

Statistique spatiale

Abdoul Razac SANE

10/03/2022

Contents

Chargement des packages	1
1. Import des données et matrice de poids	2
2. Cartographie de la proportion d'enfants vivant dans la pauvreté	2
3. Test de Moran	3
4. Test de Moran Local	5
5. Corrélation des variables explicatives	11
6. Modèle linéaire	12
7. Remarque sur la variable	13
8. Etude de la multicolinéarité des variables explicatives	13
9. Mise en oeuvre du deuxième modèle linéaire.	13
10. 11. 12. Choix entre modèle LAD et SEM	17
14. Pour le modèle SDM au lieu du modèle SEM ou LAG	20
15. Mise en oeuvre du modèle SDM	20
16. Comparaison des modèles SEM et SDM	20
17. Estimation des effets directs et indirects du modèle SDM	21
18. Le modèle SDM est-il justifié ?	22

Chargement des packages

```
library(dplyr)
library(sf)
library(cartography)
library(RColorBrewer)
library(spdep)
library(GGally)
library(spatialreg)
```

1. Import des données et matrice de poids

Chargement des fichiers de données

```
df <- st_read("south00/south00.shp")
```

```
## Reading layer 'south00' from data source
##   '/media/abdoul/Data/Aix-Marseille/MASS-POP-2020-2021/Stat_Spatiale/Devoir_Maison/south00/south00.shp'
##   using driver 'ESRI Shapefile'
## Simple feature collection with 1387 features and 17 fields
## Geometry type: MULTIPOLYGON
## Dimension:      XY
## Bounding box:  xmin: -957551.2 ymin: -1609326 xmax: 1747589 ymax: 175426.2
## CRS:            NA
```

Création du voisinage de type Queen et de la matrice de poids

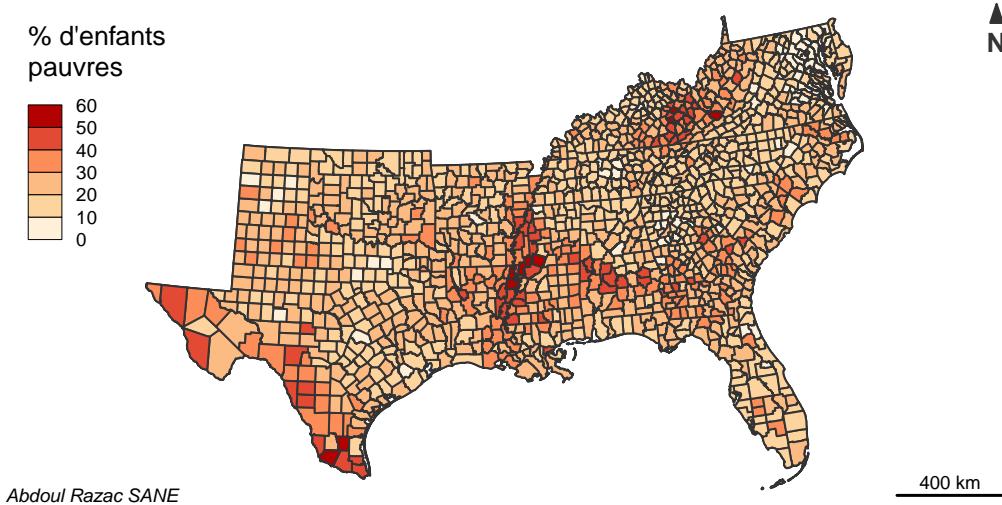
```
vois_queen <- poly2nb(df, row.names = df$FIPS, queen = TRUE)
poids_queen <- nb2listw(vois_queen, style = "W", zero.policy = TRUE)
```

2. Cartographie de la proportion d'enfants vivant dans la pauvreté

```
# Conversion de PPOV en pourcentage
df <- df %>% mutate(PPOV2 = 100*PPOV)

# Cartographie
plot(st_geometry(df))
choroLayer(df,
           var = "PPOV2",
           # method = "quantile",
           # nclass = 5,
           breaks = 0:6*10,
           col = brewer.pal(6, 'OrRd'),
           legend.title.txt = "% d'enfants\nnpauvres",
           legend.title.cex = .8, legend.pos = "topleft",
           add = TRUE)
title("Proportion d'enfants vivant dans la pauvreté\nndans le Sud et le Sud-Est des Etats-Unis")
layoutLayer( title = "", author = "Abdoul Razac SANE", north = TRUE, frame = F)
```

Proportion d'...enfants vivant dans la pauvreté dans le Sud et le Sud-Est des Etats-Unis



3. Test de Moran

Le test de Moran permet de tester l'existence d'une autocorrélation spatiale entre les comtés.

