Projet d'econometrie spatiale

20217765: Omar Saip SY

2024-04-03

SOMMAIRE

Introduction

L'économétrie spatiale permet de comprendre les relations entre les variables économiques et sociales à travers un espace géographique.

Pour arriver à cette compréhension, nous avons décidé d'étudier quelles variables influencent la production d'électricité dans le monde pour l'année 2015.

Notre analyse se concentre sur un ensemble de variables socio-économiques et démographiques, telles que la densité de population, le pourcentage de militaires par rapport à la population active, le PIB/hab, etc.

Nous utiliserons des techniques d'économétrie spatiale pour explorer les relations spatiales et comprendre les dynamiques des différents pays du monde.

I) Jeu de données

```
# A tibble: 6 × 16
  Country_Name Country_Code  Time Electricity_production Access_to_electricity
  <chr>        <chr>        <dbl>                  <dbl>                 <dbl>
1 Albania      ALB           2015                      0                 100.
2 Algeria      DZA           2015                   99.7                  99.4
3 Angola       AGO           2015                   46.8                  42
4 Argentina    ARG           2015                   66.9                  99.7
5 Armenia      ARM           2015                   35.9                  100
6 Australia    AUS           2015                   86.4                  100
# ℹ 11 more variables: Adolescent_fertility_rate <dbl>,
#   Agricultural_land <dbl>, Alternative_and_nuclear_energy <dbl>,
#   Armed_forces_personnel <dbl>, CO2_emissions <dbl>, GDP_per_capita <dbl>,
#   High_technology_exports <dbl>, Population_15_64 <dbl>,
#   Population_density <dbl>, Research_and_development_expenditure <dbl>,
#   Trade <dbl>
```

Ci-dessus, nous pouvons voir les premières lignes de la base de données. Notre jeu de données est composé de 16 colonnes pour 135 individus correspondant aux différents pays du monde. Parmi les 16 colonnes, nous en avons 3 qui représentent le nom du pays, le code du pays et l'année 2015. Nous avons, parmi les 13 colonnes restantes :

• **Electricity_production** : cette variable représente le pourcentage d'électricité produite à partir d'énergies fossiles, de gaz naturel et/ou de charbon par rapport à l'électricité produite totale (en %) (électricité produite à partir de EF, GN et/ou C).

• **Access_to_electricity** : elle représente le pourcentage de la population ayant accès à l'électricité (en %).

• **Adolescent_fertility_rate** : cela représente le nombre de naissances pour 1000 femmes âgées de 15 à 19 ans (en pour mille).

• **Agricultural_land** : elle représente le pourcentage de la superficie du pays réservé à l'agriculture (en %).

• **Alternative_and_nuclear_energy** : c'est le pourcentage d'énergie verte utilisée dans la consommation totale du pays (en %).

• **Armed_forces_personnel** : c'est le pourcentage du personnel de l'armée parmi la population active (en %).

• **CO2_emissions** : représente les émissions de CO2 par habitant (en tonnes/hab).

• **GDP_per_capita** : représente le PIB par habitant (en $/hab).

• **High_technology_exports** : représente le pourcentage de matériaux high-tech exportés parmi les produits manufacturés du pays (en %).

• **Population_15_64** : représente le pourcentage de la population active (en %).

• **Population_density** : représente la densité de la population (en hab/km²).

• **Research_and_development_expenditure** : représente le pourcentage du PIB dédié à la Recherche et au Développement (R&D) (en %).

• **Trade** : représente le pourcentage du PIB dédié aux échanges (imports et exports) entre nations (en %).

Cependant, les données nécessitent un traitement pour pouvoir les analyser plus tard.

Traitement de données

```
data$Electricity_production <- na.fill(data$Electricity_production,0)
data$Agricultural_land <- na.fill(data$Agricultural_land,20)
data[["Alternative_and_nuclear_energy"]] <- na.fill(data[["Alternative_and_nuclear_energy"]],0)
data <- data[which( !(is.na(data$GDP_per_capita)) ),]
data$High_technology_exports <- na.fill(data$High_technology_exports,0)
data$Research_and_development_expenditure <- na.fill(data$Research_and_development_expenditure,0)

data$Access_to_electricity <- na.locf(data$Access_to_electricity)
data$Armed_forces_personnel <- na.locf(data$Armed_forces_personnel)
data[["Trade"]] <- na.locf(data[["Trade"]])
```

Nous nous sommes focalisés sur le traitement de données manquantes. Comme vu ci-dessus, nous avons traité certaines données manquantes en les remplissant par une valeur fixe (arbitraire) car : cela ne concerne que 2 ou 3 pays, ou la donnée manquante pouvait être remplacée par 0.

L'autre moyen de traitement est par répétition de la dernière valeur non nulle.

Jointure entre la carte et nos données

Afin d'avoir des données spatiales, il nous fallait joindre notre jeu de données à une base géographique. Nous l'avons joint avec la base "World" disponible sur RStudio.

Résumé des données

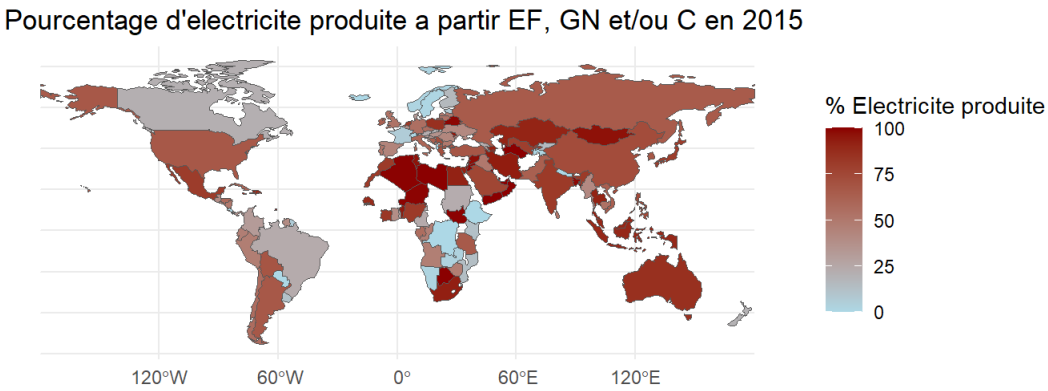
Summary Statistics

Variable	N	Mean	Std. Dev.	Min	Pctl. 25	Pctl. 75	Max
Time	135	2015	0	2015	2015	2015	2015
Electricity_production	135	59	33	0	35	91	100
Access_to_electricity	135	88	23	4.8	90	100	100
Adolescent_fertility_rate	135	44	38	2.9	12	70	177
Agricultural_land	135	40	21	0.56	27	53	83
Alternative_and_nuclear_energy	135	3.9	9.3	0	0	0.47	49
Armed_forces_personnel	135	1.3	1.2	0.0026	0.47	1.7	6.4
CO2_emissions	135	5.1	5.5	0.041	1.2	7.1	35
GDP_per_capita	135	15405	19527	481	2752	19014	105462
High_technology_exports	135	9.5	10	0	1.4	15	52
Population_15_64	135	65	6.3	48	62	68	85
Population_density	135	202	700	1.9	31	138	7807
Research_and_development_expenditure	135	0.72	0.96	0	0	1.1	4.2
Trade	135	84	52	18	51	99	351

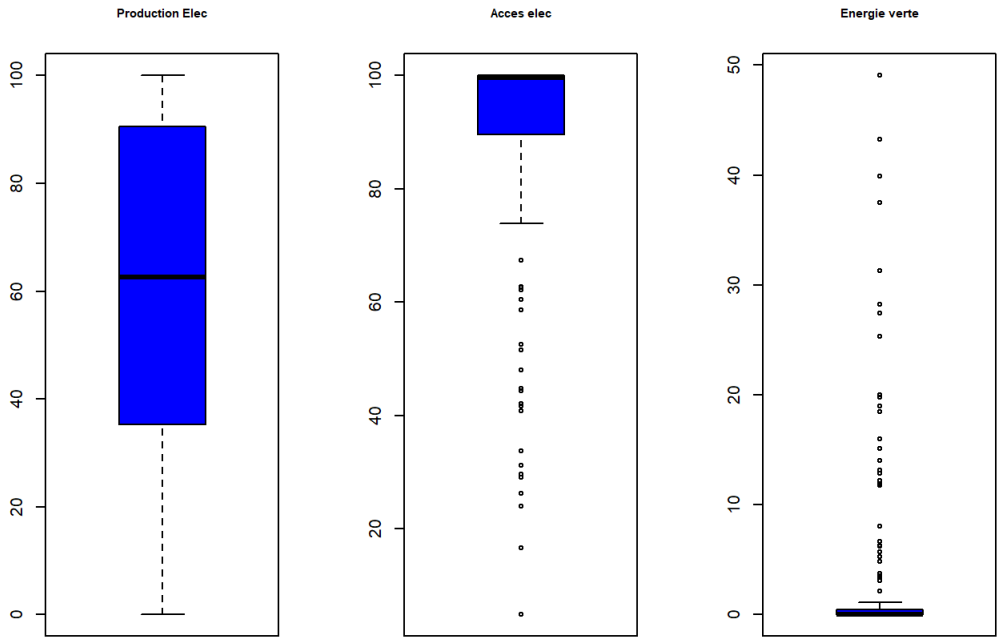
On peut noter que notre jeu de données ne contient pas de valeurs manquantes. Afin de mieux percevoir les distributions des variables, nous emploierons des boxplots et une cartographie pour la variable endogène.

