Auto-correlation et régression spatiales

Objectifs Réaliser une étude d'auto-corrélation spatiale et une modélisation linéaire spatiale avec le logiciel utilisé est **R**et ses packages *spdep, spatialreg* et *fields*.

Données

Les espèces épibenthiques (qui vivent sur les fonds marins) sont des biomarqueurs de l'état environnemental et des sources d'alimentation des poissons de la baie du Saint Laurent. Depuis plusieurs années des campagnes scientifiques sont menées pour mesurer la biomasse de ces espèces. Le Golfe du Saint Laurent est découpé en strates de milieux homogènes.

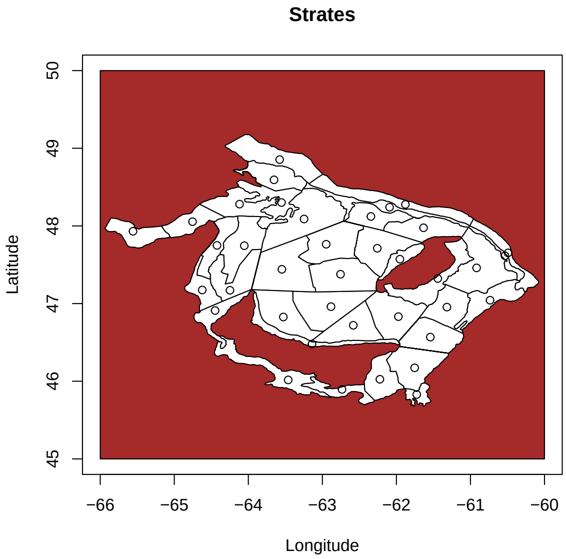
Le fichiers StLaurent.csv contient pour l'année 2000 les valeurs des variables suivantes :

strate :	numéro de la strate
longitude :	longitude du centre de la strate
latitude :	latitude du centre de la strate
BH :	présence (1) absence (0) de bernard l'ermite
totconsum :	indice de prédation
depth :	profondeur
temperature :	température

Le fichier "boundaries38s.dat" donne pour les 38 strates les coordonnées des points formant les bords :

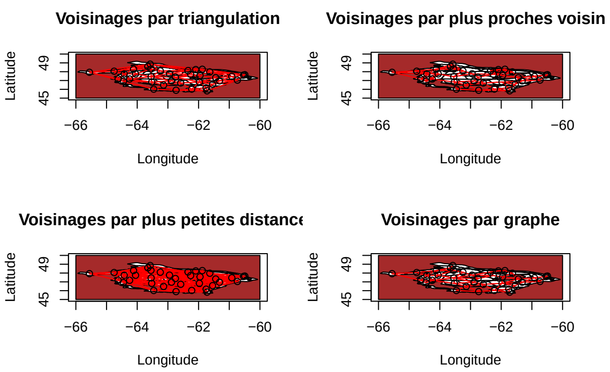
colonne 1 :	numéro de la strate (de 401 à 438)
colonne 2 :	longitude du point
colonne 3 :	latitude du point

```
    strate        longitude        latitude           BH
 Min.   : 1.00   Min.   :-65.56   Min.   :45.83   Min.   :0.0000
 1st Qu.:10.25   1st Qu.:-63.57   1st Qu.:46.85   1st Qu.:0.0000
 Median :19.50   Median :-62.74   Median :47.45   Median :1.0000
 Mean   :19.50   Mean   :-62.75   Mean   :47.36   Mean   :0.6579
 3rd Qu.:28.75   3rd Qu.:-61.78   3rd Qu.:47.96   3rd Qu.:1.0000
 Max.   :38.00   Max.   :-60.49   Max.   :48.85   Max.   :1.0000
   totconsum            depth          temperature
 Min.   :0.02763   Min.   : 17.50   Min.   :-0.1750
 1st Qu.:0.10862   1st Qu.: 35.48   1st Qu.: 0.7264
 Median :0.20314   Median : 54.85   Median : 2.2900
 Mean   :0.22550   Mean   : 75.99   Mean   : 3.9179
 3rd Qu.:0.27933   3rd Qu.: 69.11   3rd Qu.: 5.4694
 Max.   :0.69466   Max.   :295.90   Max.   :15.8333
```