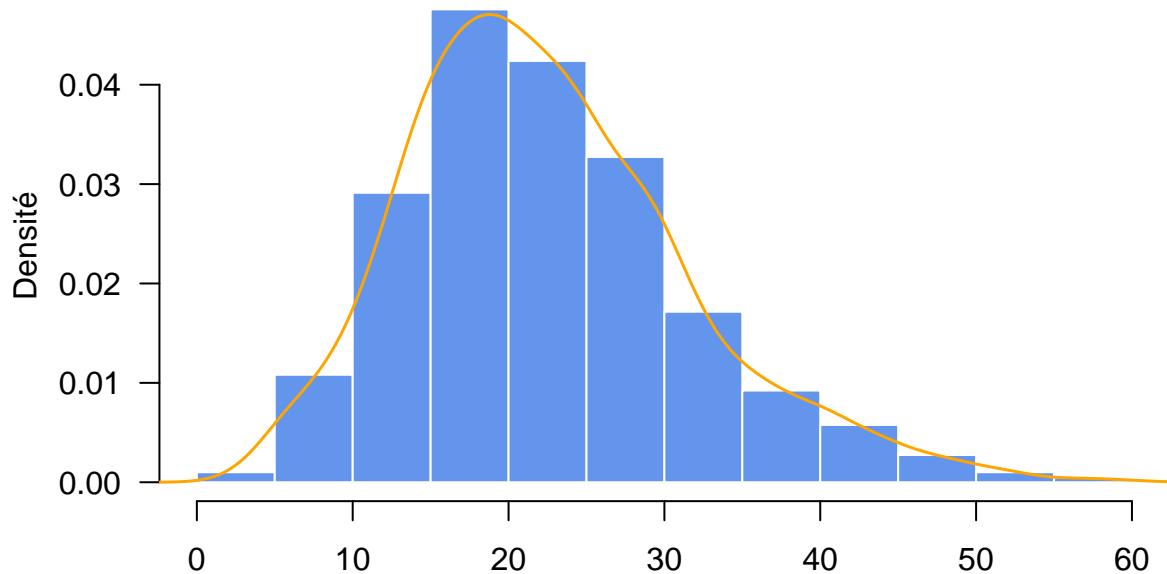
Les hypothèses de ce test sont :

- H_0 : Pas d'autocorrélation spatiale
- H_1 : Existence d'une autocorrélation spatiale

Pour ce faire nous vérifions tout d'abord si la part des enfants pauvres (PPOV) suit une distribution normale.

```
# Histogramme de PPOV2
hist(df$PPOV2, border = "white",
      probability = TRUE,
      las = 1,
      col = "cornflowerblue",
      xlab = "", ylab = "Densité",
      main = "Distribution de la part d'enfants vivant dans la pauvreté")
lines(density(df$PPOV2), col = "orange", lwd=1.5)
```

Distribution de la part d'enfants vivant dans la pauvreté



```
# Test de Shapiro-Wilk sur PPOV
shapiro.test(df$PPOV)
```

```
##
##  Shapiro-Wilk normality test
##
## data: df$PPOV
## W = 0.97175, p-value = 7.651e-16
```

Le test de Shapiro-Wilk rejette de l'hypothèse de normalité. On conclut que le vecteur d'observation n'est pas gaussien. Cela nous conduit à un test de Moran avec bootstrap.

```
# Test de Moran avec Bootstrap
test_moran <- moran.mc(df$PPOV, poids_queen, 999, zero.policy=FALSE, alternative="greater")
test_moran
```

```
##
##  Monte-Carlo simulation of Moran I
##
## data: df$PPOV
## weights: poids_queen
## number of simulations + 1: 1000
##
## statistic = 0.5893, observed rank = 1000, p-value = 0.001
## alternative hypothesis: greater
```

La p-valeur est très petite (inférieure à 0.001) donc nous rejetons l'hypothèse H_0 alors il y a une autocorrélation spatiale positive de proportion d'enfants vivant dans la pauvreté entre les comtés. Plus cette proportion est élevée dans un comté, plus elle est élevée dans les comtés voisins. L'indice de Moran vaut 0.5893.

