DATOS en una RED

Parte 1 : Autocorrelacion

Marzo 2019

Consecuencia de la autocorrelacion

$X_1, X_2, \ldots X_n$ sucesion de v.a. correlacionadas,

$\text{Var}(X_i) = \sigma^2$, $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

- $\text{Var}(\overline{X}) > \frac{\sigma^2}{n}$ si $\text{cov}(X_i, X_j) > 0$
- $\text{Var}(\overline{X}) < \frac{\sigma^2}{n}$ si $\text{cov}(X_i, X_j) < 0$

Consecuencias

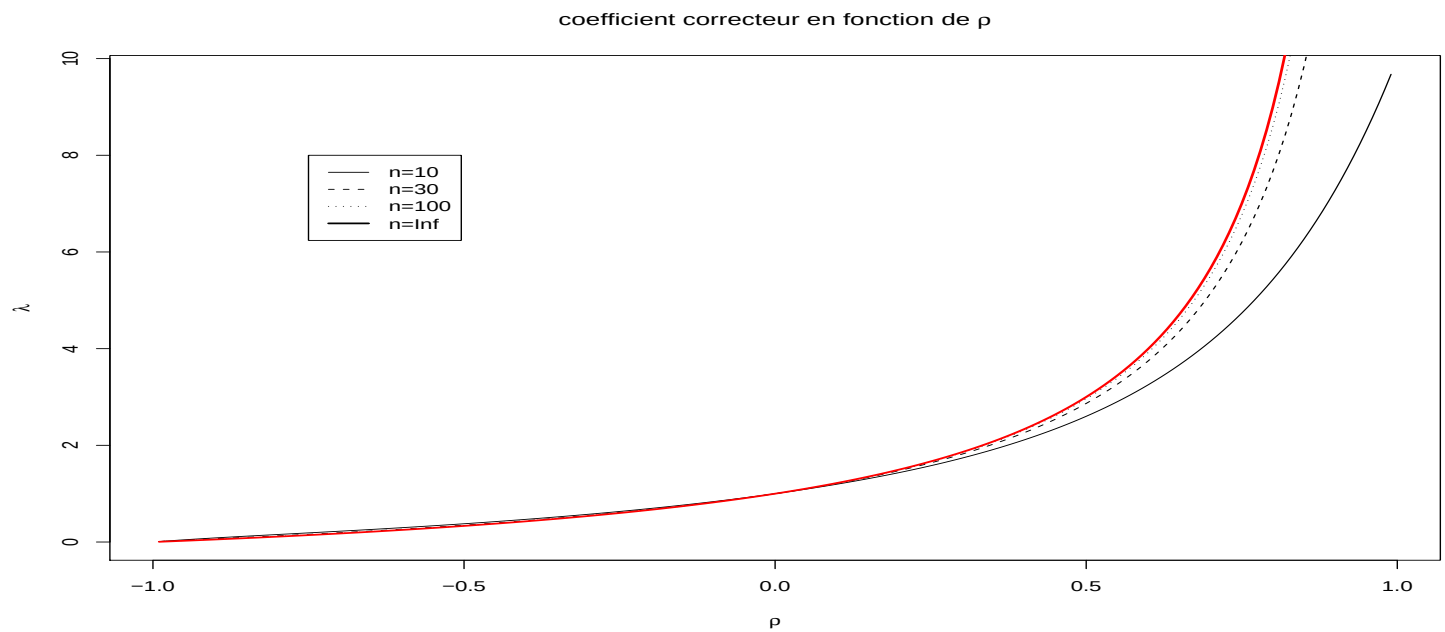
- los intervalos de confianza pueden ser mayores o menores que en el caso independiente
- las p-values de prueba para dos muestras pueden ser mayores o menores que en el caso independiente.

Consecuencia de la autocorrelacion

Ejemplo 1

$$X_t = \rho X_{t-1} + \varepsilon_t \qquad |\rho| < 1 \qquad \text{cov}(X_t, X_{t+h}) = \sigma_X^2 \rho^h$$

$$V(\overline{X}) = \frac{\sigma_X^2}{n}\left(1 + 2\left(\frac{\rho}{1-\rho}\right)\left(1 - \frac{1}{n}\right) - 2\left(\frac{\rho}{1-\rho}\right)^2\left(\frac{1-\rho^{n-1}}{n}\right)\right) = \lambda\frac{\sigma_X^2}{n}$$



Intervalos de confianza de nivel 95%

n=10, $\rho = 0.25$, $\lambda = 1.67$, $I = [\overline{x} - 2.49\frac{\sigma}{\sqrt{n}}\ ;\ \overline{x} + 2.49\frac{\sigma}{\sqrt{n}}]$

n=10, $\rho = -0.25$, $\lambda = 0.63$, $I = [\overline{x} - 1.23\frac{\sigma}{\sqrt{n}}\ ;\ \overline{x} + 1.23\frac{\sigma}{\sqrt{n}}]$