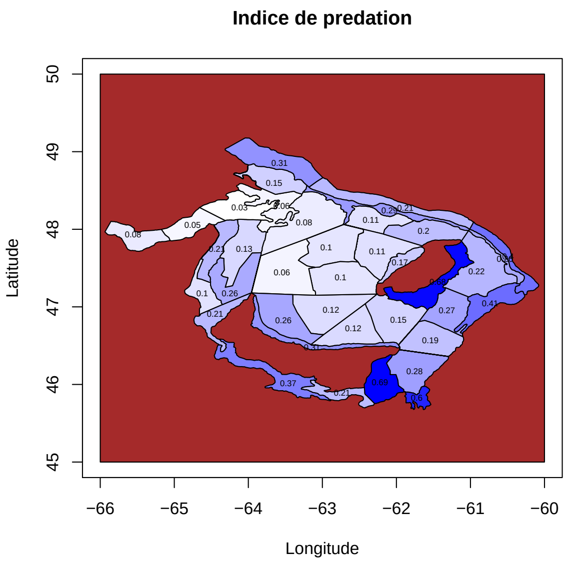


1 Auto-corrélation spatiale

L'autocorrélation spatiale mesure la corrélation d'une variable avec elle-même, lorsque les observations sont décalées dans l'espace.
Pour obtenir l'auto-corrélation et réaliser un test d'autocorrélation, il faut spécifier une matrice de poids W. On cherche une matrice de voisinage permettant de décrire de manière adéquate les relations entre les entités géographiques tout en estimant un modèle de régression identifiable (choix pas toujours facile).
Des indices d'autocorrélation spatiale (Moran ou Geary) évaluent la dépendance spatiale d'une même variable en différents sites et teste la significativité de la structure spatiale identifiée. Ils s'intéressent à la proximité spatiale et à la ressemblance ou dissemblance des sites étudiés.



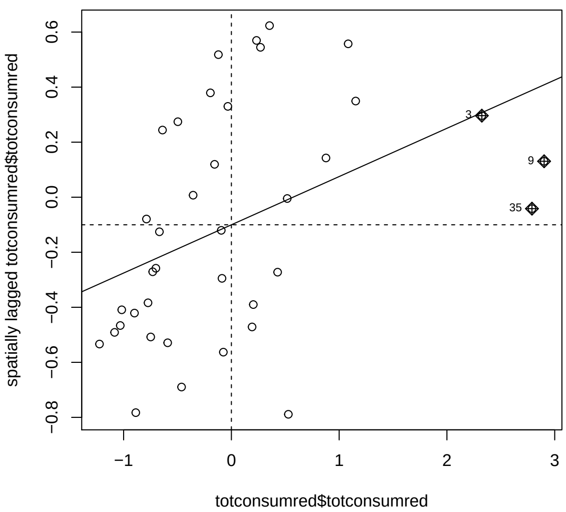
1.1 Indice de Moran ou Geary



2 Diagramme de Moran

Le diagramme de Moran (nuage de points entre les valeurs de la variable Z étudiée centrée et valeurs moyennes de la variable pour les observations voisines (WZ) en ordonnée, avec W la matrice de poids normalisée) permet une lecture rapide de la structure spatiale. A ce nuage, on ajoute la droite de régression linéaire de WZ par Z, les droites $Z = 0$ et $WZ = 0$ qui délimitent 4 quadrants.
Si la répartition des données est aléatoire, pas de relation particulière entre Z et WZ (répartition uniforme dans les quadrants). Sinon, la pente de la régression linéaire est non nulle ; chaque quadrant donne un type d'association spatiale. Ce diagramme permet également de relever les sites les plus atypiques, loins de la structure spatiale étudiée.

- Le quadrant 1 (en haut à droite) donne des valeurs de Y plus grandes que la moyenne, dans un voisinage qui leur ressemblent (autocorrélation positive et indice élevé).
- Le quadrant 2 (en bas à droite) donne des valeurs de Y plus grandes que la moyenne, dans un voisinage qui ne leur ressemble pas (autocorrélation négative et indice élevé).
- Le quadrant 3 (en bas à gauche) donne des valeurs de Y plus faibles que la moyenne, dans un voisinage qui leur ressemble (autocorrélation positive et indice faible).
- Le quadrant 4 (en haut à gauche) donne des valeurs de Y plus faibles que la moyenne, dans un voisinage qui ne leur ressemble pas (autocorrélation négative et indice faible).