4. Test de Moran Local

```
# Calcul des LISA
lisa <- localmoran(df$PPOV, poids_queen, zero.policy=T) %>%
  as.data.frame()
names(lisa) <- c("Ii","Ei","Vari","Zi","pvi")
rownames(lisa) <- df$CNTY_ST
head(lisa)
```

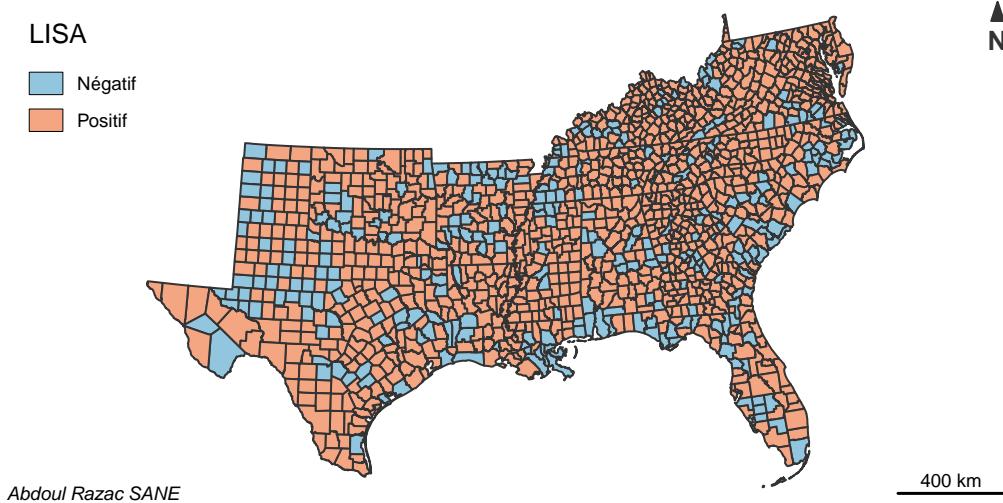
	Ii	Ei	Vari	Zi	pvi
## Autauga County AL	-0.6293437	-0.0006641639	0.18358496	-1.4672724	0.142301971
## Baldwin County AL	-0.3504966	-0.0007117863	0.16383056	-0.8641787	0.387489752
## Barbour County AL	1.5512537	-0.0018945986	0.32619668	2.7194004	0.006540039
## Bibb County AL	0.1798426	-0.0002757427	0.06349489	0.7148064	0.474728682
## Blount County AL	0.3939768	-0.0006938142	0.15969682	0.9876128	0.323342333
## Bullock County AL	2.7692836	-0.0044016457	1.21213115	2.5193144	0.011758361

- Cartographie des indices locaux de Moran

```
# Création des indices dans la base de données
df$indice_local <- lisa$Ii
df$indice_local_positif <- ifelse(lisa$Ii >= 0, "Positif", "Négatif")
df$indice_local_signif <- if_else(lisa$pvi < 0.05 , "Significatif", "Non significatif", missing = "Non significatif")
```