Consecuencia de la autocorrelacion

Ejemplo 2

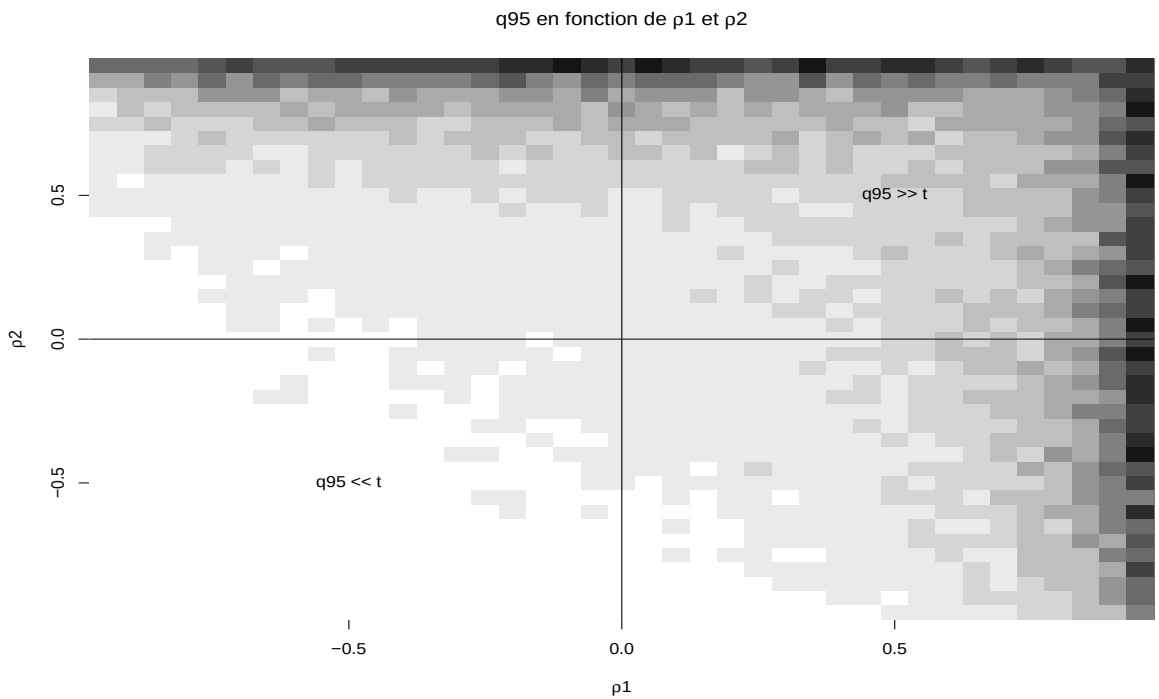
$X^i \sim \mathcal{N}(\mu_i, \sigma)$, $X_t^i = \rho_i X_{t-1}^i + \varepsilon_t^i \; i = 1, 2 \quad t = 1, T_i$