Le diagramme précédent met en évidence une structure spatiale particulière (cadrants 1, 3 et 4). On calcule donc l'indice d'autocorrélation spatiale et effectue un test afin de répondre à la question : 1) les observations voisines sont-elles sensables ou pas par hasard ? Si la réponse est non il y a de l'autocorrélation

spatiale mais dans quel sens. On teste la nullité de l'indice sous l'hypothèse de normalité des données ou non (dans ce dernier cas on utilise la randomisation ; la statistique des observations est comparée à la distribution obtenue en permuttant au hasard les données).
Le programme suivant montre qu'il y a une auto-corrélation spatiale positive (indide de Moran positive et hypothèse de nullité de l'indice rejetée) avec un seul décalage.

```
        Moran I test under randomisation

data:  totconsumred$totconsumred
weights: poids.vois

Moran I statistic standard deviate = 3.9927, p-value = 3.267e-05
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
      0.175289338      -0.027027027       0.002567662
        Moran I test under randomisation

data:  totconsum
weights: poids.vois

Moran I statistic standard deviate = 3.9927, p-value = 3.267e-05
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
      0.175289338      -0.027027027       0.002567662
        Moran I test under normality

data:  totconsum
weights: poids.vois

Moran I statistic standard deviate = 3.8743, p-value = 5.346e-05
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
      0.175289338      -0.027027027       0.002726872
        Monte-Carlo simulation of Moran I

data:  totconsum
weights: poids.vois
number of simulations + 1: 100

statistic = 0.17529, observed rank = 100, p-value = 0.01
alternative hypothesis: greater
```

Calcul des intervales de l'indice de Moran :

```
[1] -0.3545677  1.0195219
```

Indice de Geary : montre également une autocorrélation spatiale positive (indice plus petite que 1)

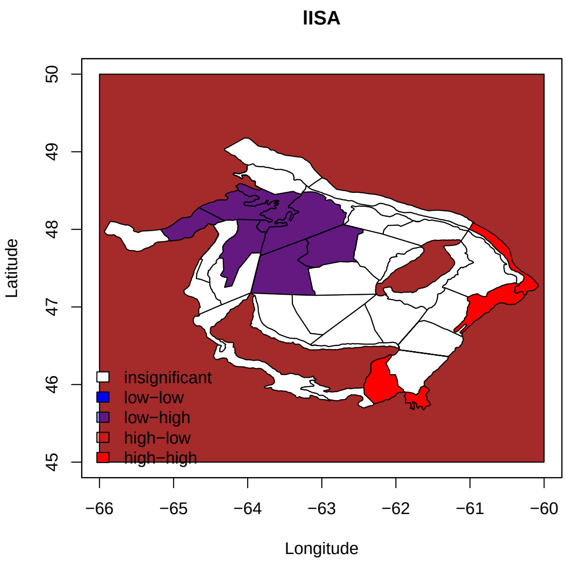
```
        Monte-Carlo simulation of Geary C

data:  totconsum
weights: poids.vois
number of simulations + 1: 100

statistic = 0.77677, observed rank = 1, p-value = 0.01
alternative hypothesis: greater
        Geary C test under randomisation

data:  totconsum
weights: poids.vois

Geary C statistic standard deviate = 3.301, p-value = 0.0004816
alternative hypothesis: Expectation greater than statistic
```

```
sample estimates:
Geary C statistic       Expectation          Variance
      0.776772218       1.000000000       0.004572916
        Geary C test under normality

data:  totconsum
weights: poids.vois

Geary C statistic standard deviate = 3.7115, p-value = 0.000103
alternative hypothesis: Expectation greater than statistic
sample estimates:
Geary C statistic       Expectation          Variance
       0.77677222        1.00000000        0.00361743
```