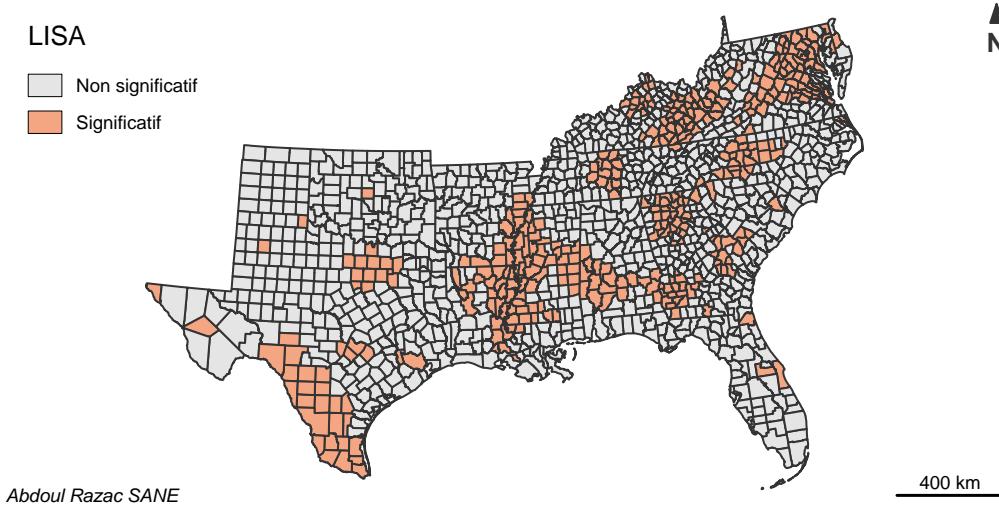
```
# Cartographie des LISA
plot_sf(df)
typoLayer(df, var = "indice_local_positif",
          legend.title.txt = "LISA",
          col = c("#92C5DE", "#F4A582"),
          legend.title.cex = .8, legend.pos = "topleft",
          add = TRUE)
title("Indice de Moran local par comté", cex = 0.6)
layoutLayer( title = "", author = "Abdoul Razac SANE", north = TRUE, frame = F)
```

Indice de Moran local par comté



```
# Cartographie des LISA
plot_sf(df)
typoLayer(df, var = "indice_local_signif",
          legend.title.txt = "LISA",
          col = c("gray90", "#F4A582"),
          legend.title.cex = .8, legend.pos = "topleft",
          add = TRUE)
title("Indice de Moran local significatif par comté")
layoutLayer( title = "", author = "Abdoul Razac SANE", north = TRUE, frame = F)
```

Indice de Moran local significatif par comté



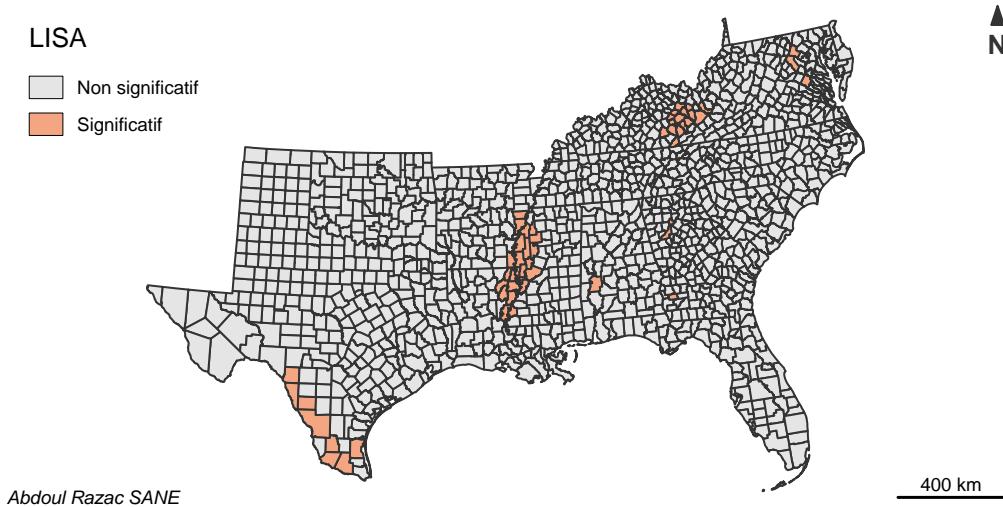
Le test moran.mc a tendance à surestimer la p-valeur (car la loi est simulée par Bootstrap). Ce qui conduit à beaucoup de valeurs significatives. On peut ajuster ces p-valeurs par la méthode de Holm.

```
# Création de la variable p_holm et p_holm_signif
df <- df %>% mutate(
  p_holm = p.adjust(lisa$pvi, method="holm"), # Ajustement de Holm
  p_holm_signif = if_else(p_holm < 0.05, "Significatif", "Non significatif",
})
```



```
# Carte des comtés avec un p_holm significatif
plot_sf(df)
typoLayer(df, var = "p_holm_signif",
  legend.title.txt = "LISA",
  col = c("gray90", "#F4A582"),
  legend.title.cex = .8, legend.pos = "topleft",
  add = TRUE)
title("Indice de Moran local significatif avec la méthode Holm")
layoutLayer( title = "", author = "Abdoul Razac SANE", north = TRUE, frame = F)
```

