
- `rThomas` ....
- `rmh` : simulation d'un PP à partir de son modèle de densité (Strauss, noyau dur, etc)

# Ajustement *finpines* sur un Neymann – Scott

- **(a) Données réelles** : modèle de NS à 3 paramètres  $\theta = (\lambda, \mu, \sigma^{**2})$ 
  - 1 - parents Poisson  $\lambda$
  - 2 - nombre de descendants d'un père Poisson  $\mu$
  - 3 - répartition des fils autour d'un père Gaussienne sphérique  $\sigma^{**2}$→ Ajustement par MCO (cf. poly) puis
- **(b) Simulation du NS estimé** ( $\Delta$  parents et  $\bullet$  descendants)





# PP doublement Poissonien

- PPP à intensité *aléatoire*  $\{\Lambda(s), s \text{ dans } S\}$
- **Exemple : PP de Cox log-gaussien  $X$** 
  - $\Lambda$  suit le modèle *log-linéaire à effet aléatoire*

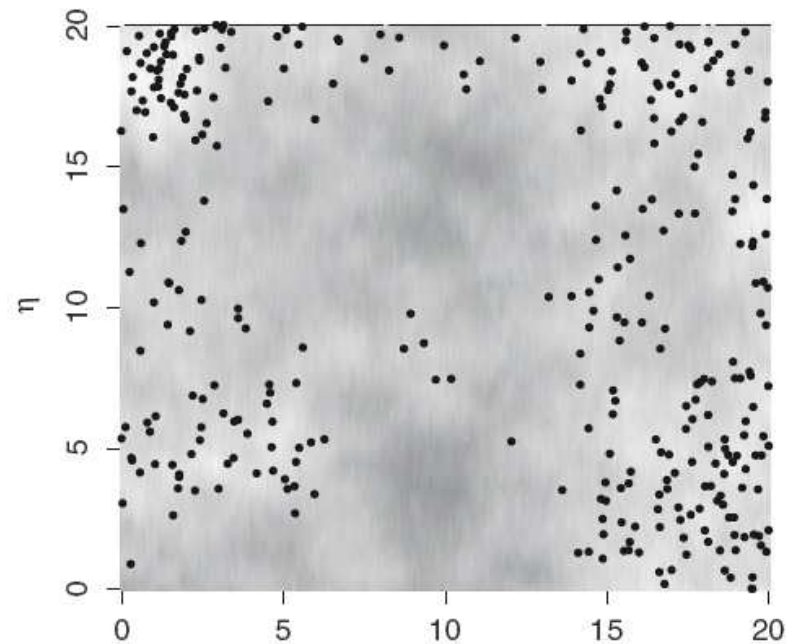
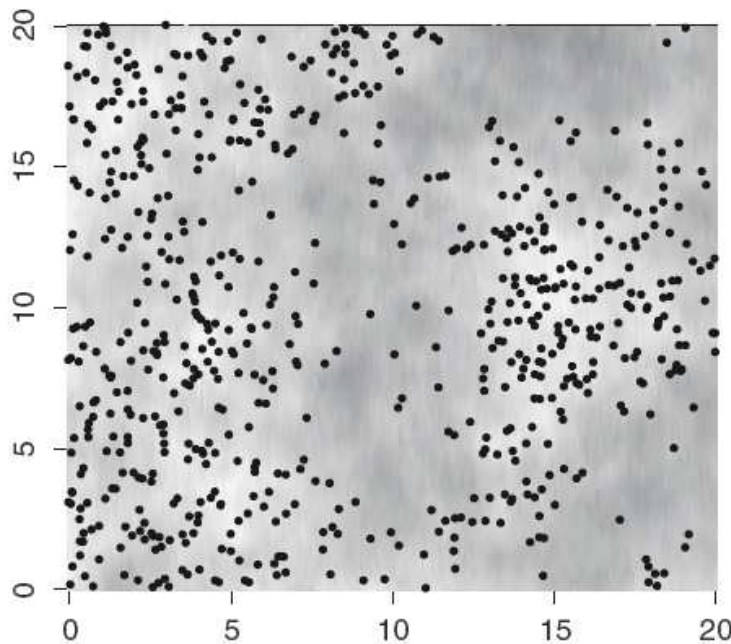
$$\log \Lambda(\xi) = {}^t z(\xi)\beta + \Psi(\xi).$$

- $\psi$  un champ Gaussien centré de covariance  $c$
  - $c$  contrôle la corrélation spatiale de  $X$
- (Moller – Waagepetersen)*

## Deux exemples de PP de Cox

- intensité  $\Lambda$  en fond grisé ( $\Lambda(s)$  grand, fond clair)
- Modèle log-Linéaire :  $\beta = z \equiv 1$  partout
- Deux covariances  $c$  pour l'intensité aléatoire  $\Lambda$

(a)  $c(\xi, \eta) = 3 \exp\{-\|\xi - \eta\|/10\}$ ; (b)  $c(\xi, \eta) = 3 \exp\{-\|\xi - \eta\|^2/10\}$ .