II) Statistiques descriptives univariées

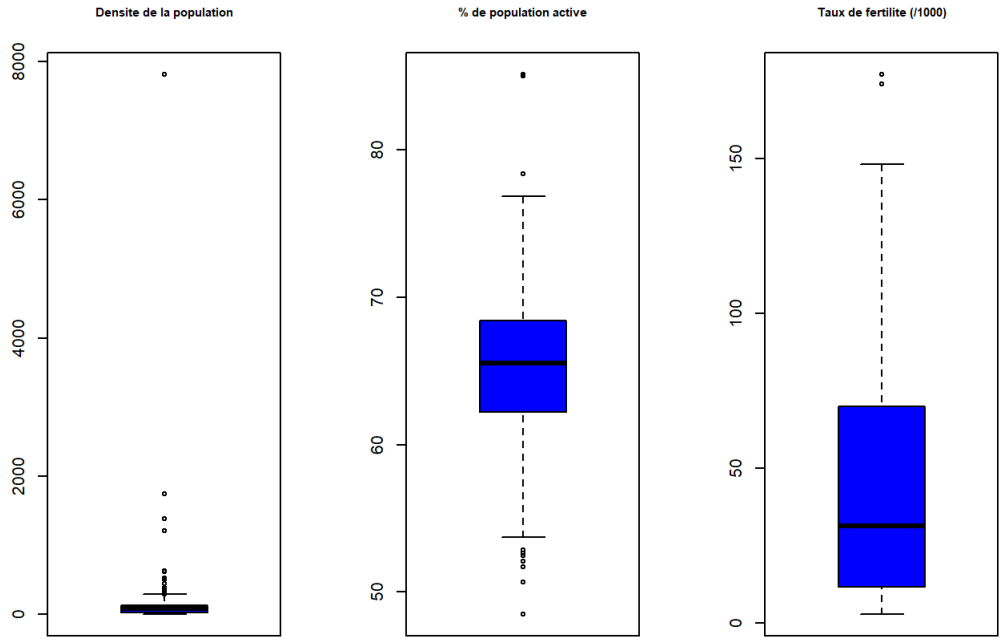
Cartographie du pourcentage d'électricité produite à partir EF, GN et/ou C



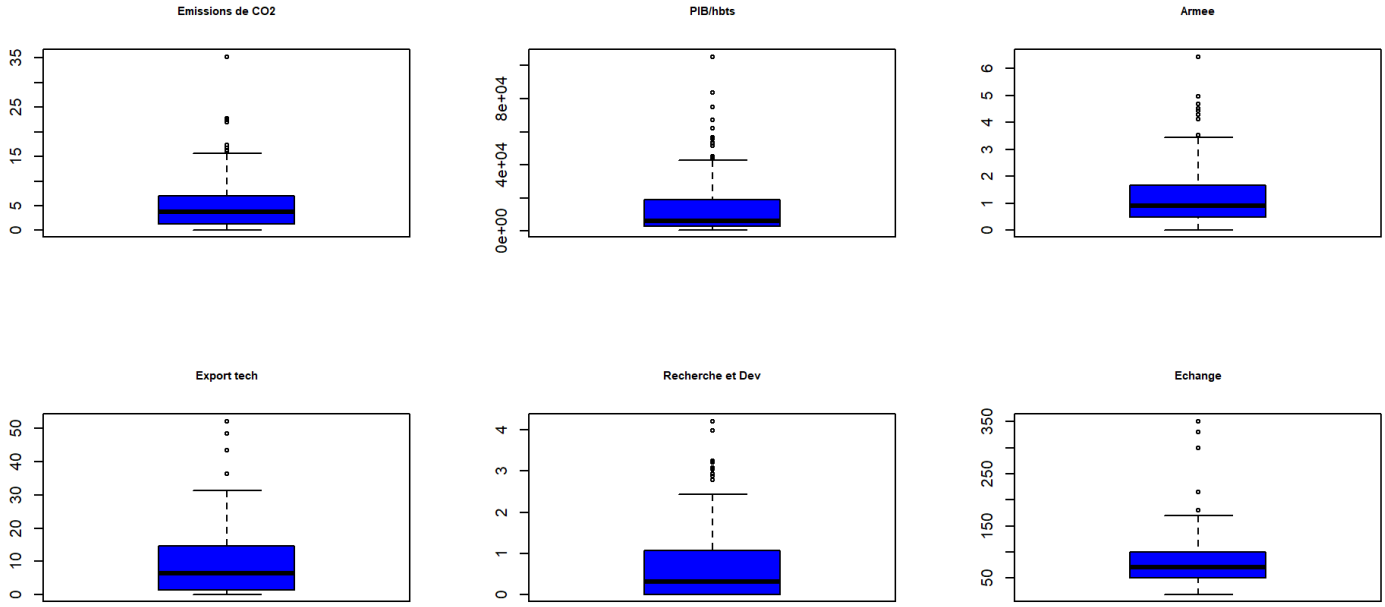
On remarque que la production d'électricité à partir d'énergies fossiles, de gaz naturel et/ou de charbon est plutôt importante en Afrique du Nord, en Afrique du Sud et dans la péninsule Arabique. De manière générale, elle semble plus élevée pour les pays en voie de développement.



La distribution des variables sur les pays est hétérogène, surtout en ce qui concerne l'accès à l'électricité et l'utilisation d'énergie verte.



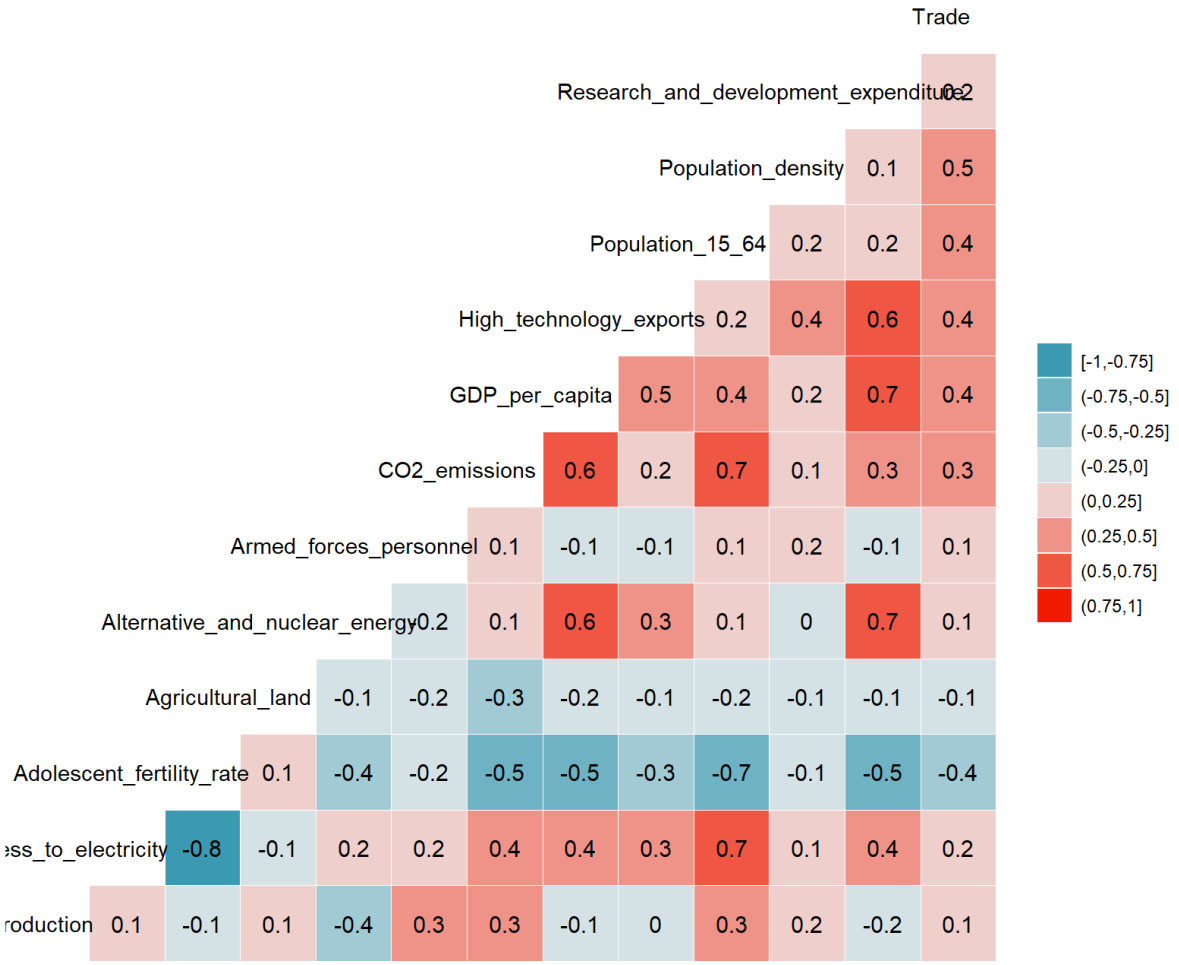
La distribution des variables sur les pays est également hétérogène, surtout en ce qui concerne la densité de la population et le pourcentage de la population âgée entre 15 et 64 ans



On observe quelques valeurs atypiques sur les boxplots, cependant, les variables mettent en évidence les différences entre les pays.

III) Statistiques descriptives bivariées

Matrice de Corrélation



On observe que la variable GDP_per_capita est très fortement corrélée avec les variables CO2_emissions, Alternative_nuclear_energy et Research_and_development_expenditure; ce qui est normal car plus un pays est riche, plus il utilise de l'énergie verte et dépense dans la R&D.

On voit aussi que la fécondité des jeunes filles (âgées de 15 à 19 ans) est synonyme de non accès à l'électricité ; En effet, ces 2 variables sont négativement corrélées.

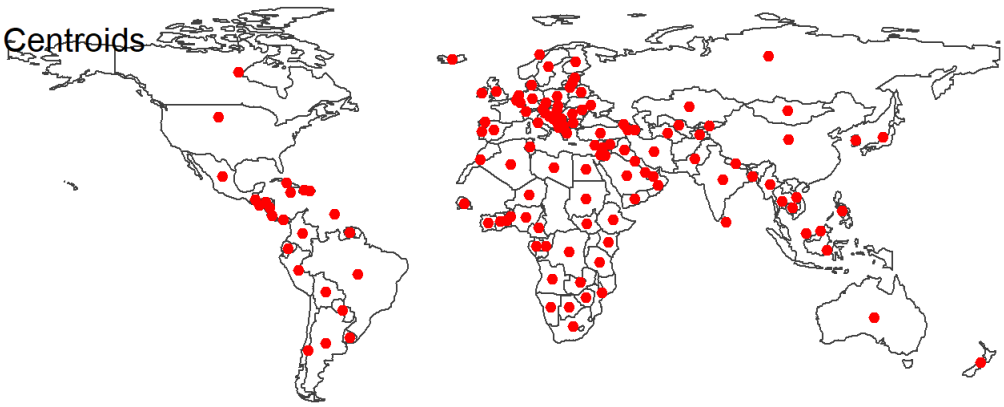
On ne peut pas déduire grand-chose sur la production d'électricité à partir de EF, GN et/ou C grâce à cette matrice. Ce qui est évident vu qu'aucune variable n'est fortement corrélée à la production d'électricité, que ce soit positivement ou négativement.