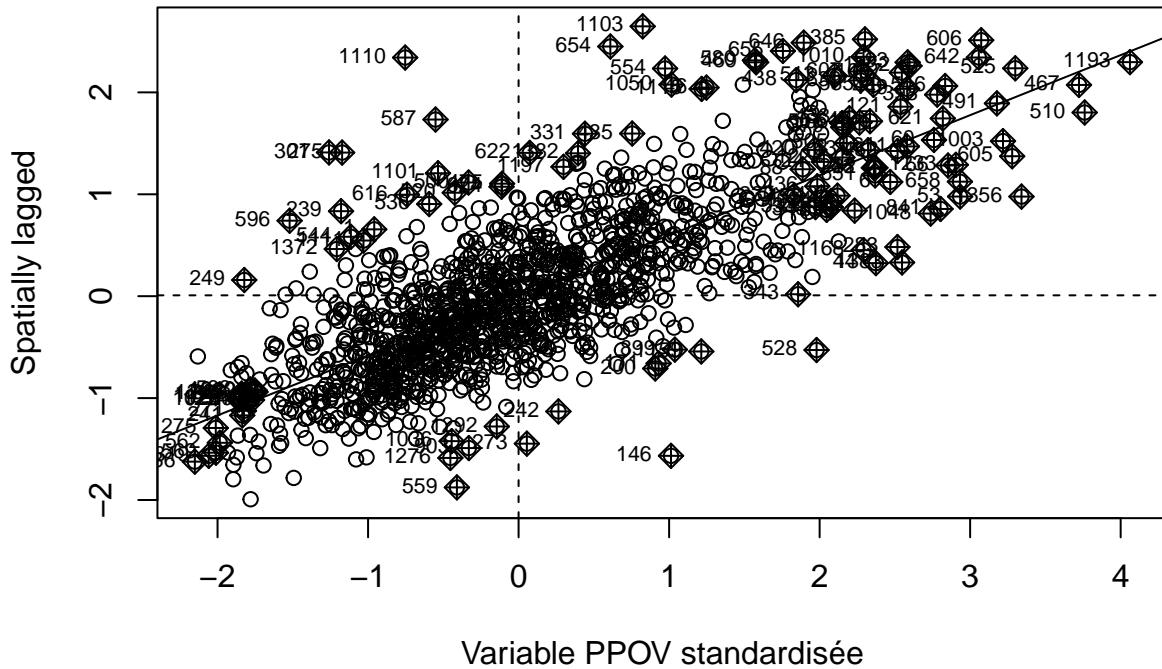
Indice de Moran local significatif avec la méthode Holm



- Diagramme de moran

```
# Diagramme de Moran
df$PPOV_std <- scale(df$PPOV)
moran.plot(as.vector(df$PPOV_std), poids_queen, zero.policy=F,
           xlab = "Variable PPOV standardisée", ylab = "Spatially lagged",
           main = "Diagramme de Moran")
```