# PP marqué (PPM)

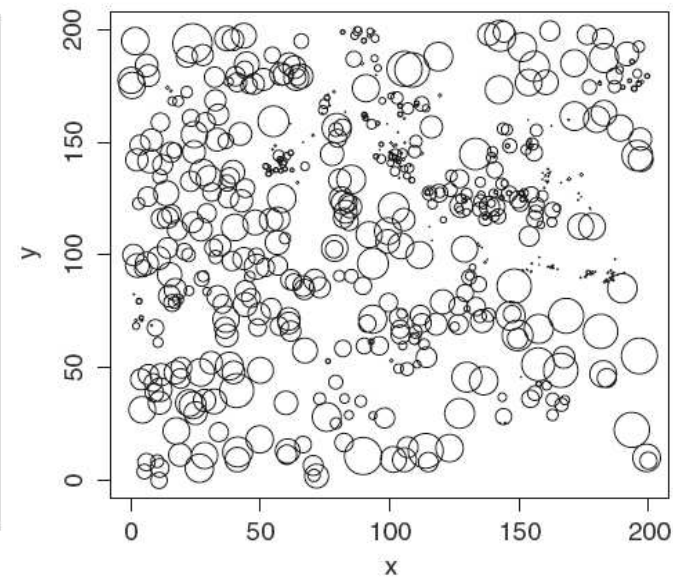
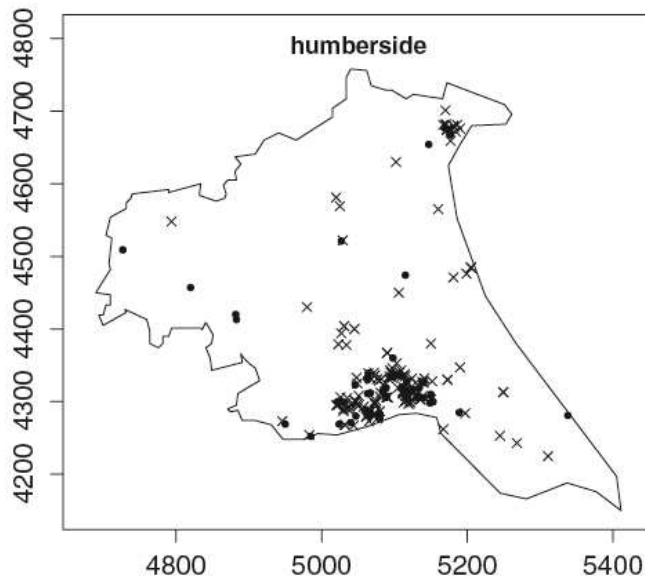
- Une marque  $m(x(i))$  s'ajoute en chaque  $x(i)$  de  $X$
- **Exemples :**
  - marque « *disque* » centré en  $x(i)$  (diamètre arbre)
  - rayon  $r$  du disque est fixé ou aléatoire  $R$
  - marques *fibres* curvilignes attachées à  $x(i)$  (système racinaire d'une plante, segment de longueur et orientation aléatoires)
  - nombre fini  $K$  de marques :  $K$  états possibles, une couleur est associée à chaque  $x(i)$  (i.e.  $K=2$  deux états « sain » ou « malade »)

# Deux exemples de PPM

(a) **Marques binaires** : localisation des 62 cas (●) de *Leucémie* d'un canton et de 141 résidences (x) d'enfants sains ( $K=2$ , données *humberside*)

(b) **Marques continues** : positions et tailles des 584 aiguilles de pin d'un sous bois (données *longleaf*)

**Question sur (a)** : effet spatial influençant la maladie ?



# Densité $f$ d'un PP

- $f$  : densité de probabilité par rapport à un  $PPP(1)$
- $f : E \rightarrow R$ ,  $E$  = espace exponentiel de toute les configurations  $x$
- $f$  admissible si intégrable, d'intégrale  $1$
- En général on définit  $f$  à une constante près :  
 $f(x) = c g(x)$  où  $g$  explicite (mais pas  $c$ !)
- Inutile connaître  $c$  pour la simulation (Metropolis)
- Mais il faut connaître  $c = c(\theta)$  pour l'estimation du MV de  $\theta$