À ce stade, nous n'avons pas de résultats concluants qui nous permettraient de choisir des variables de contrôle. De ce fait, nous les prendrons toutes puis enlèverons celles non significatives.

IV) Matrice de poids spatiaux

Matrice de continuité

Visualisation des centroïdes de chaque pays



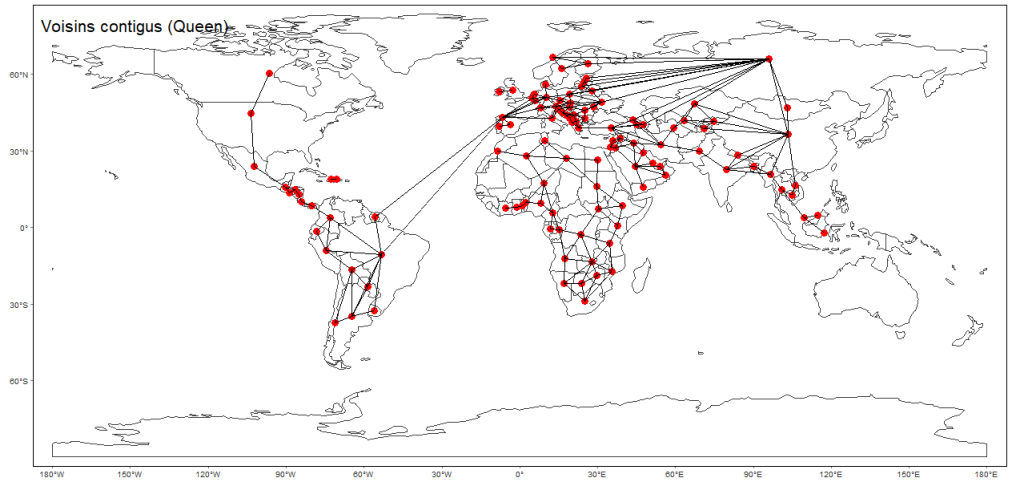

```
Neighbour list object:
Number of regions: 131
Number of nonzero links: 428
Percentage nonzero weights: 2.494027
Average number of links: 3.267176
12 regions with no links:
6 29 30 58 61 63 68 72 92 97 106 119
16 disjoint connected subgraphs
Link number distribution:

 0  1  2  3  4  5  6  7  8  9 13
12 13 30 17 31 13  2  8  2  2  1
13 least connected regions:
17 19 23 33 34 44 51 53 55 76 99 101 128 with 1 link
1 most connected region:
103 with 13 links
```

Nous retrouvons 131 pays sur nos 135 initiaux, avec un total de 428 liens non nuls entre eux. Cela représente environ 2.49 % de poids non nuls, indiquant une connectivité modérée entre les pays.

En moyenne, chaque pays est relié à environ 3.27 autres pays.

On note l'existence de 16 sous-graphes connectés suggérant la présence de 16 groupes distincts.

Représentation des liens entre ces pays sur la carte (contiguité)



La distance géographique du nord du Canada par rapport aux autres pays pourrait se traduire par une connectivité spatiale plus faible dans l'analyse des voisins.

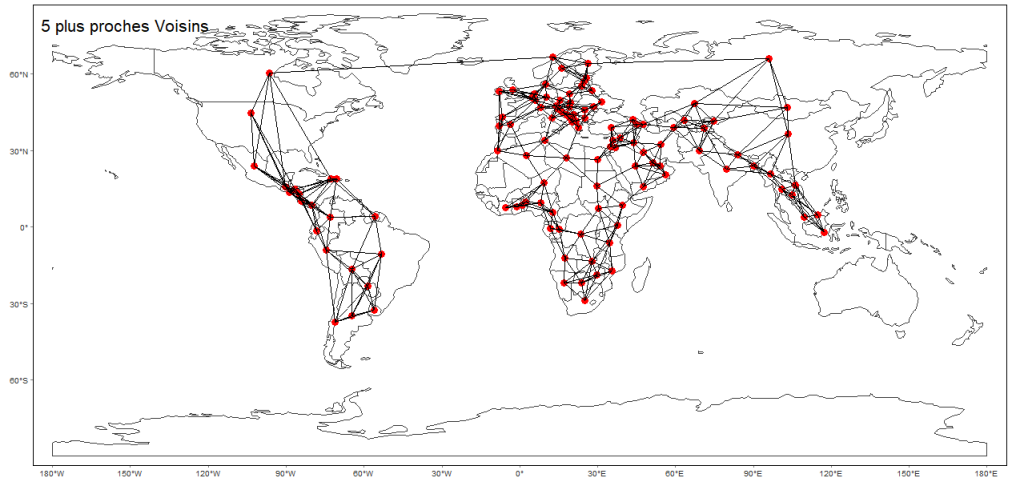
En effet, sa position est isolee du reste du monde.

Matrice des plus proches voisins

Déterminons les plus proches voisins de chaque pays, nous avons choisi d'en sélectionner 5.

```
     [,1] [,2] [,3] [,4] [,5]
[1,]   17   24   25   77  118
[2,]   12   44   71   73   97
[3,]   53   62   84   91   94
[4,]   14   15   20   90  112
[5,]    7   42   54  103  109
[6,]   28   47   49  100  101
```

Représentation des liens entre ces pays sur la carte (5 plus proches voisins)



Matrice de poids sur la distance euclidienne

```
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
135.8   272.3   458.1   572.7   785.3  2166.0 
```

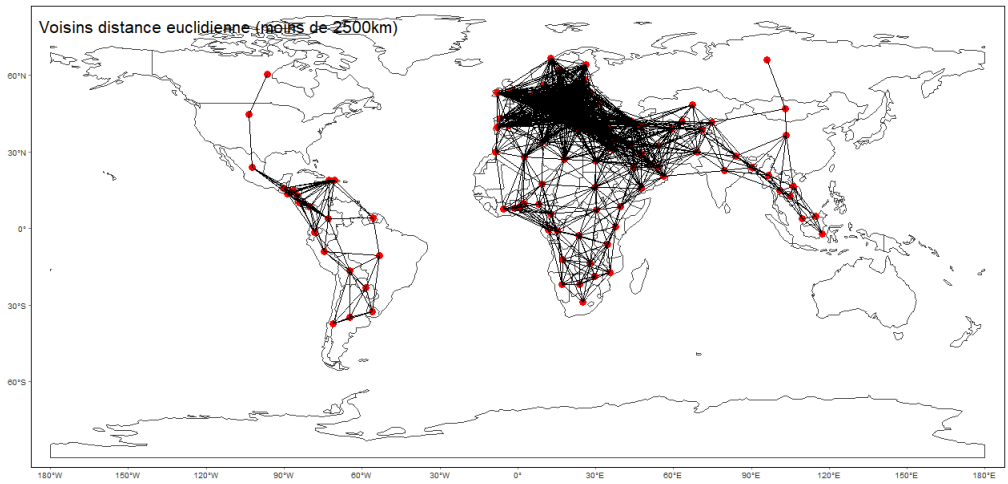
```
Neighbour list object:
Number of regions: 119
Number of nonzero links: 2566
Percentage nonzero weights: 18.12019
Average number of links: 21.56303
2 disjoint connected subgraphs
```

Pour 2500km on peut voir que chaque point ne compte pas le même nombre de voisin. C'est assez variable.

On note au total 2566 liens entre 119 pays. En moyenne chaque pays a 21.56 voisins.

Le poids de ces liens est d'environ 18.12%.

Représentation des liens entre ces pays sur la carte (distance euclidienne)



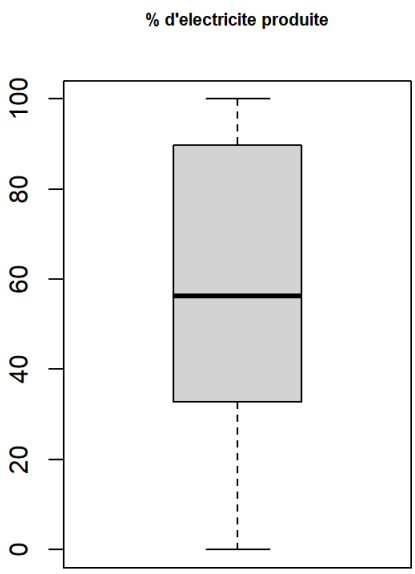
Nous observons qu'il y a plus de liens pour la distance euclidienne (voisins de moins de 2500 km).

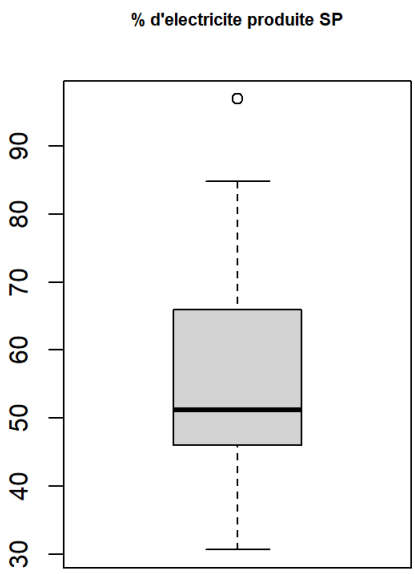
De ce fait, nous continuerons avec la distance euclidienne dans la suite de notre étude.

V) Analyse de l'autocorrélation spatiale

Création du spatial lag avec la matrice de continuité normalisée

Comparaison sous forme de boxplot du pourcentage d'électricité produite à partir de EF, GN et/ou C et le spatial lag de cette variable.