Prueba de dos muestras : H_0: $\mu_1 = \mu_2$ contra H_1: $\mu_1 \neq \mu_2$

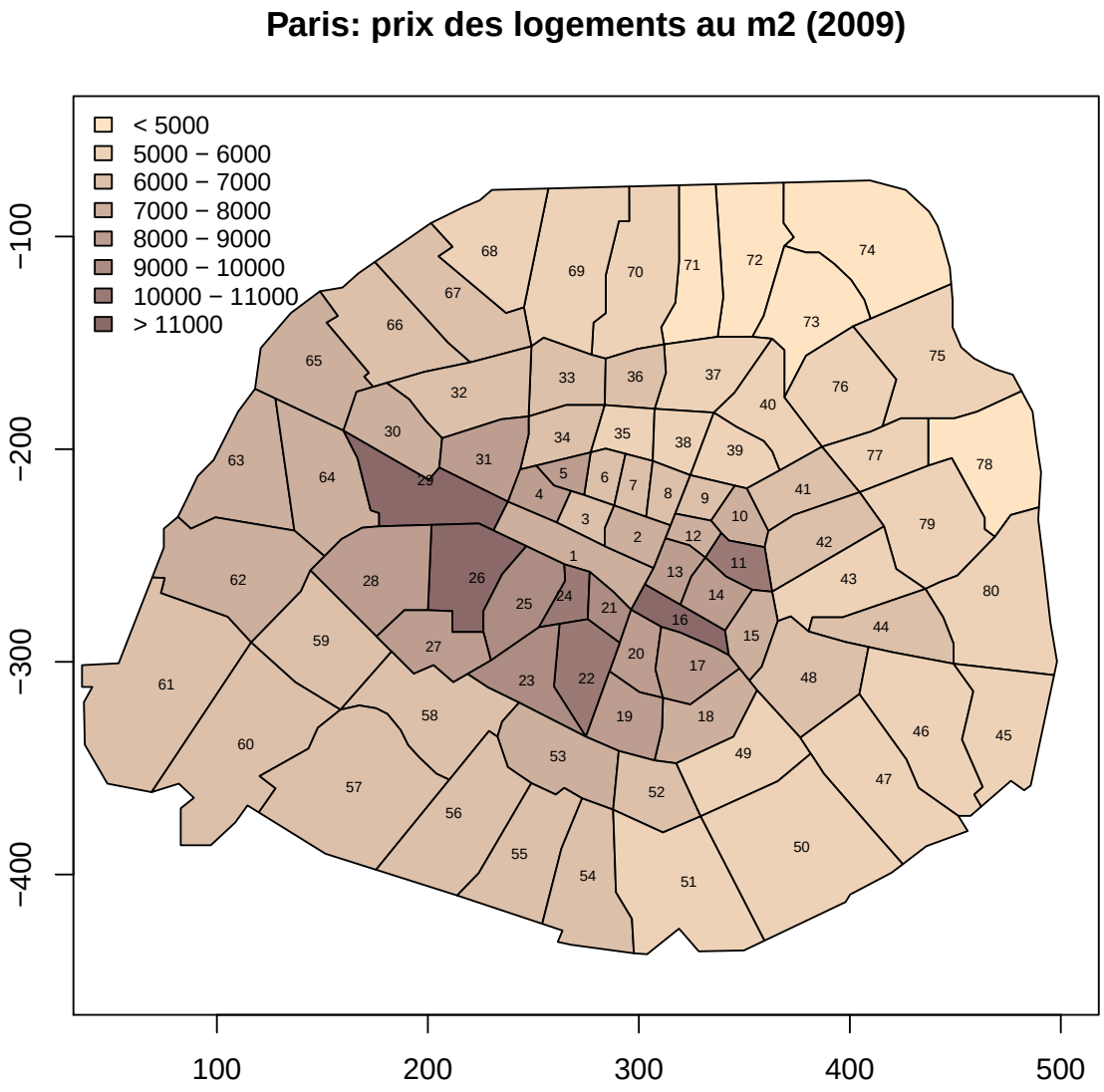
Si las muestras son independientes

$Z = \frac{\overline{X^1} - \overline{X^2}}{S\sqrt{\frac{1}{T_1} + \frac{1}{T_2}}} \sim \mathcal{T}(T_1 + T_2 - 2)$, $R = \{|Z| > z\}$

Caso dependiente: estimacion de $\widehat{q}$ cuantil de nivel 95% (simulacion)



Autocorrelacion espacial



Vecindad

$S = (s_i)_{i=1,n}$, nodos (lugares) de D.
Grafo $\mathcal{G}$: relacion binaria de $S \times S$: s_i esta relacionado con s_j.
Matriz W **de adyacencia**: matriz cuadrada de tamaño n^2

$$w_{ij} = \begin{cases} 0 & \text{si } s_i = s_j \\ 0 & \text{si } (s_i, s_j) \notin \mathcal{G} \end{cases}$$