# Exemple de PP à densité : PP de Gibbs

- $U(x) \leftarrow$  potentiel de Gibbs :  $U(x) = \sum \Phi(A)(x)$
- Admissibilité de  $\exp \{U(x)\}$
- **Exemple** : famille exponentielle

$$f(x) = c(\theta) \exp \{ \langle \theta, T(x) \rangle \}$$

# PP de Strauss

- $U(x)$  dérive de 2 statistiques issues de  $x$  :
  - 1 -  $n(x)$  = nombre de points de  $x$
  - 2 -  $s(x)$  = nombre de couples de  $x$  à distances  $< r$
- **Energie** :  $U(x) = a n(x) + b s(x)$  (ou  $a = \log \beta$  et  $b = \log \gamma$ )

$$f_{\theta}(x) = c(\theta) \beta^{n(x)} \gamma^{s(x)}, \quad \theta = {}^t(\beta, \gamma).$$

- $\beta$  (ou  $a$ ) règle l'intensité de  $x$ ;  $\gamma$  règle la régularité spatiale :
- $\gamma < 1$  : d'autant plus régulière que  $\gamma$  petit
  - $\gamma = 1$  : *PP* de *Poisson* homogène d'intensité  $b = \log(\gamma)$
  - $\gamma > 1$  : formation d'agrégats
- *PP* à **noyau dur** :  $\gamma = 0$ , interdit les couples à distance  $< r$

# Simulation Metropolis d'un PP de Gibbs

On circule dans les espaces  $E(n)$  en autorisant à une itération :

→ soit une naissance (proba  $\frac{1}{2}$ )

→ soit une mort (proba  $\frac{1}{2}$ )

suivant la règle suivante :

$$f_{\theta}(x) = c(\theta)\beta^{n(x)}\gamma^{s(x)}, \quad r(x, x \cup \xi) = \frac{\nu(S)f(x \cup \xi)}{n(x)f(x)}.$$

- **Naissance**  $\xi$  retenue avec la proba  $\inf\{1, r(x, x \cup \xi)\}$  ;
- sinon rester en  $x$ .
- **Mort**  $\eta$  retenue avec la probabilité  $\inf\{1, r(x \setminus \eta, x)^{-1}\}$  ;
- sinon rester en  $x$ .

Cet algorithme simule le PP de densité  $f$ .



# Quelques outils statistiques

- **Moments d'ordre 1** ou intensité  
(modèle sur la moyenne)
- Moments d'ordre 2, corrélation repondérée  
(indépendance spatiale ou non)
- **Moment réduit  $K$**  d'ordre 2 de Ripley
- **Distances aux plus proches voisins**

# Moments (intensités) d'ordre 1 et 2 d'un PP

## Moment d'ordre 1

si  $B$  borélien borné :  $\lambda(B) = E(N(B))$

*Intensité  $\rho$  d'ordre 1* :  $\lambda(dx) = \rho(x)dx$

## Intensité d'ordre 2

$$\rho_2(x, y) = \frac{P(N(dx)=1 \text{ et } N(dy)=1)}{dx \, dy}$$

## Corrélation de paires repondérée

$$g(x, y) = 1 + \frac{Cov(N(dx), N(dy))}{\rho(x)\rho(y) \, dx \, dy}$$

1.  $g(\xi, \eta) = 1$  si les points apparaissent indépendamment
2.  $g(\xi, \eta) > 1$  traduit une attraction entre les points (c positive).
3.  $g(\xi, \eta) < 1$  traduit une répulsion entre les points (c

## Moment K de Ripley (cas isotropique)

Soit  $X$  un PP isotropique de densité  $\rho$ .

Soit  $B(x, h)$  la boule de centre  $x$  et de rayon  $h$ .

Soit  $x$  un point de la réalisation  $X$ . Alors :

$$\rho \times K(h) = E(\text{nb points } X \text{ dans } B(x, h))$$

$K$  est liée à  $\rho_2$  : si  $d = 2$ ,  $\rho^2(h)K(h) = 2\pi \int_0^h u\rho_2(u)du$

## **K** et régularité spatiale ?

Soit  $X$  un PP de densité fixé sur  $\mathbb{R}^2$

et la fonctionnelle  $L(h) = \sqrt{\frac{K(h)}{\pi}}$ . On a :

$h \mapsto L(h)$  est  $\left\{ \begin{array}{l} \equiv h \text{ si } X \text{ est un PP de Poisson} \\ \textit{convexe} \text{ si } X \text{ est plus régulier (noyau dur)} \\ \textit{concave} \text{ si } X \text{ présente des agrégats (NS)} \end{array} \right.$