Le spatial lag représente la moyenne pondérée des valeurs de la variable (pourcentage d'électricité produite à partir de EF, GN et/ou C dans notre cas) dans les pays voisins, reflétant ainsi les interactions spatiales et l'influence de l'environnement géographique sur les valeurs de la variable.

Indice de Moran (matrice de distance euclidienne)

```
Call:
lm(formula = wx_prod_elec ~ x_prod_elec)

Residuals:
    Min      1Q  Median      3Q     Max
-26.560  -9.250  -2.485  10.126  39.736

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 47.77335    2.45013  19.498  < 2e-16 ***
x_prod_elec  0.14366    0.03755   3.826 0.000211 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.31 on 117 degrees of freedom
Multiple R-squared:  0.1112,    Adjusted R-squared:  0.1036
F-statistic: 14.64 on 1 and 117 DF,  p-value: 0.0002109
```

La régression montre une relation positive entre la production d'électricité à partir de EF, GN et/ou C dans un pays et celle dans les pays voisins (le coefficient de x_prod_elec est très significatif dans le modèle).

Le R^2 multiple est de 0.11, ce qui signifie que le modèle explique environ 11.12 % de la variabilité de la variable dépendante.

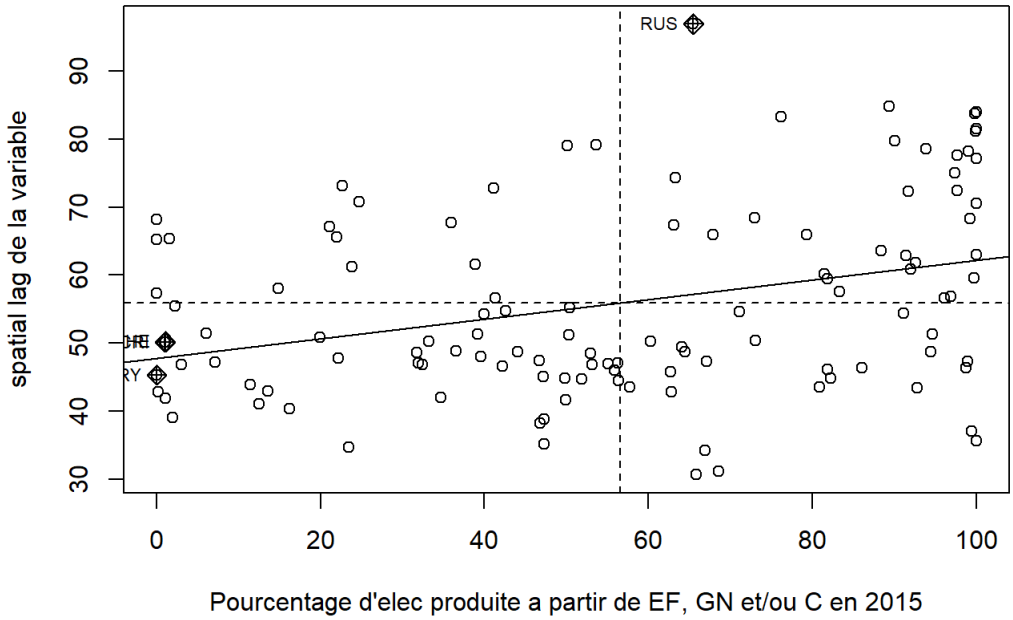
La F-statistique est de 14.64 avec un p-value de 2.1e-04, ce qui indique que le modèle est statistiquement significatif.

L'analyse montre alors que la production d'électricité à partir de EF, GN et/ou C d'un pays est très liée à celle des nations avoisinantes.

```
    Moran I test under normality

data:  data_spatiale$Electricity_production
weights: nb2500km.ids.w

Moran I statistic standard deviate = 3.6119, p-value = 0.0003039
alternative hypothesis: two.sided
sample estimates:
Moran I statistic       Expectation          Variance
      0.143663078      -0.008474576       0.001774185
```

La statistique de Moran I est de 3.611, ce qui implique une autocorrélation spatiale significative dans les données.

De plus, la p-value de 3e-04 est inférieure au seuil de 0.05, ce qui confirme l'existence d'une autocorrélation spatiale de la production d'électricité à partir de EF, GN et/ou C entre les pays.



Dans la partie superieure droite du nuage de points de Moran, on remarque une autocorrélation spatiale positive pour la Russie (RUS).

```
   Monte-Carlo simulation of Moran I

data:  data_spatiale$Electricity_production
weights: nb2500km.ids.w
number of simulations + 1: 100

statistic = 0.14366, observed rank = 100, p-value < 2.2e-16
alternative hypothesis: two.sided
```

Dans le cadre de la simulation de Monte-Carlo de Moran I, la statistique observée est comparée à une distribution nulle obtenue par des itérations de simulations aléatoires sous l'hypothèse nulle d'absence d'autocorrélation spatiale.

Avec un p-value < 2.2e-16, inférieur au seuil de significativite de 0.05, nous rejetons l'hypothèse nulle d'absence d'autocorrélation spatiale.

Cela indique qu'il existe une autocorrélation spatiale significative dans les données de la production d'électricité à partir de EF, GN et/ou C.

LISA(les indicateurs locaux d'autocorrélation spatiale)

Données non centrées

```
       Ii                E.Ii               Var.Ii              Z.Ii         
 Min.   :-0.85923   Min.   :-2.571e-02   Min.   :0.0000109   Min.   :-3.6476  
 1st Qu.:-0.06864   1st Qu.:-1.444e-02   1st Qu.:0.0087154   1st Qu.:-0.6185  
 Median : 0.08638   Median :-6.988e-03   Median :0.0529499   Median : 0.7864  
 Mean   : 0.14366   Mean   :-8.475e-03   Mean   :0.1024673   Mean   : 0.5578  
 3rd Qu.: 0.33785   3rd Qu.:-9.517e-04   3rd Qu.:0.1247344   3rd Qu.: 1.5601  
 Max.   : 1.12551   Max.   :-2.820e-07   Max.   :1.1933003   Max.   : 3.6990  
 Pr(z != E(Ii))
 Min.   :0.0002164
 1st Qu.:0.0835758
 Median :0.2130140
 Mean   :0.2998051
 3rd Qu.:0.4349847
 Max.   :0.9943553
```

Données centrées

```
      Ii               E.Ii             Var.Ii            Z.Ii
Min.   :-0.85923   Min.   :-2.571e-02   Min.   :0.0000109   Min.   :-3.6476
1st Qu.:-0.06864   1st Qu.:-1.444e-02   1st Qu.:0.0087154   1st Qu.:-0.6185
Median : 0.08638   Median :-6.988e-03   Median :0.0529499   Median : 0.7864
Mean   : 0.14366   Mean   :-8.475e-03   Mean   :0.1024673   Mean   : 0.5578
3rd Qu.: 0.33785   3rd Qu.:-9.517e-04   3rd Qu.:0.1247344   3rd Qu.: 1.5601
Max.   : 1.12551   Max.   :-2.820e-07   Max.   :1.1933003   Max.   : 3.6990
Pr(z != E(Ii))
Min.   :0.0002164
1st Qu.:0.0835758
Median :0.2130140
Mean   :0.2998051
3rd Qu.:0.4349847
Max.   :0.9943553
```

Les valeurs de “Ii” représentent les indices de Moran locaux pour chaque observation. Ils mesurent l’autocorrélation spatiale locale pour chaque pays.

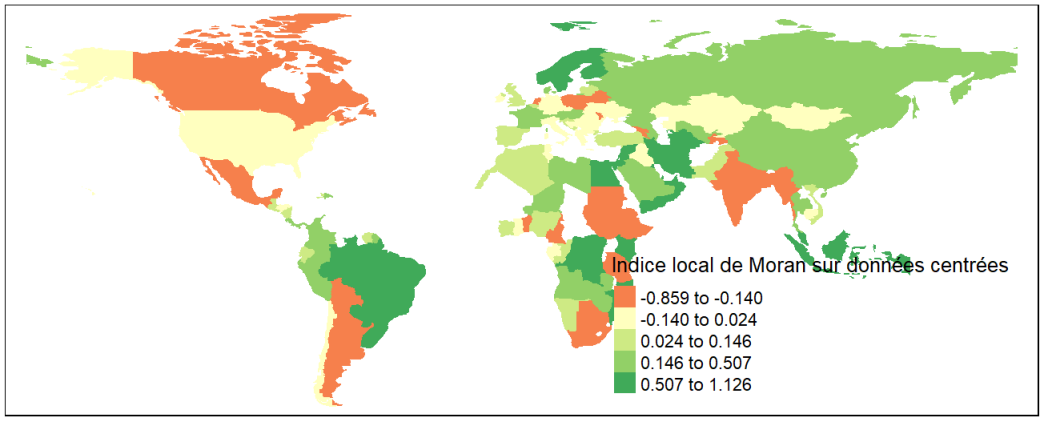
Une valeur positive indique une autocorrélation spatiale positive locale, ce qui signifie que les pays avec des pourcentages d’électricité produite à partir de EF, GN et/ou C similaires ont tendance à être proches.

Tandis qu’une valeur négative indique une autocorrélation spatiale négative locale, indiquant un regroupement de valeurs opposées dans l’espace.

Les colonnes “E.Ii” et “Var.Ii” représentent l’espérance et la variance de la distribution des indices de Moran locaux. La colonne “Z.Ii” présente les valeurs z-standardisées des indices de Moran locaux. Une valeur z élevée (positivement ou négativement) indique une forte autocorrélation spatiale locale.