Ejemplos

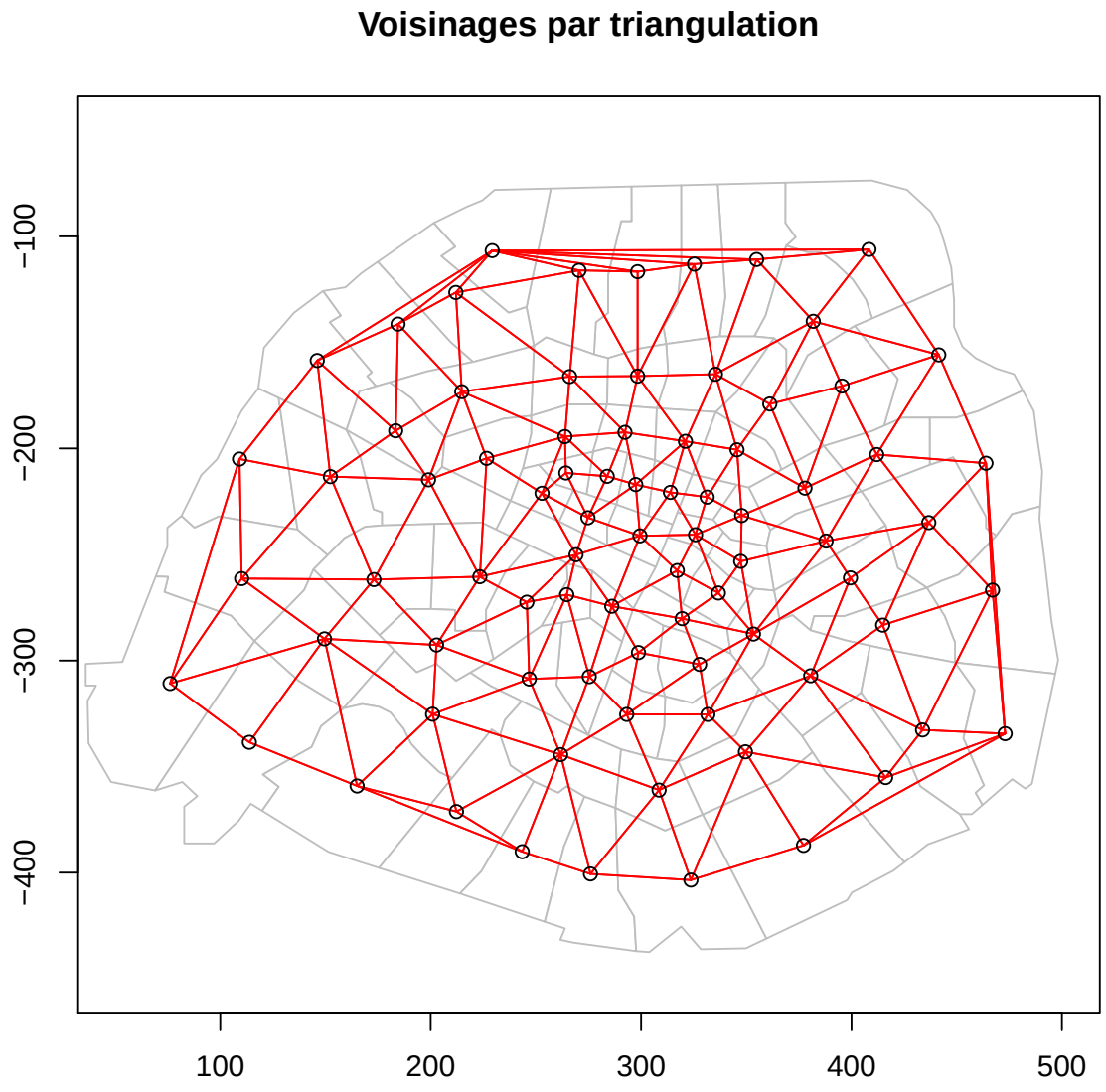
W con coeficientes binarios

$$w_{ij} = \begin{cases} 1 & \text{si } (s_i, s_j) \in \mathcal{G} \\ 0 & \text{sino} \end{cases}$$

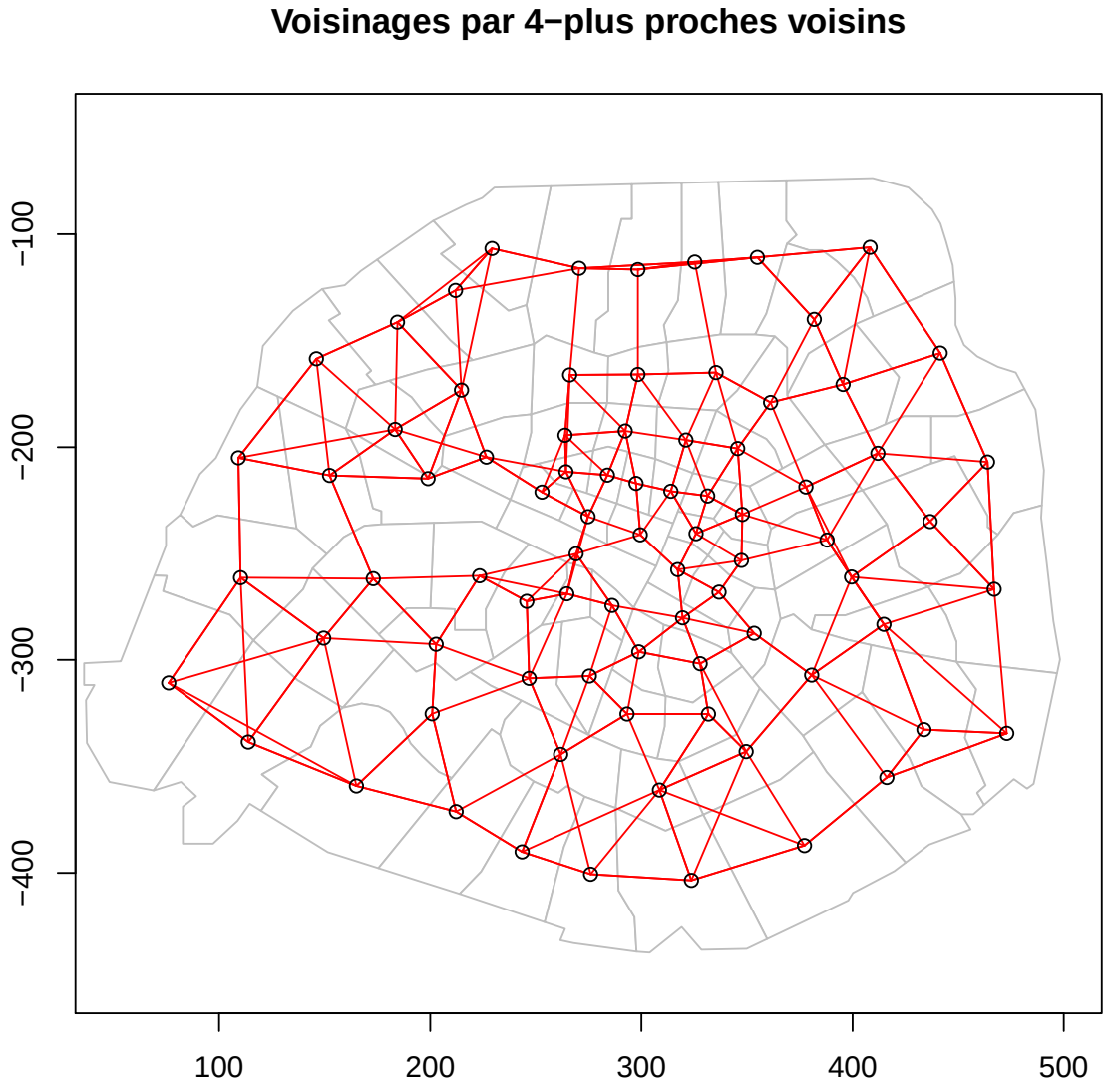
W con coeficientes basados en las distancias

$$w_{ij} = \begin{cases} \frac{1}{d_{ij}} & \text{si } (s_i, s_j) \in \mathcal{G} \qquad d_{ij} = \text{dist}(s_i, s_j) \\ 0 & \text{sino} \end{cases}$$