# Distances aux plus proche voisins (ppv)

1 - d'un point  $\bullet$  de  $X$

2 - d'un point  $\circ$  de la fenêtre d'observation

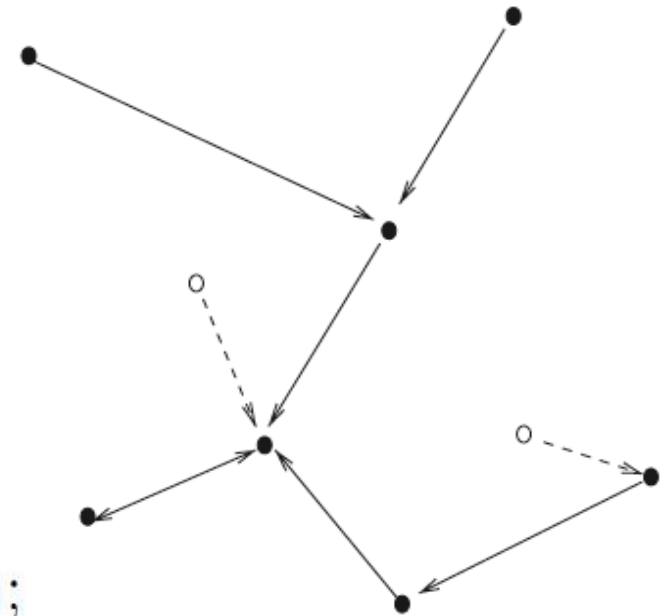
*Deux distributions de distance au ppv :*

**1** - d'un point  $x = \bullet$  de  $X$  :

$$G_x(r) = G(r) = P(d(x, X \setminus \{x\}) \leq r);$$

**2** - d'un point  $x = \circ$  de  $S$  ( $x \notin X$ ) :

$$F(r) = P(d(x, X \setminus \{x\}) \leq r).$$



# Distances aux ppv et régularité spatiale ?

Soit la fonctionnelle associée aux 2 distances :

$$J(r) = \frac{1 - G(r)}{1 - F(r)}$$

Si  $X$  est stationnaire :

Alors si  $J$   $\left\{ \begin{array}{l} = 1, X \text{ est un PP de } Poisson \\ < 1 \text{ indique } plus \text{ de régularité (noyau dur)} \\ > 1 \text{ indique } moins \text{ de régularité (NS)} \end{array} \right.$

$r \mapsto J(r)$  bonne statistique pour évaluer la dépendance spatiale

# Questions de base pour un PP

- La répartition spatiale **X** est-elle due uniquement au hasard ? (*CSR = Complete Spatial Randomness*) ou non ?
- Si non, y a t'il compétition ? coopération ?
- Quel modèle de PP spatial proposer pour **X** ?
- Comment estimer le modèle ? Comment le valider ?
  - Méthodes paramétriques
  - ou méthodes de Monte Carlo

# Test de « CSR » : $X$ est un PPP

Utilisation des distances aux PPV et fonctionnelle  $J$   
Estimation de  $J \leftarrow$  estimations empiriques de  $G$  et  $F$

Indicateur de régularité spatiale :  $J(r) = \frac{1-G(r)}{1-F(r)}$ ,

$\Rightarrow$

Valeur de $J$	$J < 1$	$J = 1$	$J > 1$
répartition	+ régulière	au hasard (PPP)	- régulière

$X \sim PPP(\lambda)$  homogène sur  $\mathbb{R}^2$  :  $G(r) = F(r) = 1 - \exp\{-\lambda\pi r^2\}$ .