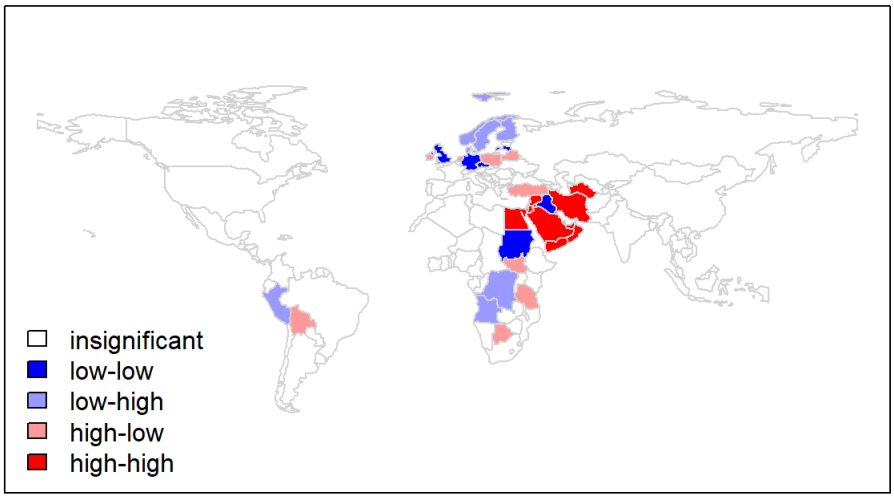
La colonne “Pr(z != E(Ii))” fournit les p-values associées à chaque indice de Moran local. Ils testent si les valeurs observées des indices de Moran locaux sont significativement différentes de leur espérance sous l’hypothèse nulle d’absence d’autocorrélation spatiale locale.

Cartographie des indices locaux de Moran sur données centrées



Cette cartographie nous montre que :

- Le pourcentage d’électricité produite à partir de EF, GN et/ou C, pour un pays, a une tendance inverse de celui des pays voisins en Amérique du Nord (Canada), Une partie de l’Amérique du Sud (sud-ouest), l’Afrique du sud, le Soudan;
- Le pourcentage d’électricité produite à partir de EF, GN et/ou C, pour un pays, a la même tendance que celui des pays voisins en Europe, une partie de l’Amérique du Sud (nord, centre et nord-est), en Afrique centrale, le Maghreb, en Océanie et la péninsule Arabique et en Asie;
- Le pourcentage d’électricité produite à partir de EF, GN et/ou C, pour un pays, ne dépend pas de celui des pays voisins aux États-Unis.



On centre les indices locaux par rapport à leur moyenne et on définit un niveau de significativité (ici = 10%).

Par conséquent, sur cette cartographie, nous avons mis en place un système qui nous permet de filtrer les pays. En effet :

- insignificant : représente les pays dont la p-value associée à l'indice de Moran est supérieure à 10%. On considère qu'il y a une absence d'autocorrélation spatiale.
- low-low : ce sont les pays dont le pourcentage d'électricité produite à partir de EF, GN et/ou C est inférieur à la moyenne et idem pour l'indice de Moran. Ce sont les pays avec un pourcentage d'électricité produite faible et cette production d'électricité est inverse à celle des voisins.
- low-high : ce sont les pays dont le pourcentage d'électricité produite à partir de EF, GN et/ou C est inférieur à la moyenne et l'indice de Moran est supérieur à la moyenne des indices de Moran. Ce sont les pays avec un pourcentage d'électricité produite faible et cette production d'électricité a la même tendance que celle des voisins.
- high-low : ce sont les pays dont le pourcentage d'électricité produite à partir de EF, GN et/ou C est supérieur à la moyenne et l'indice de Moran est inférieur à la moyenne des indices de Moran. Ce sont les pays avec un pourcentage d'électricité produite élevé et cette production d'électricité est inverse à celle des voisins.
- high-high : ce sont les pays dont le pourcentage d'électricité produite à partir de EF, GN et/ou C est supérieur à la moyenne et idem pour l'indice de Moran. Ce sont les pays avec un pourcentage d'électricité produite élevé et cette production d'électricité a la même tendance que celle des voisins.

VI) Estimation des modèles

OLS

OLS (Ordinary Least Squares) est une méthode de régression linéaire qui a une approche d'estimation des paramètres d'un modèle de régression linéaire en minimisant la somme des carrés des résidus.

```
Call:
lm(formula = equation, data = data_spatiale)

Residuals:
    Min      1Q  Median      3Q     Max
-57.796 -13.870   2.869  17.415  61.086

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
Agricultural_land                        0.3023716  0.1260274   2.399 0.018141 *
Alternative_and_nuclear_energy          -0.8067376  0.4007296  -2.013 0.046584 *
Armed_forces_personnel                   7.6265737  2.2378814   3.408 0.000921 ***
CO2_emissions                            2.9362465  0.7516783   3.906 0.000164 ***
GDP_per_capita                          -0.0003741  0.0002490  -1.502 0.135996
Population_density                       0.0448875  0.0173715   2.584 0.011102 *
Research_and_development_expenditure     0.1215128  4.5030026   0.027 0.978522
Access_to_electricity                   -0.1069901  0.1781431  -0.601 0.549374
Adolescent_fertility_rate                0.0149363  0.0958659   0.156 0.876479
Population_15_64                         0.5257465  0.3478457   1.511 0.133598
Trade                                   -0.0229812  0.0636702  -0.361 0.718849
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.1 on 108 degrees of freedom
Multiple R-squared:  0.8548,    Adjusted R-squared:   0.84
F-statistic: 57.79 on 11 and 108 DF,  p-value: < 2.2e-16
```

En analysant les résultats de ce modèle, nous remarquons qu'il y a beaucoup de variables non significatives.

De ce fait, nous diminuerons les variables explicatives et ne garderons que celles avec les meilleures p-value.

```
Call:
lm(formula = equation, data = data_spatiale)

Residuals:
    Min      1Q  Median      3Q     Max
-55.744 -14.355   2.305  19.601  74.724

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
Agricultural_land              0.6115876  0.0821760   7.442 2.09e-11 ***
Alternative_and_nuclear_energy -0.6364516  0.3526389  -1.805  0.07377 .
Armed_forces_personnel         10.5822865  1.9461173   5.438 3.16e-07 ***
CO2_emissions                   3.7918934  0.6092427   6.224 8.47e-09 ***
GDP_per_capita                 -0.0004235  0.0002171  -1.951  0.05355 .
Population_density              0.0479577  0.0175777   2.728  0.00738 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.82 on 113 degrees of freedom
Multiple R-squared:  0.8396,    Adjusted R-squared:  0.831
F-statistic: 98.55 on 6 and 113 DF,  p-value: < 2.2e-16
```

Notre modèle ne contient dorénavant que des variables significatives au niveau 10%.

- "Agricultural_land" : Le coefficient est de 0.61. Ce qui signifie que si le pourcentage de terre agricole augmente d'1% pour un pays, alors le pourcentage d'électricité produite à partir de EF, GN et/ou C augmente de 0.61%. C'est une relation positive.
- "Alternative_and_nuclear_energy" : Le coefficient est de -0.64. Ce qui signifie que si le pourcentage d'énergie verte consommée augmente d'1% pour un pays, alors le pourcentage d'électricité produite à partir de EF, GN et/ou C diminue de 0.64%. C'est une relation négative.
- "Armed_forces_personnel" : Le coefficient est de 10.58 Ce qui signifie que si le pourcentage de militaires dans la population active augmente d'1% pour un pays, alors le pourcentage d'électricité produite à partir de EF, GN et/ou C augmente de 10.58%. C'est une relation positive et forte.
- "CO2_emissions" : Le coefficient est de 3.79. Ce qui signifie que si les émissions de CO2 augmentent d'1 tonne/hbts pour un pays, alors le pourcentage d'électricité produite à partir de EF, GN et/ou C augmente de 3.79%. C'est une relation positive.
- "GDP_per_capita" : Le coefficient est de -0.0004. Ce qui signifie que si le PIB augmente d'1 $/hbts pour un pays, alors le pourcentage d'électricité produite à partir de EF, GN et/ou C diminue de 0.0004%. C'est une relation négative et faible.
- "Population_density" : Le coefficient est de 0.048. Ce qui signifie que si la densité de la population augmente d'1 hbts/km2 pour un pays, alors le pourcentage d'électricité produite à partir de EF, GN et/ou C augmente de 0.048%. C'est une relation positive.

Le R-carré est de 83.96%, indiquant que le modèle explique environ 83.96% de la variance de la variable dépendante. C'est un très bon ajustement pour les données de production, où une grande variabilité est souvent observée.

La F-statistique est de 98.55, ce qui est significatif (p < 2.2e-16), indiquant que le modèle est statistiquement significatif et que les variables explicatives, dans leur ensemble, ont un effet significatif sur la variable à expliquer.

Test d'hétéroscedasticité

```
    studentized Breusch-Pagan test

data:  prod_elecs_OLS
BP = 11.446, df = 5, p-value = 0.04321
```

Le test de Breusch-Pagan s'avere significatif. Donc il y a bien une hétéroscedasticité.

De ce fait nous corrigerons ce probleme d'hétéroscedasticité en appliquant la correction de White.

correction de l'hétéroscedasticité

```
Call:
lm(formula = equation, data = data_spatiale, weights = 1/e2chap)

Weighted Residuals:
    Min      1Q  Median      3Q     Max
-11.5469  -2.4592  -0.3939   1.3428  15.5646

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
Agricultural_land              0.6227825  0.1232522   5.053 1.69e-06 ***
Alternative_and_nuclear_energy -0.4347995  0.3064690  -1.419   0.1587
Armed_forces_personnel         20.6863018  2.6170631   7.904 1.96e-12 ***
CO2_emissions                  3.1169448  0.4690662   6.645 1.11e-09 ***
GDP_per_capita                -0.0003934  0.0002200  -1.788   0.0765 .
Population_density             0.0397198  0.0227056   1.749   0.0829 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.315 on 113 degrees of freedom
Multiple R-squared:  0.835, Adjusted R-squared:  0.8262
F-statistic: 95.28 on 6 and 113 DF,  p-value: < 2.2e-16
```

```
    studentized Breusch-Pagan test

data:  prod_elecs_OLS
BP = 0.50553, df = 5, p-value = 0.9919
```

Le test de Breusch-Pagan n'est pas significatif. Donc il y a bien une homoroscedasticité.

De ce fait, notre modèle est un peu différent du précédent.