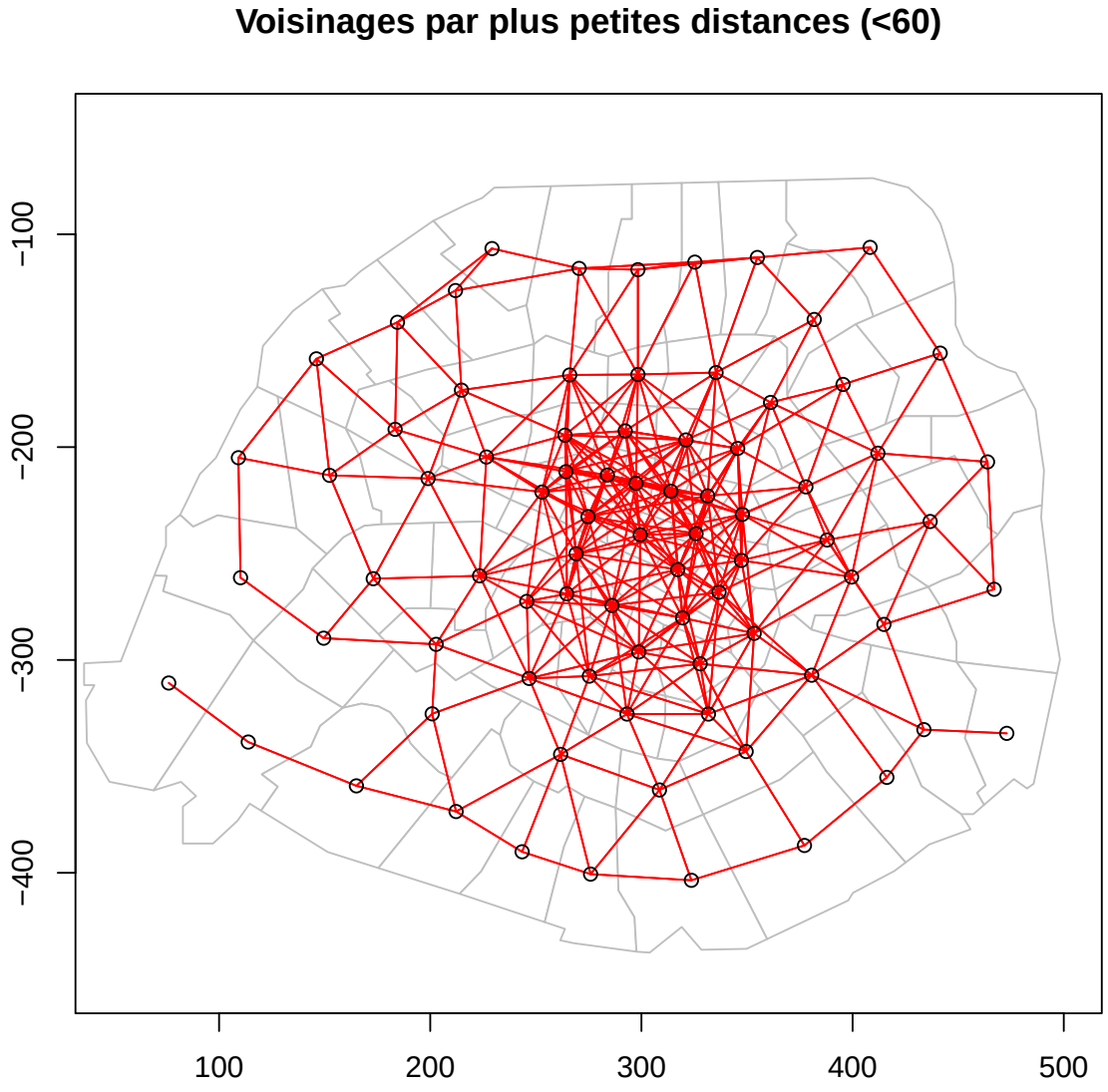
Vecindad


Voisinages par triangulation
−100
−200
−300
−400
100
200
300
400
500

Vecindad


Voisinages par 4−plus proches voisins
−100
−200
−300
−400
100
200
300
400
500

Vecindad


Voisinages par plus petites distances (<60)
−100
−200
−300
−400
100
200
300
400
500

Vecindad

$$S_0 = \sum_{i \neq j} w_{ij} \quad S_1 = \frac{1}{2} \sum_{i \neq j} (w_{ij} + w_{ji})^2 \quad S_2 = \sum_{i=1}^{n} (w_{i+} + w_{+i})^2$$

```
Characteristics of weights list object:
Neighbour list object:
Number of regions: 80
Number of nonzero links: 446
Percentage nonzero weights: 6.96875
Average number of links: 5.575

Weights style: B
Weights constants summary:
   n   nn  S0  S1    S2
B 80 6400 446 892 10304
```