2.1 Autocorrélation spatiale locale

Les indices précédents mesurent l'autocorrélation globale, pour étudier l'autocorrélation spatiale locale (LISA ; Local Indicators of Spatial Association), des versions locales des indices précédents existents et permettent de mesurer pour une variable donnée la dépendance locale entre une unité spatiale et les unités spatiales voisines. Ils permettent d'identifier les regroupements similaires autour d'un site donné et les zones de non-stationnarité spatiale locale.
L'Indice de Moran locale en un site i est :

$$I_i = \sum_j w_{ij}(Z_j - \overline{Z})(Z_i - \overline{Z})$$

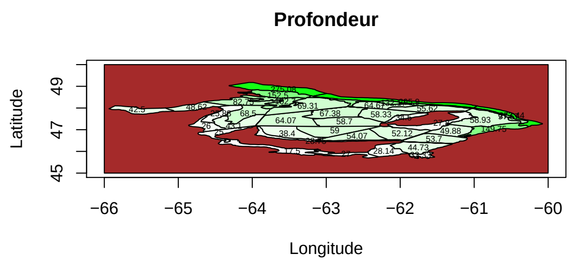


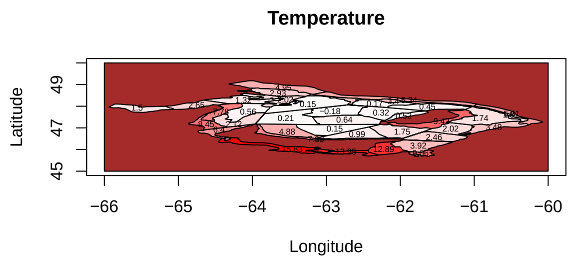
3 Régression spatiale

3.1 Modèle linéaire

```
Call:
lm(formula = totconsum ~ temperature + depth)

Residuals:
     Min       1Q   Median       3Q      Max
-0.26612 -0.06910 -0.01072  0.06687  0.32157

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.1028305  0.0381121   2.698   0.0107 *
temperature 0.0259357  0.0050844   5.101 1.18e-05 ***
depth       0.0002771  0.0003021   0.917   0.3652
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1258 on 35 degrees of freedom
Multiple R-squared:  0.4272,        Adjusted R-squared:  0.3945
F-statistic: 13.05 on 2 and 35 DF,  p-value: 5.817e-05
Analysis of Variance Table

Response: totconsum
            Df  Sum Sq Mean Sq F value    Pr(>F)
temperature  1 0.39986 0.39986 25.2645 1.481e-05 ***
depth        1 0.01332 0.01332  0.8417    0.3652
Residuals   35 0.55394 0.01583
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Call:
lm(formula = totconsum ~ temperature)

Residuals:
     Min       1Q   Median       3Q      Max
-0.27395 -0.06573 -0.01683  0.06186  0.31137

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.126135   0.028351   4.449 7.98e-05 ***
temperature 0.025363   0.005035   5.037 1.34e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1255 on 36 degrees of freedom
Multiple R-squared:  0.4135,        Adjusted R-squared:  0.3972
F-statistic: 25.38 on 1 and 36 DF,  p-value: 1.34e-05
Analysis of Variance Table

Response: totconsum
            Df  Sum Sq Mean Sq F value   Pr(>F)
temperature  1 0.39986 0.39986  25.376 1.34e-05 ***
Residuals   36 0.56727 0.01576
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```



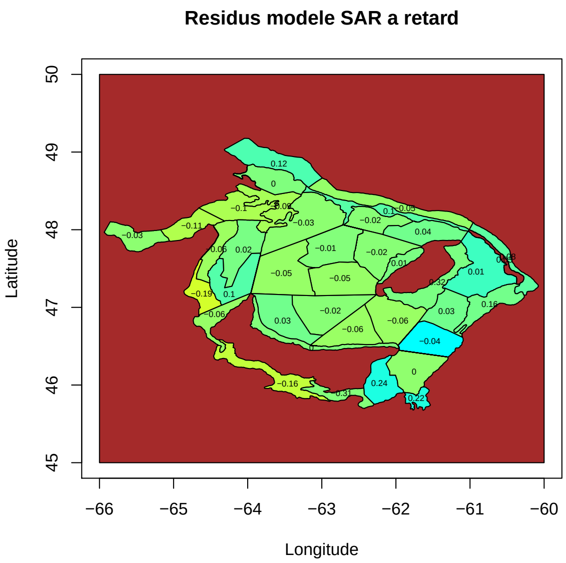


3.2 Modèle SAR

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

$\varepsilon \sim N(0, \sigma^2)$, ou pas $(I - \rho W)$ inversible

```
Call:lagsarlm(formula = totcons.lm, listw = poids.vois, type = "lag",
    method = "eigen")

Residuals:
      Min        1Q    Median        3Q       Max
-0.307128 -0.055084 -0.010262  0.034457  0.316731