- "Agricultural_land" : Le coefficient est de 0.62. Ce qui signifie que si le pourcentage de terre agricole augmente d'1% pour un pays, alors le pourcentage d'électricité produite à partir de EF, GN et/ou C augmente de 0.62%. C'est une relation positive.
- "Alternative_and_nuclear_energy" : Le coefficient est de -0.43 Ce qui signifie que si le pourcentage d'énergie verte consommée augmente d'1% pour un pays, alors le pourcentage d'électricité produite à partir de EF, GN et/ou C diminue de 0.43%. C'est une relation négative.
- "Armed_forces_personnel" : Le coefficient est de 20.69 Ce qui signifie que si le pourcentage de militaires dans la population active augmente d'1% pour un pays, alors le pourcentage d'électricité produite à partir de EF, GN et/ou C augmente de 20.69%. C'est une relation positive et forte.
- "CO2_emissions" : Le coefficient est de 3.12 Ce qui signifie que si les émissions de CO2 augmentent d'1 tonne/hbts pour un pays, alors le pourcentage d'électricité produite à partir de EF, GN et/ou C augmente de 3.12%. C'est une relation positive.
- "GDP_per_capita" : Le coefficient est de -0.0004. Ce qui signifie que si le PIB augmente d'1 $/hbts pour un pays, alors le pourcentage d'électricité produite à partir de EF, GN et/ou C diminue de 0.0004%. C'est une relation négative et faible.
- "Population_density" : Le coefficient est de 0.0397 Ce qui signifie que si la densité de la population augmente d'1 hbts/km2 pour un pays, alors le pourcentage d'électricité produite à partir de EF, GN et/ou C augmente de 0.0397%. C'est une relation positive.

Le R-carré est de 83.5%, indiquant que le modèle explique environ 83.5% de la variance de la variable dépendante. C'est un très bon ajustement pour les données de production, où une grande variabilité est souvent observée.

La F-statistique est de 95.28, ce qui est significatif (p < 2.2e-16), indiquant que le modèle est statistiquement significatif et que les variables explicatives, dans leur ensemble, ont un effet significatif sur la variable à expliquer.

Autocorrélation spatiale

```
    Global Moran I for regression residuals

data:
model: lm(formula = equation, data = data_spatiale, weights = 1/e2chap)
weights: PPV2.w

Moran I statistic standard deviate = 1.1816, p-value = 0.1187
alternative hypothesis: greater
sample estimates:
Observed Moran I      Expectation         Variance
     0.041372629     -0.018845626      0.002597115
```

On note la présence d'autocorrélation spatiale positive, significative au niveau 15% (p-value=0.119 < 0.15).

SLX

Le modèle spatial de SLX (Spatial Lag Cross-Regression) est une extension du modèle de régression spatiale qui prend en compte à la fois l'effet spatial de la variable dépendante et l'effet spatial des variables explicatives. La formule générale du modèle SLX peut être écrite comme suit :

$$Y = \rho WY + X\beta + \varepsilon$$

où :

Y est le vecteur des observations de la variable dépendante.

ρ est le coefficient de l'effet spatial de la variable dépendante.

W est une matrice de pondération spatiale représentant la relation spatiale entre les observations.

$X\beta$ est le produit matriciel des variables explicatives et de leurs coefficients. ε est le terme d'erreur.

```
Call:
lm(formula = Electricity_production ~ Agricultural_land + W_Agricultural_land +
    Alternative_and_nuclear_energy + W_Alternative_and_nuclear_energy +
    Armed_forces_personnel + W_Armed_forces_personnel + CO2_emissions +
    W_CO2_emissions + Population_density + W_Population_density -
    1, data = data_spatiale)

Residuals:
    Min      1Q  Median      3Q     Max
-54.581 -16.253   1.553  16.590  60.304

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
Agricultural_land                0.3232836  0.1288967   2.508 0.013612 *
W_Agricultural_land              0.3173453  0.1932628   1.642 0.103464
Alternative_and_nuclear_energy  -1.0825446  0.3438177  -3.149 0.002117 **
W_Alternative_and_nuclear_energy -0.3233146  0.7194191  -0.449 0.654028
Armed_forces_personnel           5.8069205  2.4760718   2.345 0.020826 *
W_Armed_forces_personnel        10.8423642  6.1334466   1.768 0.079902 .
CO2_emissions                    2.3316545  0.6055647   3.850 0.000199 ***
W_CO2_emissions                 -0.3026784  1.4168233  -0.214 0.831233
Population_density               0.0342019  0.0190209   1.798 0.074925 .
W_Population_density            -0.0006032  0.0405503  -0.015 0.988158
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.25 on 109 degrees of freedom
Multiple R-squared:  0.8518,    Adjusted R-squared:  0.8382
F-statistic: 62.64 on 10 and 109 DF,  p-value: < 2.2e-16
```

```
   Global Moran I for regression residuals

data:
model: lm(formula = Electricity_production ~ Agricultural_land +
W_Agricultural_land + Alternative_and_nuclear_energy +
W_Alternative_and_nuclear_energy + Armed_forces_personnel +
W_Armed_forces_personnel + CO2_emissions + W_CO2_emissions +
Population_density + W_Population_density - 1, data = data_spatiale)
weights: nb2500km.ids.w

Moran I statistic standard deviate = 0.81579, p-value = 0.2073
alternative hypothesis: greater
sample estimates:
Observed Moran I      Expectation         Variance
    0.002937758     -0.029156680      0.001547739
```

Ce modèle n'est pas mieux que celui OLS car on a ici, aucun retard de significatif.

Cependant, on note un R2 de 85.18%, ce qui est plutot un bon signe.

Malheureusement, on note une absence d'autocorrélation spatiale positive.

SAR

Le modèle spatial autorégressif est un modèle qui incorpore l'autorégression spatiale dans le modèle. Voici la formule générale du modèle SAR :

$$Y = \rho WY + X\beta + \varepsilon$$

où :

Y est le vecteur des observations de la variable dépendante. ρ est le coefficient de l'autorégression spatiale de la variable dépendante. W est une matrice de pondération spatiale représentant la relation spatiale entre les observations. $X\beta$ est le produit matriciel des variables explicatives et de leurs coefficients. ε est le terme d'erreur.

```
Call:lagsarlm(formula = equation, data = data_spatiale, listw = nb2500km.ids.w)

Residuals:
    Min      1Q  Median      3Q     Max
-56.615 -13.538   3.451  19.978  61.876

Type: lag
Coefficients: (asymptotic standard errors)
                                Estimate  Std. Error z value  Pr(>|z|)
Agricultural_land               0.43651292  0.09755368  4.4746 7.656e-06
Alternative_and_nuclear_energy -0.73888197  0.33426270 -2.2105 0.0270717
Armed_forces_personnel          7.58111680  2.09152644  3.6247 0.0002893
CO2_emissions                   3.01721690  0.63707323  4.7361 2.179e-06
GDP_per_capita                 -0.00034512  0.00020665 -1.6701 0.0949010
Population_density              0.04296552  0.01670649  2.5718 0.0101175

Rho: 0.28841, LR test value: 7.034, p-value: 0.0079979
Asymptotic standard error: 0.098434
    z-value: 2.93, p-value: 0.00339
Wald statistic: 8.5847, p-value: 0.00339

Log likelihood: -553.6617 for lag model
ML residual variance (sigma squared): 637.42, (sigma: 25.247)
Number of observations: 119
Number of parameters estimated: 8
AIC: NA (not available for weighted model), (AIC for lm: 1128.4)
LM test for residual autocorrelation
test value: 1.9198, p-value: 0.16588
```

Toutes les variables de ce modèle sont significatives au niveau 5%, sauf "GDP_per_capita".

Le pourcentage de terre agricole, le pourcentage de militaires parmi la population active, les émissions de CO2 et la densité de la population ont un effet positif sur la production d'électricité à partir de EF, GN et/ou C.

Tandis que le pourcentage d'énergie verte consommée et le PIB/hbts ont un effet négatif sur la production d'électricité à partir de EF, GN et/ou C.

Rho est le coefficient d'autocorrelation spatiale pour les erreurs. Il a une valeur de 0.288 avec une p-value significative au niveau 5% (p=0.008 > 0.05), indiquant que l'autocorrelation spatiale est bien présente dans le modele.

SEM

Le modèle d'erreur spatiale (SEM - Spatial Error Model) est un modèle de régression spatiale qui incorpore l'autocorrélation spatiale dans les erreurs du modèle. Voici la formule générale du modèle SEM :

$$Y = X\beta + \varepsilon\varepsilon = \lambda W\varepsilon + U$$

où : Y est le vecteur des observations de la variable dépendante. X est une matrice des variables explicatives. β est un vecteur de coefficients des variables explicatives. ε est le terme d'erreur. λ est le coefficient d'autocorrélation spatiale de l'erreur. W est une matrice de pondération spatiale représentant la relation spatiale entre les observations. U est un terme d'erreur spatiale.

```
Call:
errorsarlm(formula = equation, data = data_spatiale, listw = nb2500km.ids.w)

Residuals:
    Min      1Q  Median      3Q     Max
-55.7255 -14.3661   2.2905  19.6457  74.4382

Type: error
Coefficients: (asymptotic standard errors)
                                Estimate  Std. Error z value  Pr(>|z|)
Agricultural_land               0.61072728  0.08049846  7.5868 3.286e-14
Alternative_and_nuclear_energy -0.63414207  0.34385991 -1.8442  0.065156
Armed_forces_personnel          10.56345873  1.90161083  5.5550 2.776e-08
CO2_emissions                    3.79272973  0.59559205  6.3680 1.915e-10
GDP_per_capita                  -0.00042363  0.00021190 -1.9992  0.045587
Population_density               0.04798943  0.01717700  2.7938  0.005209

Lambda: 0.015794, LR test value: 0.0058228, p-value: 0.93917
Asymptotic standard error: 0.19038
    z-value: 0.082957, p-value: 0.93389
Wald statistic: 0.0068819, p-value: 0.93389

Log likelihood: -557.1757 for error model
ML residual variance (sigma squared): 682.93, (sigma: 26.133)
Number of observations: 119
Number of parameters estimated: 8
AIC: 1130.4, (AIC for lm: 1128.4)
```

Pour ce modele, toutes les variables ne sont pas significatives. Nous interpreterons alors que les variables significatives au niveau 5%

Le pourcentage de militaires parmi la population active, Les emissions de CO2 et la densite de la population ont un effet positif sur la production d'electricite a partir de EF, GN et/ou C.