Datos binarios

$Z_i = Z(s_i) = 0$ (blanco) o 1 (negro) , $\mathrm{P}(Z_i = 1) = p$

$$NN = \frac{1}{2}\sum_{i,j} w_{ij} Z_i Z_j \text{ numero de pares de vecinos negros}$$

$$NB = \frac{1}{2}\sum_{i,j} w_{ij} (Z_i - Z_j)^2 \text{ numero de pares de vecinos blanco-negro}$$

$$BB = \frac{1}{2}\sum_{i,j} w_{ij} - NN - NB$$

Distribucion bajo la hypotesis de independencia :

- modelo binomial : $\mathcal{B}(n, p)$
- modelo hypergeometrico : $\mathcal{H}(n, n_1)$

Calculo de los momentos

$$R = \sum_{i \neq j} w_{ij} Y_{ij}$$

$$\begin{aligned}
\mathrm{E}(R) &= \sum_{i \neq j} w_{ij} \mathrm{E}(Y_{ij}) \\
\mathrm{V}(R) &= \sum_{i \neq j} w_{ij}(w_{ij} + w_{ji}) \mathrm{V}(Y_{ij}) \\
&\quad + \sum_{i \neq j \neq k} (w_{ij} + w_{ji})(w_{ik} + w_{ki}) \mathrm{cov}(Y_{ij}, Y_{ik}) \\
&\quad + \sum_{i \neq j \neq k \neq \ell} w_{ij} w_{k\ell} \mathrm{cov}(Y_{ij}, Y_{k\ell})
\end{aligned}$$

Calculo de los momentos

Modelo binomial :

$$\mathrm{E}(NN) = \tfrac{1}{2} S_0 p^2$$
$$\mathrm{V}(NN) = \tfrac{1}{4}\left(S_1(p^2 - p^4) + (S_2 - 2S_1)(p^3 - p^4)\right)$$
$$\mathrm{E}(BN) = S_0 p(1-p),$$
$$\mathrm{V}(BN) = S_1 p(1-p) + \tfrac{1}{4}\big(S_2 p(1-p)(1 - 4p(1-p))\big)$$

Modelo hypergeometrico : $n^{(p)} = \frac{n!}{(n-p)!}$

$$\mathrm{E}(NN) = \tfrac{1}{2} S_0 \tfrac{n_1^{(2)}}{n^{(2)}}$$
$$4\mathrm{V}(NN) = S_1\big(\tfrac{n_1^{(2)}}{n^{(2)}} - 2\tfrac{n_1^{(3)}}{n^{(3)}} + \tfrac{n_1^{(4)}}{n^{(4)}}\big) + S_2\big(\tfrac{n_1^{(3)}}{n^{(3)}} - \tfrac{n_1^{(4)}}{n^{(4)}}\big) + S_0^2 \tfrac{n_1^{(4)}}{n^{(4)}} - \big(S_0 \tfrac{n_1^{(2)}}{n^{(2)}}\big)^2$$

Prueba de independencia

Teorema

Si los Y_{ij} estan uniformemente acotados en D entonces la estadistica $R = \sum_{i \neq j} w_{ij} Y_{ij}$ converge a una distribucion de probabilidad normal si $S_0^{-2} Var(R)$ es exactemente de orden n^{-1}.