Type: lag
Coefficients: (asymptotic standard errors)
             Estimate Std. Error z value  Pr(>|z|)
(Intercept) 0.0247105  0.0515772  0.4791    0.6319
temperature 0.0234929  0.0047247  4.9724 6.613e-07

Rho: 0.5196, LR test value: 4.3249, p-value: 0.037558
Asymptotic standard error: 0.22894
    z-value: 2.2696, p-value: 0.02323
Wald statistic: 5.1512, p-value: 0.02323

Log likelihood: 28.12851 for lag model
ML residual variance (sigma squared): 0.012969, (sigma: 0.11388)
Number of observations: 38
Number of parameters estimated: 4
AIC: -48.257, (AIC for lm: -45.932)
LM test for residual autocorrelation
test value: 0.31833, p-value: 0.57261
        Moran I test under randomisation

data:  res
weights: poids.vois

Moran I statistic standard deviate = 0.86453, p-value = 0.1936
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
      0.017065351      -0.027027027       0.002601188
```



Les mesures importantes de pertinence sont la valeur de log-vraisemblance, les critères d'Akaïke et de Schwarz. Ces critères permettront de déceler la meilleure spécification spatiale : Significativité du paramètre autorégressif ; Variation des valeurs estimées des coefficients des variables explicatives.

Interprétation du modèle : dans un modèle linéaire classique, l'impact d'une variation d'une variable explicative sur la réponse est égale au coefficient correspondant (dérivée partiel de la réponse en fonction de la variable explcative).
Dans ce modèle SAR ; les éléments diagonaux de la matrice des dérivées partielles de z_i par rapport à x_i représentent les impacts directs qui incluent les effets de spillovers (rétombées) propres, hétérogènes pour chaque observation en présence d'autocorrélation spatiale, dûs aux poids différents de la matrice d'intéractions. Les éléments hors-diagonaux représentent les impacts indirects ou effets de spillovers (collectés dans la matrice) ils quantifient les effets globaux (sur tout l'échantillon). En sommant les impacts directes (resp. indirects), on obtient l'impact total direct (resp indirect).

```
 [1] 0.02472291 0.02475493 0.02476920 0.02468111 0.02461042 0.02453486
 [7] 0.02475523 0.02471041 0.02467303 0.02459006
```

```
[1] 0.02424721
```

```
           [,1]
[1,] 0.02465551
```

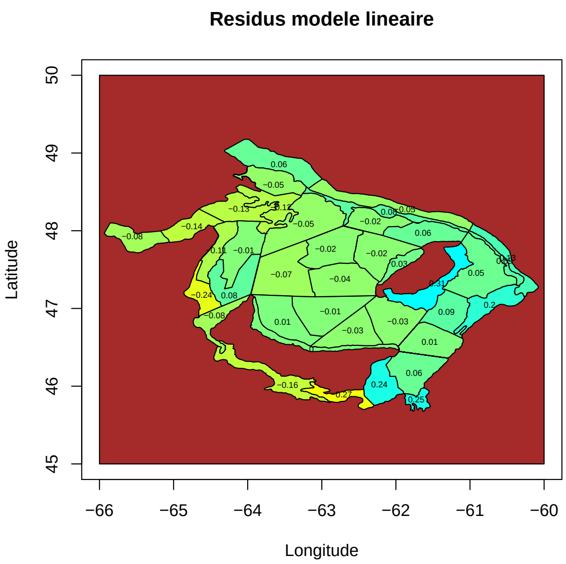
```
           [,1]
[1,] 0.04890272
```

3.3 Modèle SEM

Autocorrélation des residus (modèle SEM) : $Z = \rho W X + X\beta + \varepsilon$, avec hypothese de normalité $\varepsilon \sim N(0, \sigma^2)$ ou non

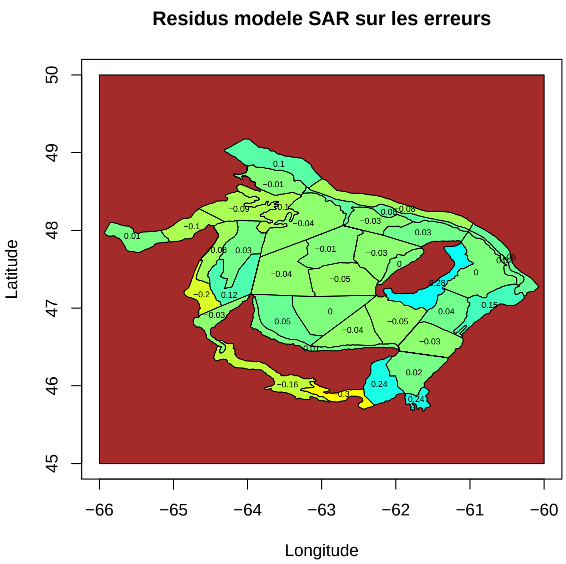
```
        Global Moran I for regression residuals