# Test de « CSR » : utilisation de K

Autre indicateur :  $L(h) = \sqrt{\frac{K(h)}{\pi}}$

⇒

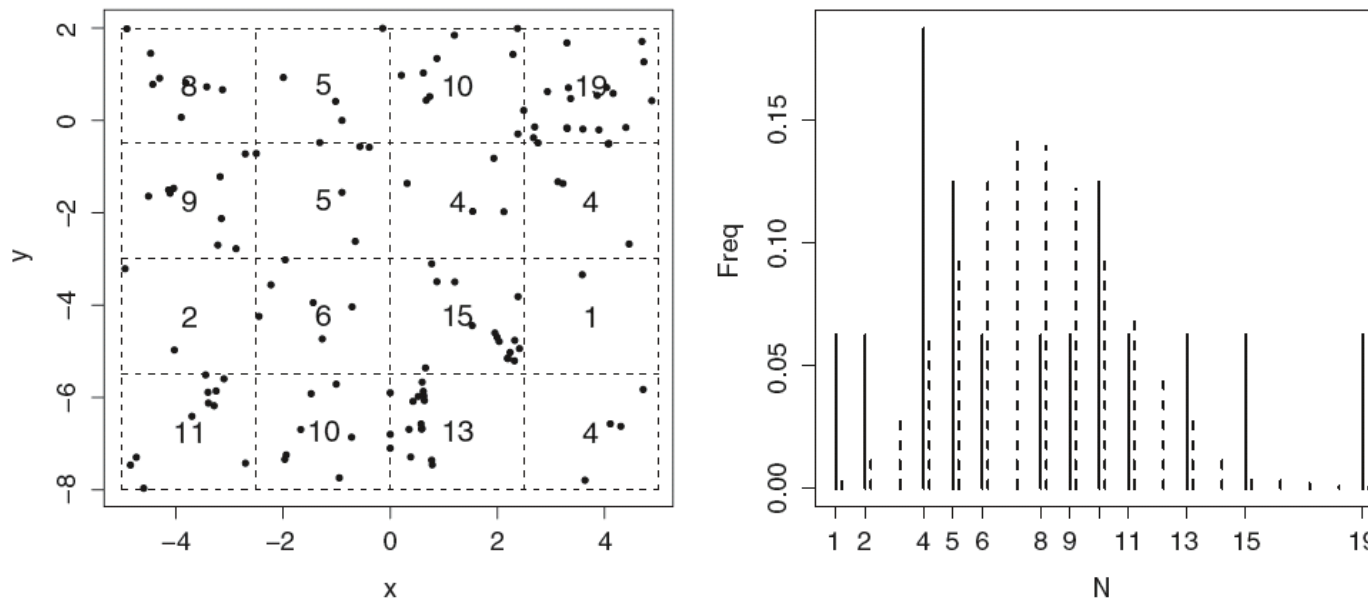
Type de $L$	convexe	linéaire	concave
répartition	+ régulière	au hasard (PPP)	- régulière

L'estimation de  $L$  découle naturellement de celle de  $K$ .

# Test de « CSR » : comptages par quadrats

Chi 2 d'ajustement sur une loi de Poisson (*finpines*)

- On forme (par exemple)  $4 \times 4 = 16$  quadrats même surface
- Comptage des effectifs  $N(i)$  pour chaque quadrat  $i$
- Distribution empirique des  $N(i)$  (à droite)
- Distance à une distribution de Poisson :  $D = 46.7 \gg \chi^2(15;5\%)$

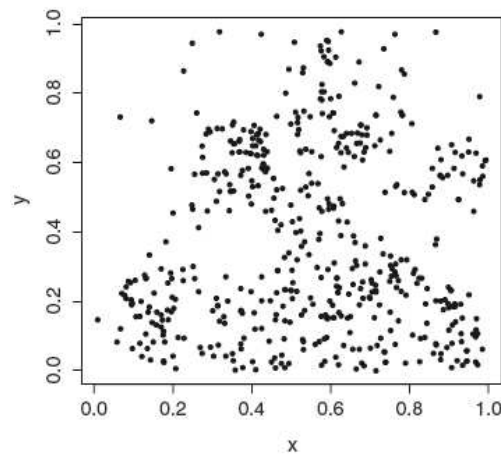


# Estimation de l'intensité du PP

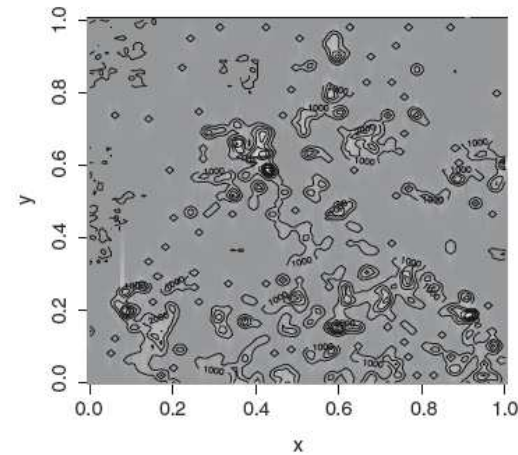
- **Méthodes non paramétriques classiques** : fenêtrage et noyau de convolution
- **Modèle paramétrique** : définir le modèle et estimer
  - par MV en supposant CSR (c'est une PV si CSR non vérifiée)
  - bonne propriétés asymptotiques si  $X$  est ergodique

# Estimation NP d'une densité d'un PP

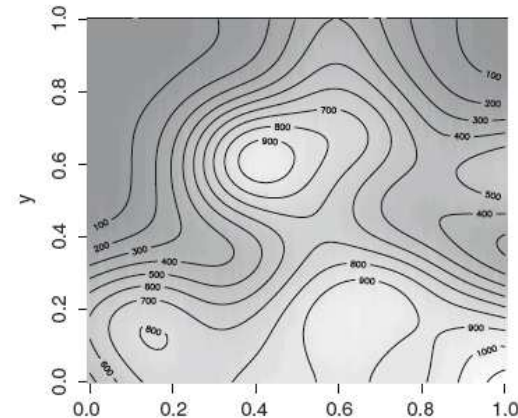
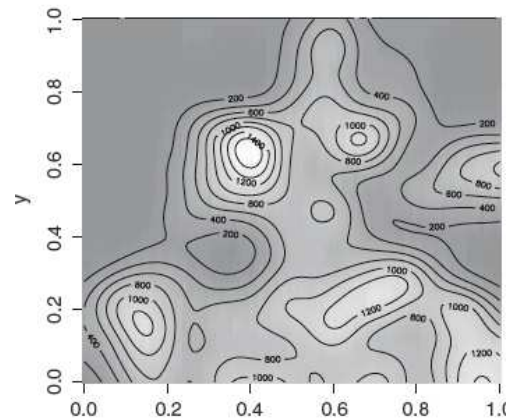
(a) données érables (*lansing*) et  
3 estimations (b-c-d) avec noyaux de + en + régularisant