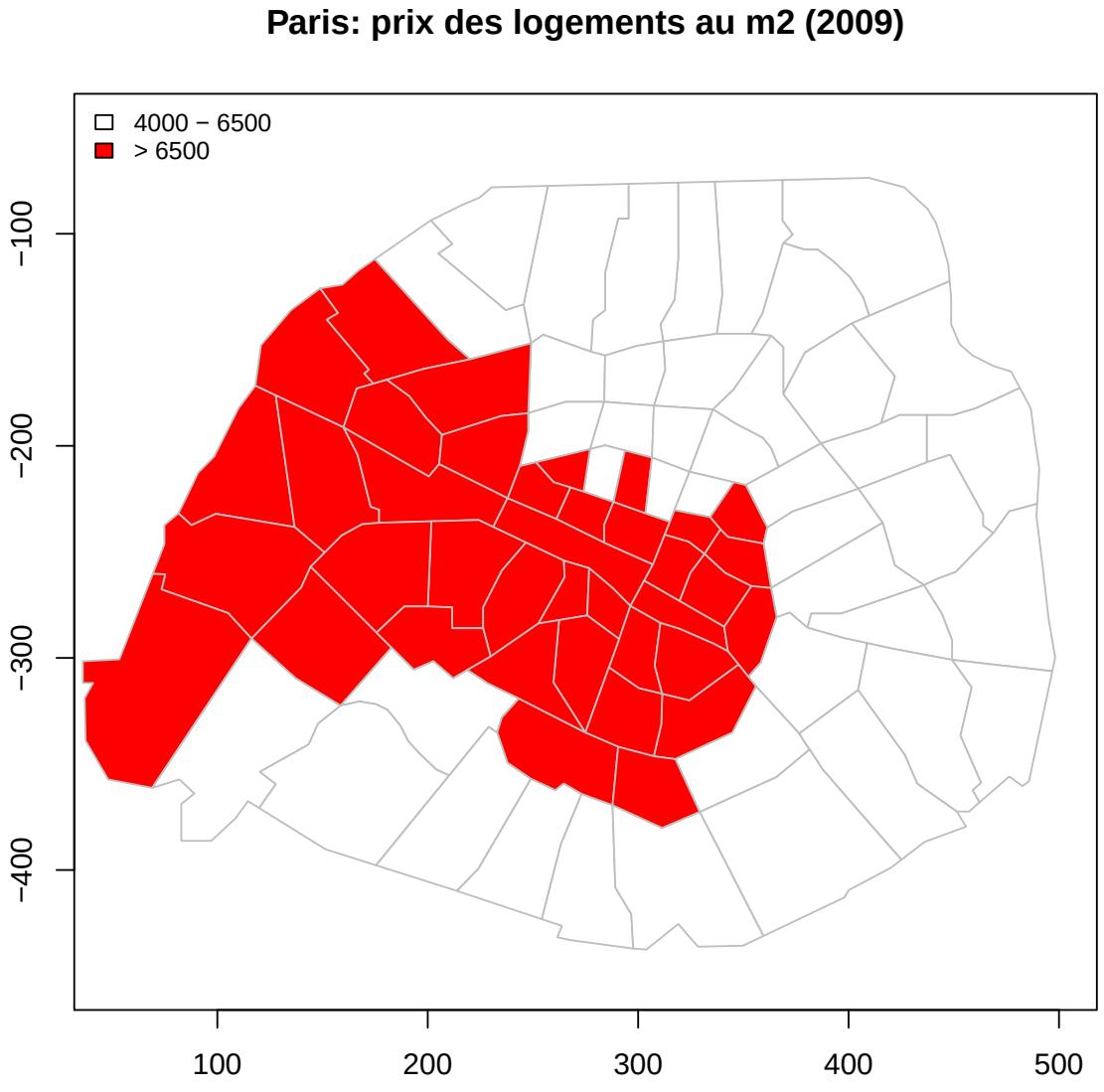
Comentario

- La condicion se verifica para grillas regulares, o si el numero de vecinos esta uniformemente acotado.
- Este resultado da una estadistica de prueba en el caso binomial.

Prueba por permutaciones

- repartir aleatoriamente n_1 celdas negras y $n - n_1$ celdas blancas,
- determinar las distribuciones empiricas de BB, BN, NN,
- comparar con los valores observados.

Datos binarios



$$B = 42$$
$$N = 38$$
$$NN = 65$$
$$BB = 80$$

Prueba de independencia

```
        Join count test under nonfree sampling

Std. deviate for faible = 4.3687, p-value = 6.251e-06
sample estimates:
Same colour statistic     Expectation          Variance
              80.00000        60.76044         19.39523


        Join count test under nonfree sampling

Std. deviate for fort = 3.6039, p-value = 0.0001568
sample estimates:
Same colour statistic     Expectation          Variance
              65.00000        49.61044         18.23553
```

Calculo de los momentos

```
        Monte-Carlo simulation of join-count statistic

number of simulations + 1: 101
Join-count statistic for faible = 80,
rank of observed statistic = 101,
p-value = 0.009901
sample estimates:
    mean of simulation variance of simulation
              60.82000                17.26020

Join-count statistic for fort = 65,
rank of observed statistic = 101,
p-value = 0.009901
sample estimates:
    mean of simulation variance of simulation
              48.84000                18.29737
```

Indice de Moran

Z un campo aleatorio real, dotado de un grafo de vecindad

Indice de Moran:

$$I = \frac{n}{\sum_{i \neq j} w_{ij}} \frac{\sum_{i,j} w_{ij}(Z_i - \overline{Z})(Z_j - \overline{Z})}{\sum_i (Z_i - \overline{Z})^2}$$

si $I > 0$ hay correlacion positiva (clusters)
si $I < 0$ hay correlacion negativa (repulsion)
si $I \approx 0$ no hay correlacion

Indice de Geary

Indice de Geary:

$$C = \frac{n-1}{2\sum_{i\neq j} w_{ij}} \frac{\sum_{i,j} w_{ij}(Z_i - Z_j)^2}{\sum_i (Z_i - \overline{Z})^2}$$

si $C \approx 0$ hay correlacion positiva (clusters)
si $C \gg 0$ hay correlacion negativa (repulsion)

Distribucion bajo la hypotesis de independencia

- modelo Gaussiano : $Z_i \sim \mathcal{N}(\mu, \sigma)$
- modelo remuestrado : $Z_i \in \{z_1, z_2, \ldots z_n\}$ equiprobables

Teorema
Si $(X_1, \ldots, X_n)$ *i.i.d,* $X_i \sim \mathcal{N}(0,1)$ *y* $h : \mathbb{R}^n \to \mathbb{R}$ *es invariante por cambio de escala, entonces* $h(X_1, \ldots, X_n)$ *es independiente de* $Q = \sum_{i=1}^n X_i^2$.