data:
model: lm(formula = totconsum ~ temperature)
weights: poids.vois

Moran I statistic standard deviate = 4.8907, p-value = 5.023e-07
alternative hypothesis: greater
sample estimates:
Observed Moran I      Expectation         Variance
     0.216121473     -0.031968887      0.002573186
```

```
        Moran I test under randomisation

data:  res0
weights: poids.vois

Moran I statistic standard deviate = 4.6978, p-value = 1.315e-06
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
      0.216121473      -0.027027027       0.002678911
```



On accepte l'existence d'une auto-corrélation positive

```
Call:errorsarlm(formula = totconsum ~ temperature, data = donnees,
    listw = poids.vois)

Residuals:
       Min         1Q     Median         3Q        Max
-0.2990746 -0.0498816 -0.0059529  0.0366786  0.2841549

Type: error
Coefficients: (asymptotic standard errors)
             Estimate Std. Error z value  Pr(>|z|)
(Intercept) 0.1291709  0.0501543  2.5755   0.01001
temperature 0.0257756  0.0048375  5.3283 9.913e-08

Lambda: 0.59644, LR test value: 5.5791, p-value: 0.018176
Asymptotic standard error: 0.21997
    z-value: 2.7115, p-value: 0.006698
Wald statistic: 7.3522, p-value: 0.006698

Log likelihood: 28.75558 for error model
ML residual variance (sigma squared): 0.012413, (sigma: 0.11142)
Number of observations: 38
Number of parameters estimated: 4
AIC: -49.511, (AIC for lm: -45.932)
        Moran I test under randomisation

data:  res1
weights: poids.vois

Moran I statistic standard deviate = 0.096528, p-value = 0.4616
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
     -0.022090961      -0.027027027       0.002614879
```

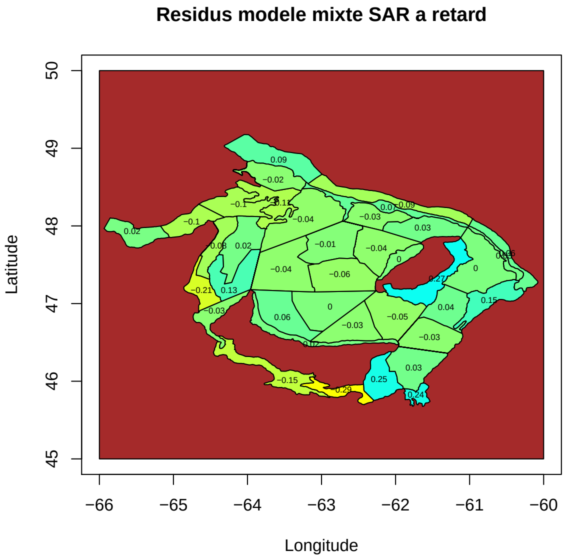


3.4 Modèle SDM : Modele mixte SAR à retard

Modèle avec décalages des variables explicatives) : $Z = \rho_1 W_1 Z + \rho_2 W_2 X + X\beta + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$. C'est un modèle SAR avec le décalage spatial des variables explicatives en plus

```
Call:
lagsarlm(formula = totconsum ~ temperature, data = donnees, listw = poids.vois,
    type = "mixed")

Residuals:
       Min         1Q     Median         3Q        Max
```

```
-0.2912936 -0.0469297 -0.0072158  0.0394243  0.2748867

Type: mixed
Coefficients: (asymptotic standard errors)
                  Estimate Std. Error z value  Pr(>|z|)
(Intercept)      0.0683143  0.0617441  1.1064    0.2685
temperature      0.0263167  0.0051316  5.1283 2.923e-07
lag.temperature -0.0204269  0.0175523 -1.1638    0.2445