(a)



(b)



# Estimations NP des distances aux ppv et $K$

- 1 – à partir de statistiques empiriques de comptage
- 2 -- lissage postérieur éventuel

**Moment K de Ripley** : sur le disque  $B$  centré en  $O$  de rayon  $h$  si  $X$  est de densité  $\tau$  :

$$\tau^2 \hat{\mathcal{K}}(B) = \frac{1}{\nu(A)} \sum_{\xi, \eta \in X \cap A}^{\neq} \mathbf{1}_B(\xi - \eta)$$

**Distribution G aux ppv** :  $x$  un point de  $X$ ,  $h(i)$  les  $n$  distances  $d(x, x(i))$  pour les  $n$  points de  $X$  :

$$\hat{G}(h) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_{(0, h_i]}(h)$$

# Estimation d'un modèle d'intensité $\rho(\bullet, \theta)$

- Si  $X$  est *PPP* d'intensité  $\rho(\bullet, \theta)$ , la log-vraisemblance de  $\{x(1), x(2), \dots, x(n)\}$  sur  $A$  est donnée ci-dessous si  $\rho(\bullet, \theta)$  suit un MLG, maximisation via un logiciel dédié
- Si  $X$  est *PP* de densité  $\rho(\bullet)$ , on maximise encore cette Pseudo-Vraisemblance  
Bonne propriété limite si  $X$  ergodique et  $A \rightarrow R^{**2}$ .

$$l_A(\theta) = \sum_{\xi \in x \cap A} \log \rho(\xi; \theta) + \int_A \{1 - \rho(\eta; \theta)\} d\eta.$$

# Moindres carrés pour un modèle $K(\cdot, \theta)$

- 1 – choix d'une famille  $H$  de distances identifiant  $\theta \rightarrow K(\cdot, \theta)$
- 2 – MCO sur une puissance  $c$  de  $K$

Choix (Diggle) :  $c=0.5$  si  $X$  régulier,  $0.25$  si  $X$  avec agrégats

$$D^*(\theta) = \sum_{i=1,k} w_i \{ \hat{K}(h_i)^c - K(h_i; \theta)^c \}^2;$$

$k$  distances  $\mathcal{H} = \{h_1, h_2, \dots, h_k\}$ .

# Autres méthodes paramétriques

- 1 - Pour un PP, on sait définir une densité conditionnelle comme pour un champ de Markov sur un réseau (Jensen-Moller) → PVC pour un PP
- 2 - On maximise cette PVC : sous conditions d'ergodicité et de faible dépendance, « bons résultats asymptotiques
- 3 - Estimation par MV : la difficulté est le calcul de la constante de normalisation → l'approcher par Monte Carlo



# Estimation d'un PP avec `spatstat`

Déclarer le modèle paramétrique : `Poisson`, `Strauss`, `Hardcore`, `StraussHard`, `Geyer`, `N-S`, `Thomas` ....

- `ppm` : ajuste le modèle aux données via la pseudo vraisemblance conditionnelle (si PPP, c'est la vraisemblance); consulter les exemples, i.e :  
> `data(nztrees)` puis > `plot(nztrees)`  
> `ppm(nztrees, ~ x, Strauss(13), correction="periodic")`
- `rmh` : simule le modèle estimé → procédure « visuelle » de validation.

# Estimation d'un PP avec `spatstat` (suite)

- `fitin` : donne l'interaction (0 pour un PPP) du PP estimé
- `lgcp.estK` : ajuste un PP de Cox log-gaussien à covariance exponentielle par MC sur la base de la statistique K
- `thomas.estK` : idem pour un PP de Thomas
- `log.link` : donne la log-vraisemblance de l'ajustement pour un PPP; on en déduit le critère AIC.
- `density` : lissage de densité par convolution (noyau k à choisir)

# Estimation d'un PP avec `spatstat` : statistique et bande de confiance `G`, `J` et `K`

- `Gest` : estimation de la densité cumulée de la distance d'un point de la configuration  $X$  à son ppv dans  $X$   
(`F` si distance d'un point courant à son ppv dans  $X$ )
- `Jest` :  $J(r) = (1 - G(r)) / (1 - F(r))$ .  
`J > 1`, `= 1` et `< 1` indique plus régulier, Poisson, moins régulier
- `Kest` : estime le moment d'ordre 2 `K` de Ripley.
- `Kinhom` : pour un PP inhomogène
- `envelope` : calcule les bandes de confiances par simulation des statistiques de base (`K`, `G`, `J` ...)
- `Quadrat.test` : test du chi2 de l'hypothèse CSR d'indépendance spatiale.