Calculo de los momentos

Indice de Moran

- modelo Gaussiano

$$\mathrm{E}(I) = -\frac{1}{n-1}$$

$$\mathrm{V}(I) = \frac{1}{(n-1)(n+1)S_0^2}(n^2 S_1 - nS_2 + 3S_0^2) - \frac{1}{(n-1)^2}$$

- modelo remuestrado

$$\mathrm{E}(I) = -\frac{1}{n-1}$$

$$\begin{aligned}\mathrm{V}(I) &= \frac{1}{(n-1)^{(3)}S_0^2}\Big(n\big((n^2-3n+3)S_1 - nS_2 + 3S_0^2\big) \\ &\quad -b_2\big((n^2-n)S_1 - 2nS_2 + 6S_0^2\big)\Big) - \frac{1}{(n-1)^2}\end{aligned}$$

$b_2 = \frac{m_4}{m_2^2}$ $\qquad$ $nm_4 = \sum_i (z_i - \overline{z})^4$

Calculo de los momentos

Indice de Geary

- modelo Gaussiano

$$\mathrm{E}(C) = 1$$

$$\mathrm{V}(C) = \frac{2(S_1 + S_2)(n-1) - 4S_0^2}{2(n+1)S_0^2}$$

- modelo remuestrado

$$\mathrm{E}(C) = 1$$

$$\begin{aligned}\mathrm{V}(C) \;=\; & \frac{1}{n(n-1)^{(2)}S_0^2}\Big((n-1)S_1\big(n^2-3n+3-(n-1)b_2\big) \\ & -\frac{1}{4}(n-1)S_2\big(n^2+3n-6-(n^2-n+2)b_2\big) \\ & +S_0^2\big(n^2-3-(n-1)^2b_2\big)\Big)\end{aligned}$$

Prueba de independencia

Teorema

λ_i son los valores propios de la matriz de adyacencia W. Si $\frac{\sum_i \lambda_i^j}{\left(\sum_i \lambda_i^2\right)^{1/2}} = o(1)$ para $j = 1, \ldots$ entonces bajo el modelo Gaussiano los indices I y C convergen en distribucion a una distribucion normal. Este resultado sigue siendo verdadero si las variables no son gaussianas pero de momento de orden 4 finito.

Comentario Si W est symetrica, el numero de vecinos esta uniformemente acotado y $0 < \lim \frac{S_1}{n} < \infty$ entonces la condicion del teorema se verifica.

Prueba por permutaciones

- repartir aleatoriamente los valores observados,
- deducir las distribuciones empiricas de I et C,
- comparar con los valores observados.

Pruebas Moran et Geary

```
        Moran's I test under normality

Moran I statistic standard deviate = 6.8281,
p-value = 4.301e-12
alternative hypothesis: greater sample
estimates: Moran I statistic    Expectation    Variance
                     0.42700       -0.01266     0.00415

        Geary's C test under normality

Geary C statistic standard deviate = 5.5842,
p-value = 1.174e-08
alternative hypothesis: Expectation greater than statistic
estimates: Geary C statistic    Expectation     Variance
                     0.60739        1.00000      0.00494
```

Pruebas Moran et Geary

```
        Moran's I test under randomisation

Moran I statistic standard deviate = 6.8754,
p-value = 3.091e-12
estimates: Moran I statistic    Expectation     Variance
                     0.42700       -0.01266      0.00409

        Geary's C test under randomisation

Geary C statistic standard deviate = 5.3571,
p-value = 4.227e-08
estimates: Geary C statistic    Expectation     Variance
                     0.60739        1.00000      0.00537
```

Pruebas Moran et Geary

```
        Monte-Carlo simulation of Moran's I

number of simulations + 1: 101

statistic = 0.427
observed rank = 101, p-value = 0.009901
alternative hypothesis: greater

        Monte-Carlo simulation of Geary's C

number of simulations + 1: 101

statistic = 0.6074
observed rank = 1, p-value = 0.009901
alternative hypothesis: less
```

Correlograma espacial

Vecindad de orden $\mathbf{k}$

$w_{ij}^{(k)} = 1$ si s_i y s_j son vecinos de orden k

$w_{ij}^{(k)} = \frac{1}{d_{ij}}$ si $d_{ij} \in$ clase de distancia k

$$I(k) = \frac{n}{\sum_{i,j} w_{ij}^{(k)}} \frac{\sum_{i,j} w_{ij}^{(k)} (Z_i - \overline{Z})(Z_j - \overline{Z})}{\sum_i (Z_i - \overline{Z})^2}$$

Correlograma espacial

```
Spatial correlogram for prixm2
method: Moran's I
     estimate expectation    variance     sd       Pr(I)
1  0.42495941 -0.01265823  0.00512735 6.1115 9.869e-10 ***
2  0.24510554 -0.01265823  0.00249602 5.1594 2.478e-07 ***
3  0.11222228 -0.01265823  0.00158181 3.1399   0.00169 **
4  0.04612056 -0.01265823  0.00121587 1.6857   0.09186 .
5 -0.00080356 -0.01265823  0.00109872 0.3576   0.72061
```

Correlograma espacial