Rho: 0.59134, LR test value: 5.4494, p-value: 0.019576
Asymptotic standard error: 0.22203
    z-value: 2.6633, p-value: 0.0077381
Wald statistic: 7.0931, p-value: 0.0077381

Log likelihood: 28.80455 for mixed model
ML residual variance (sigma squared): 0.012391, (sigma: 0.11132)
Number of observations: 38
Number of parameters estimated: 5
AIC: -47.609, (AIC for lm: -44.16)
LM test for residual autocorrelation
test value: 1.392, p-value: 0.23806
        Moran I test under randomisation

data:  res3
weights: poids.vois

Moran I statistic standard deviate = 0.080123, p-value = 0.4681
alternative hypothesis: greater
sample estimates:
Moran I statistic       Expectation          Variance
     -0.022922531      -0.027027027       0.002624225
Call:lagsarlm(formula = totcons.sarlm, data = donnees, listw = poids.vois,
    type = "lag", method = "eigen")

Residuals:
       Min         1Q     Median         3Q        Max
-0.2912936 -0.0469297 -0.0072158  0.0394243  0.2748867

Type: lag
Coefficients: (asymptotic standard errors)
               Estimate Std. Error z value  Pr(>|z|)
(Intercept)   0.0683143  0.0617441  1.1064    0.2685
temperature   0.0263167  0.0051316  5.1283 2.923e-07
Wtemperature -0.0204269  0.0175523 -1.1638    0.2445

Rho: 0.59134, LR test value: 5.4494, p-value: 0.019576
Asymptotic standard error: 0.22203
    z-value: 2.6633, p-value: 0.0077381
Wald statistic: 7.0931, p-value: 0.0077381

Log likelihood: 28.80455 for lag model
ML residual variance (sigma squared): 0.012391, (sigma: 0.11132)
Number of observations: 38
Number of parameters estimated: 5
AIC: -47.609, (AIC for lm: -44.16)
LM test for residual autocorrelation
test value: 1.392, p-value: 0.23806
```



Interpretation du modèle SDM

```
[1] 0.02593592
            [,1]
[1,] -0.01152341
           [,1]
[1,] 0.01441252
Impact measures (mixed, exact):
                Direct    Indirect      Total
temperature 0.02593592 -0.01152341 0.01441252
```

3.5 Modèle SLX

Modèle SLX : $Z = \rho WX + X\beta + \varepsilon$, $\varepsilon \sim N(0, \sigma^2$, décalage spatial des variables explicatives

```
Call:
lm(formula = totcons.sarlm, data = donnees)

Residuals:
     Min       1Q   Median       3Q      Max
-0.26519 -0.06498 -0.01168  0.06715  0.29310

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.150851   0.061046   2.471   0.0185 *
temperature   0.026685   0.005851   4.561    6e-05 ***
Wtemperature -0.008764   0.019111  -0.459   0.6494
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1269 on 35 degrees of freedom
Multiple R-squared:  0.417,	Adjusted R-squared:  0.3836
F-statistic: 12.51 on 2 and 35 DF,  p-value: 7.939e-05
```

3.6 Comparaison des modèles

```
        Lagrange multiplier diagnostics for spatial dependence
data:
model: lm(formula = totconsum ~ temperature + depth)
weights: poids.vois

      statistic parameter  p.value
LMerr    9.3736         1 0.002201 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
        Lagrange multiplier diagnostics for spatial dependence
data:
model: lm(formula = totconsum ~ temperature + depth)
weights: poids.vois

       statistic parameter  p.value
LMerr   9.373615         1 0.002201 **
LMlag   8.628057         1 0.003310 **
RLMerr  0.837464         1 0.360123
RLMlag  0.091906         1 0.761768
SARMA   9.465521         2 0.008802 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
        Lagrange multiplier diagnostics for spatial dependence

data:
residuals: residuals(totcons.lm)
weights: poids.vois

LMErr = 9.3736, df = 1, p-value = 0.002201
        Likelihood ratio for spatial linear models

data:
Likelihood ratio = 0.097937, df = 1, p-value = 0.7543
sample estimates:
     Log likelihood of totcons.mixte Log likelihood of totcons.errorsarlm
                           28.80455                             28.75558
                   df       AIC
totcons.mixte       5 -47.60911
totcons.errorsarlm  4 -49.51117
totcons.lagsarlm    4 -48.25701
totcons.lm          4 -44.83518
totcons.slx         4 -44.15975
```