Tandis que le pourcentage d'energie verte consomme a un effet negatif sur la production d'electricite a partir de EF, GN et/ou C.

Lambda est le coefficient d'autocorrelation spatiale pour les erreurs. Il a une valeur de 0.016 avec une p-value non significative (p=0.93 > 0.05), indiquant que l'autocorrelation spatiale n'est pas presente pour les erreurs dans le modele.

La valeur de Wald pour Lambda n'est pas significative (p=0.89 > 0.05), nous montrant encore une fois l'absence d'autocorrelation spatiale des erreurs du modele.

SDM

Le modèle de dépendance spatiale simultanée (SDM - Simultaneous Spatial Dependence Model) est un modèle de régression spatiale qui incorpore à la fois l'autorégression spatiale de la variable dépendante et l'autorégression spatiale des variables explicatives. Voici la formule générale du modèle SDM :

$$Y = \rho WY + X\beta + WX\delta + \varepsilon$$

où : Y est le vecteur des observations de la variable dépendante.

ρ est le coefficient d'autorégression spatiale de la variable dépendante.

W est une matrice de pondération spatiale représentant la relation spatiale entre les observations.

X est une matrice des variables explicatives.

β est un vecteur de coefficients des variables explicatives. δ est un vecteur de coefficients de l'autorégression spatiale des variables explicatives.

ε est le terme d'erreur.

```
Call:
lagsarlm(formula = equation, data = data_spatiale, listw = nb2500km.ids.w,
    Durbin = T)

Residuals:
     Min       1Q   Median       3Q      Max
-56.1547 -16.4447   2.2762  16.9609  61.1577

Type: mixed
Coefficients: (asymptotic standard errors)
                                  Estimate  Std. Error z value  Pr(>|z|)
Agricultural_land                0.32260676  0.12319105  2.6188  0.008825
Alternative_and_nuclear_energy    -0.85766838  0.35953189 -2.3855  0.017055
Armed_forces_personnel            5.76200077  2.35784840  2.4438  0.014535
CO2_emissions                     3.00941579  0.72637148  4.1431 3.427e-05
GDP_per_capita                   -0.00035747  0.00023077 -1.5490  0.121370
Population_density                0.03836117  0.01844897  2.0793  0.037589
lag.Agricultural_land             0.29649654  0.22013300  1.3469  0.178013
lag.Alternative_and_nuclear_energy -0.15422545  1.32628603 -0.1163  0.907428
lag.Armed_forces_personnel         9.48152647  6.33302057  1.4972  0.134352
lag.CO2_emissions                 -0.69173772  2.09832035 -0.3297  0.741655
lag.GDP_per_capita                 0.00019134  0.00081942  0.2335  0.815366
lag.Population_density             0.00358770  0.03971518  0.0903  0.928020

Rho: 0.021748, LR test value: 0.011998, p-value: 0.91278
Asymptotic standard error: 0.18583
    z-value: 0.11703, p-value: 0.90683
Wald statistic: 0.013697, p-value: 0.90683

Log likelihood: -551.2545 for mixed model
ML residual variance (sigma squared): 618.23, (sigma: 24.864)
Number of observations: 119
Number of parameters estimated: 14
AIC: NA (not available for weighted model), (AIC for lm: 1128.5)
LM test for residual autocorrelation
test value: 0.18814, p-value: 0.66447
```

Agricultural_land, Alternative_and_nuclear_energy, Armed_forces_personnel, CO2_emissions, GDP_per_capita et Population_density sont les variables explicatives directes.

lag.Agricultural_land, lag.Alternative_and_nuclear_energy, lag.Armed_forces_personnel, lag.CO2_emissions, lag.GDP_per_capita et lag.Population_density sont les retards spatiaux des variables explicatives, qui mesurent l'impact des valeurs des pays voisins sur celui d'intérêt.

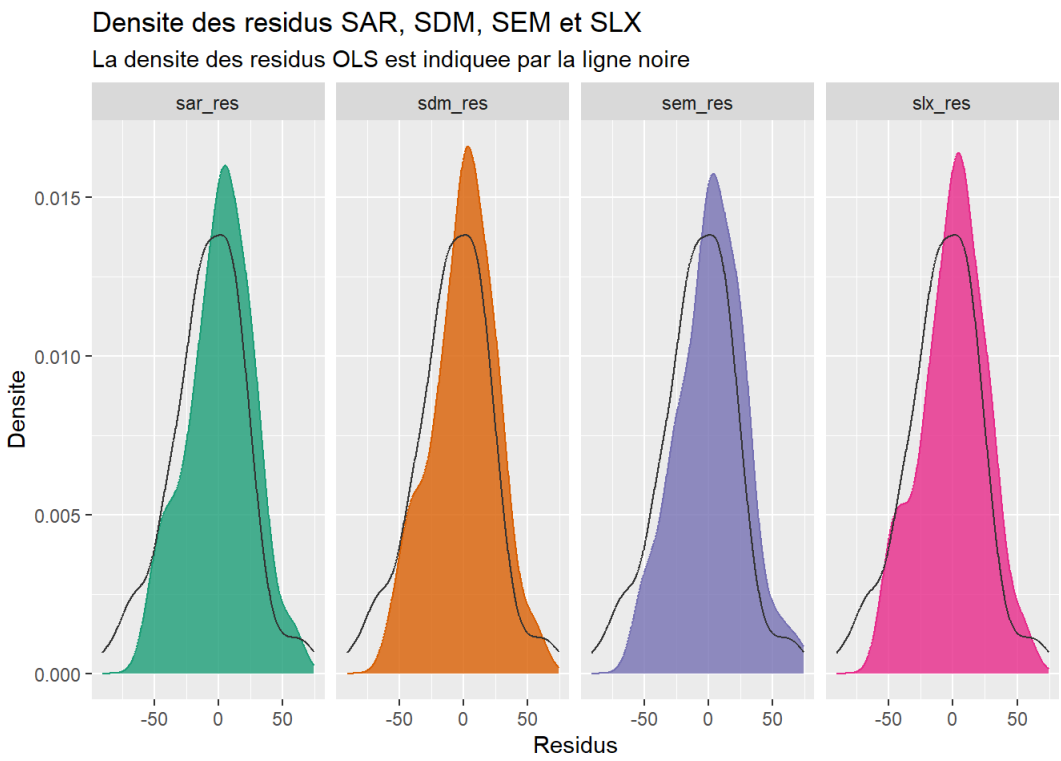
Les variables Agricultural_land, Armed_forces_personnel, Population_density et CO2_emissions ont des coefficients positifs et significatifs ($p < 0.05$), ce qui indique une forte relation positive avec la variable dépendante.

La variable Alternative_and_nuclear_energy a un coefficient négatif significatif (p , 0.05), ce qui indique une forte relation negative avec la variable dépendante.

Le reste des variables ne sont pas significatifs au niveau 5%, ce qui implique que nous avons pas assez de preuves concernant leur importance dans ce modele.

Les coefficients des variables retardées spatialement (lag_) ne sont pas statistiquement significatifs pour la plupart ($p > 0.05$), indiquant que les effets des pays voisins pour ces variables ne sont pas significatifs dans ce modèle.

Distribution des résidus



Dans tous les cas, les résidus semblent suivre une loi normale. Ce qui est rassurant.

Synthese des resultats

Résultats d'estimation des modèles spatiaux

	Dependent variable:			
	OLS	SLX	SEM	SDM
Agricultural_land	0.623***	0.323**	0.611***	0.323***
	(0.123)	(0.129)	(0.080)	(0.123)
W_Agricultural_land		0.317		
		(0.193)		
Alternative_and_nuclear_energy	-0.435	-1.083***	-0.634*	-0.858**
	(0.306)	(0.344)	(0.344)	(0.360)
W_Alternative_and_nuclear_energy		-0.323		
		(0.719)		
Armed_forces_personnel	20.686***	5.807**	10.563***	5.762**
	(2.617)	(2.476)	(1.902)	(2.358)
W_Armed_forces_personnel		10.842*		
		(6.133)		
CO2_emissions	3.117***	2.332***	3.793***	3.009***
	(0.469)	(0.606)	(0.596)	(0.726)
GDP_per_capita	-0.0004*		-0.0004**	-0.0004
	(0.0002)		(0.0002)	(0.0002)
W_CO2_emissions		-0.303		
		(1.417)		
Population_density	0.040*	0.034*	0.048***	0.038**
	(0.023)	(0.019)	(0.017)	(0.018)
W_Population_density		-0.001		
		(0.041)		
lag.Agricultural_land			0.296	
			(0.220)	
lag.Alternative_and_nuclear_energy			-0.154	
			(1.326)	
lag.Armed_forces_personnel			9.482	
			(6.333)	
lag.CO2_emissions			-0.692	
			(2.098)	
lag.GDP_per_capita			0.0002	
			(0.001)	
lag.Population_density			0.004	
			(0.040)	
Observations	119	119	119	119
R2	0.835	0.852		
Adjusted R2	0.826	0.838		
Log Likelihood			-557.176	-551.255
sigma2			682.931	618.226
Akaike Inf. Crit.			1,130.351	1,130.509
Residual Std. Error	4.315 (df = 113)	26.245 (df = 109)		
F Statistic	95.278*** (df = 6; 113)	62.643*** (df = 10; 109)		
Wald Test (df = 1)			0.007	0.014
LR Test (df = 1)			0.006	0.012
Note:				*p<0.1; **p<0.05; ***p<0.01

VII) Choix du modèle

Lesage et Pace

SDM VS SAR

$$H_0 : \delta = 0$$

$$If H_0 rejected = SDM$$

```
	Likelihood ratio for spatial linear models

data:
Likelihood ratio = 4.8143, df = 6, p-value = 0.5678
sample estimates:
Log likelihood of prod_elecs_SDM Log likelihood of prod_elecs_SAR
                       -551.2545                       -553.6617
```

Ce test nous permet de dire que le modèle SDM n'est pas meilleur que celui SAR.