# Test de Monte Carlo de $H(0)$ : « $X$ est CSR »

Test basé sur un intervalle de confiance sur  $K$

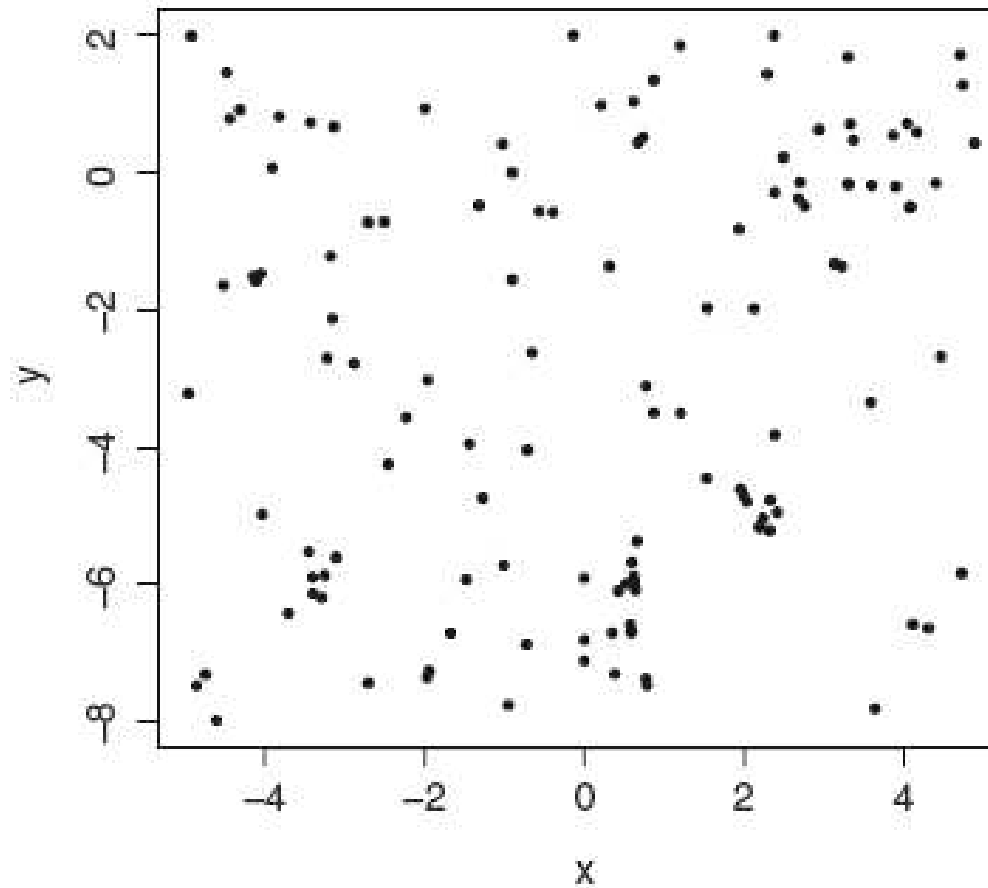
On observe  $x$ . Choisir  $L$  distances  $\{h(1), h(2), h(L)\}$ .

- Estimer (empiriquement)  $K$  par  $K^*(h(l))$  à ces  $L$  distances
- Sous  $H(0)$ , estimation  $\rho^*$  de l'intensité  $\rho$  d'un PPP
- Simulation de  $m$  réalisations  $x^*(l)$ ,  $l=1, m$ , d'un PPP( $\rho^*$ ) (i.e.  $m=20$ )
- $\rightarrow m$  estimations  $K^*(i, h(l))$  associées à chaque  $x^*(i)$
- Enveloppes *inf* et *sup* de ces  $m$  estimations
- Si  $K^*$  se trouve entre ces 2 enveloppes, accepter  $H(0)$  (ici, niveau 10%)

On peut aussi comparer les  $K^*$  à  $\pi h^2$ , valeur théorique de  $K$  pour un PPP.  
Cette méthode décrit le principe du *Bootstrap paramétrique*.