SAR VS OLS

$$H_0 : \delta = 0$$

$$If H_0 rejected = SAR$$

```
	Likelihood ratio for spatial linear models

data:
Likelihood ratio = 103.44, df = 1, p-value < 2.2e-16
sample estimates:
Log likelihood of prod_elecs_SAR Log likelihood of prod_elecs_OLS
                       -553.6617                       -605.3799
```

Ce teste nous montre que le modèle SAR est meilleur que celui OLS

SDM VS SLX

$$H_0 : \ \rho = 0, \theta \neq 0, \theta + \rho * \beta \neq 0$$

```
	Likelihood ratio for spatial linear models

data:
Likelihood ratio = 2.4128, df = 3, p-value = 0.4913
sample estimates:
Log likelihood of prod_elecs_SDM Log likelihood of prod_elecs_SLX
                       -551.2545                       -552.4609
```

Ce teste révèle que le modèle SLX est meilleur que celui SDM.

SLX VS OLS

```
	Likelihood ratio for spatial linear models

data:
Likelihood ratio = 105.84, df = 4, p-value < 2.2e-16
sample estimates:
Log likelihood of prod_elecs_SLX Log likelihood of prod_elecs_OLS
                       -552.4609                       -605.3799
```

Ce teste révèle que le modèle SLX est meilleur que celui OLS

SDM VS SEM

$$H_0 : \theta + \rho * \beta \neq 0$$

```
	Likelihood ratio for spatial linear models

data:
Likelihood ratio = 11.842, df = 6, p-value = 0.06558
sample estimates:
Log likelihood of prod_elecs_SDM Log likelihood of prod_elecs_SEM
                       -551.2545                       -557.1757
```

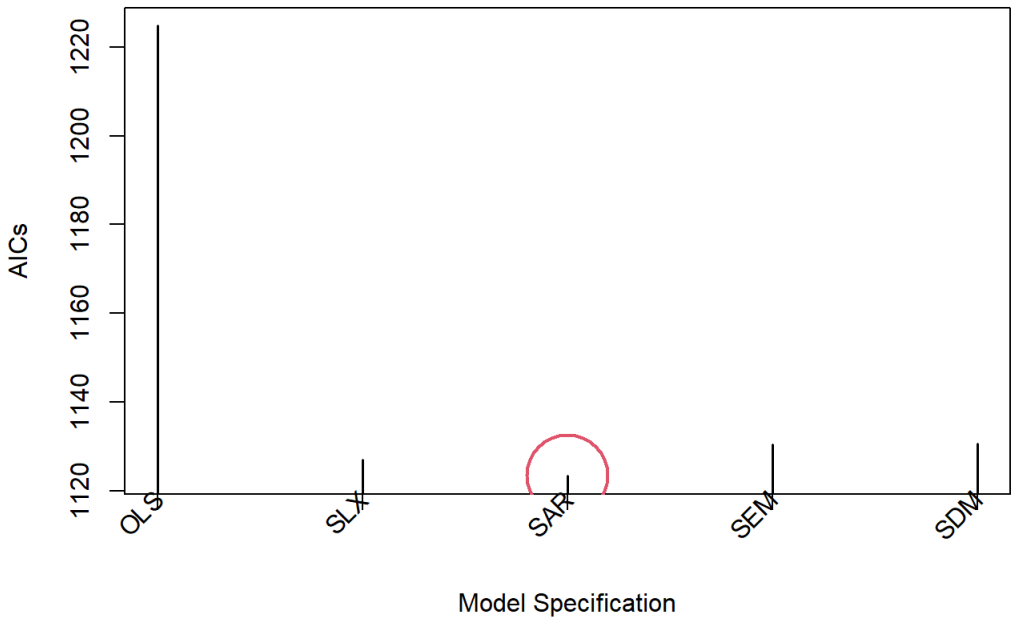
Ce teste révèle que le modèle SDM est meilleur que celui SEM

SEM VS OLS

```
	Likelihood ratio for spatial linear models

data:
Likelihood ratio = 96.408, df = 1, p-value < 2.2e-16
sample estimates:
Log likelihood of prod_elecs_SEM Log likelihood of prod_elecs_OLS
                       -557.1757                        -605.3799
```

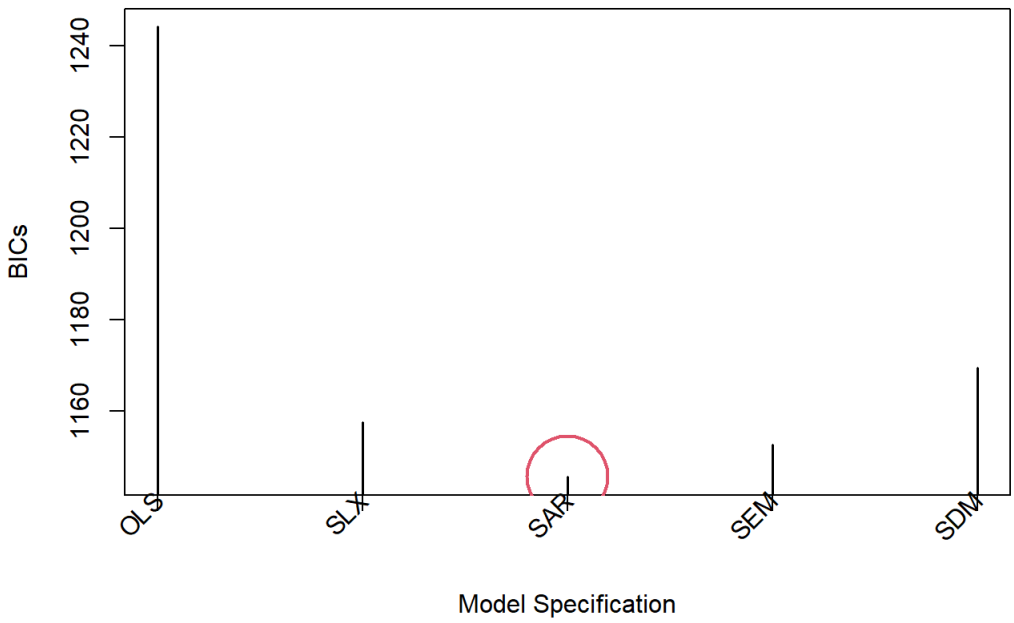
Ce teste révèle que le modèle OLS est meilleur que celui SEM

AIC

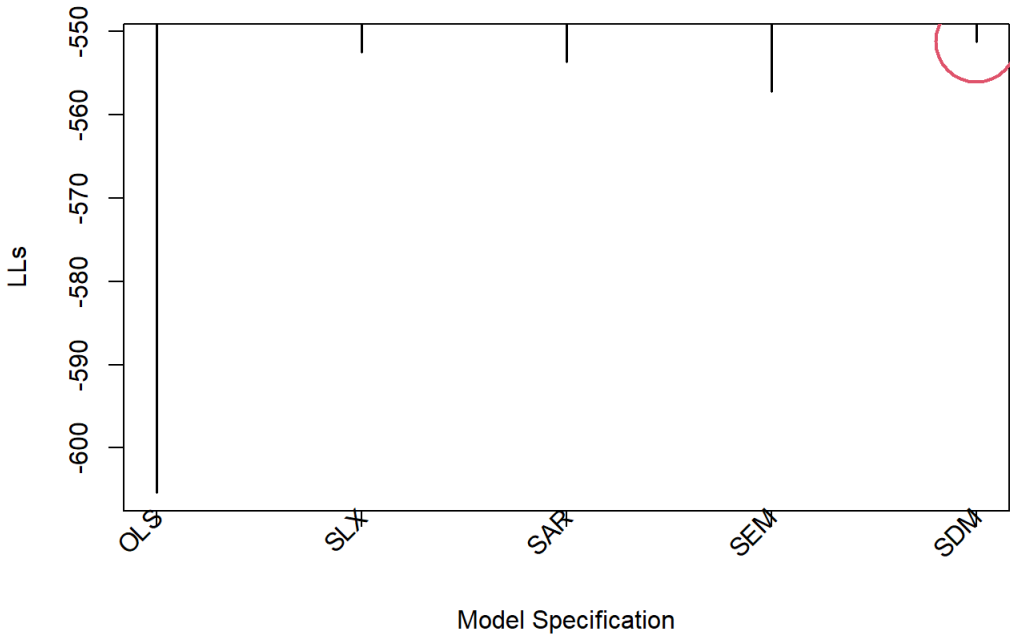


Pour le critère de l'AIC le modèle SAR est le plus efficace

BIC



Pour le critère du BIC c'est également le modèle SAR qui est le plus performant.



Concernant le maximum de vraisemblance, le modèle SDM est meilleur; mais pas de loin par rapport au modèle SAR.

```
    Likelihood ratio for spatial linear models

data:
Likelihood ratio = 4.8143, df = 6, p-value = 0.5678
sample estimates:
Log likelihood of prod_elecs_SDM Log likelihood of prod_elecs_SAR
                       -551.2545                        -553.6617
```

En comparant le modèle SDM et SAR, la p-value indique que celui SAR est meilleur.

Conclusion

Dans le cadre de cette étude portant sur le pourcentage d'electricite produite a partir d'energies fossiles, de gaz naturel et/ou de charbon dans le monde en 2015 , nous avons exploré les relations spatiales entre différentes variables socio-économiques et démographiques.

Les caractéristiques socio-économiques sont distribuées de manière spatialement autocorrélée. On peut penser que les processus socio-économiques locaux influencient le comportement des pays voisins.

Parmi les differents modèles testés, nous avons choisi le modèle SAR qui est globalement satisfaisant et possède les meilleures métriques (AIC et BIC).

Notre étude met en avant l'importance de prendre en compte les interactions spatiales dans l'analyse des phénomènes socio-économiques. Cela peut être utile pour mieux comprendre pourquoi certains pays ont un niveau de developpement different malgre des caracteristiques similaires.