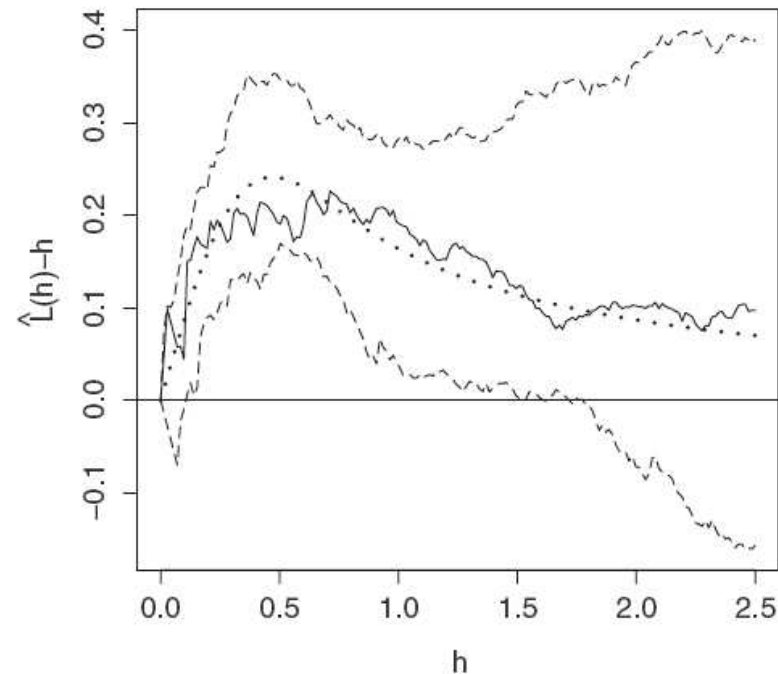
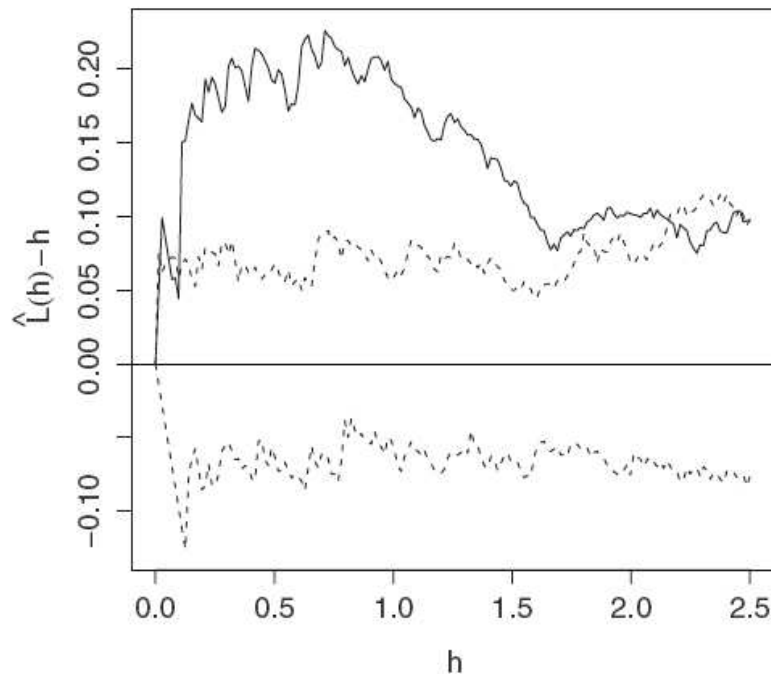
# Les 156 pins d'une forêt finlandaise

*(finpines)*



# « Indépendance $H(0)$ » contre « Neyman-Scott » (données *finpines*)

- En continu, le graphe estimé de  $h \rightarrow L(h) - h$  ( $=0$  si  $(H(0))$ ) pour  $x$
- À gauche : bande de confiance pour  $m=40$  simulations  $x^*$  sous  $H(0)$
- À droite : bande sous l'alternative N-S, avec en pointillé la courbe  $h \rightarrow K(h)$  théorique pour les paramètres de N-S estimés.



---

## *spatstat*: le package **R** pour les PP spatiaux

(spatial PP analysis, model fitting, simulation, ...)

- *Anova.ppm*: déviance (s) entre deux ou plus que deux modèles
- *Jest*: estimation (empirique) de l'indicateur J du caractère (J=1) PPP ou non
- *Kest*: idem pour la fonction moment réduit d'ordre 2, K
- *nearest.neighbour*: distances aux ppv
- *ppm*: ajuste un modèle de PP à des données,
- *rmh.ppm*: simulation par Metropolis-Hastings (mh) d'un modèle de PP,
- *rpoispp*: simulation d'un PPP homogène ou non)
- Données : *cells*, *finpines*, .....