Diagramme de Moran

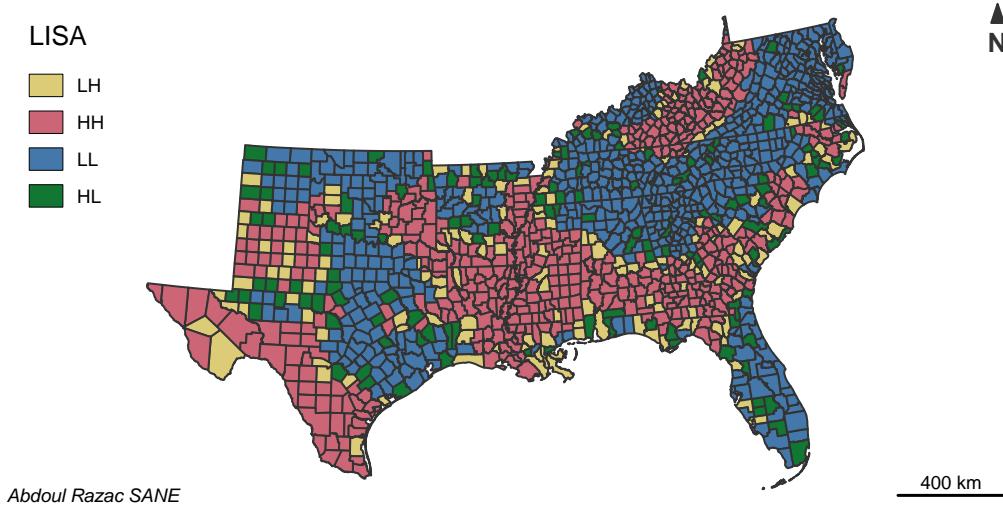


Nous avons cartographié les slots. Seuls les comtés avec un indice de Moran local significatif ont été représentés.

```
# Création des catégories de slots et des slots avec un p_holm significatif
df <- df %>% mutate(
  indice_local_lag = lag.listw(poids_queen, PPOV_std, zero.policy=F),
  slots = case_when(
    PPOV_std > 0 & indice_local_lag > 0 ~ "HH",
    PPOV_std < 0 & indice_local_lag < 0 ~ "LL",
    PPOV_std < 0 & indice_local_lag > 0 ~ "LH",
    PPOV_std > 0 & indice_local_lag < 0 ~ "HL"
  ),
  slots_signif = if_else(p_holm < 0.05, slots, "Non signif")
)
```

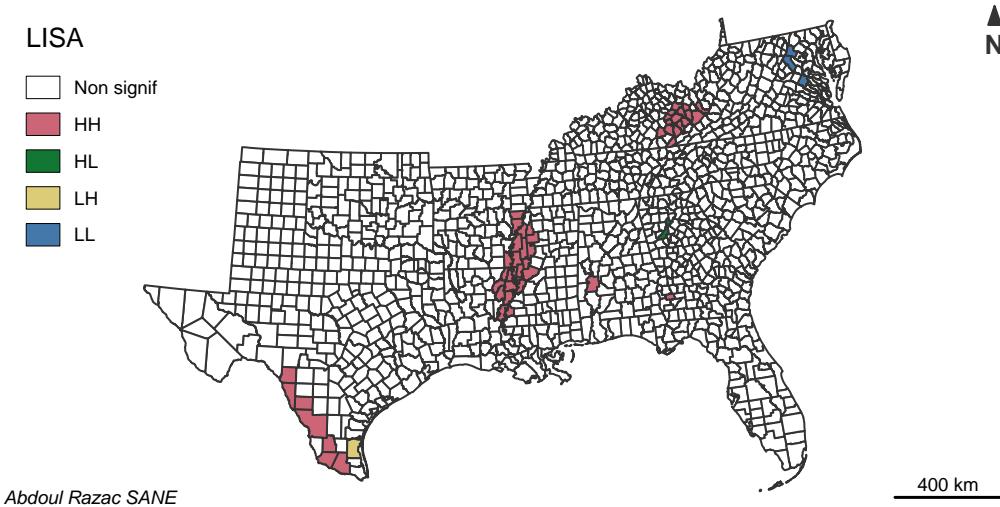
```
# Cartographie des slots
plot_sf(df)
typoLayer(df, var = "slots",
          legend.title.txt = "LISA",
          col = c("#DDCC77", "#CC6677", "#4477AA", "#117733"),
          legend.title.cex = .8, legend.pos = "topleft",
          add = TRUE)
title("Slots des comtés avec la méthode Holm")
layoutLayer( title = "", author = "Abdoul Razac SANE", north = TRUE, frame = F)
```

Slots des comtés avec la méthode Holm



```
# Cartographie des slots avec un p_holm significatif
plot_sf(df)
typoLayer(df, var = "slots_signif",
          legend.title.txt = "LISA",
          col = c("#FFFFFF", "#CC6677", "#117733", "#DDCC77", "#4477AA"),
          legend.title.cex = .8, legend.pos = "topleft",
          add = TRUE)
title("Slots des comtés significatifs avec la méthode Holm")
layoutLayer( title = "", author = "Abdoul Razac SANE", north = TRUE, frame = F)
```

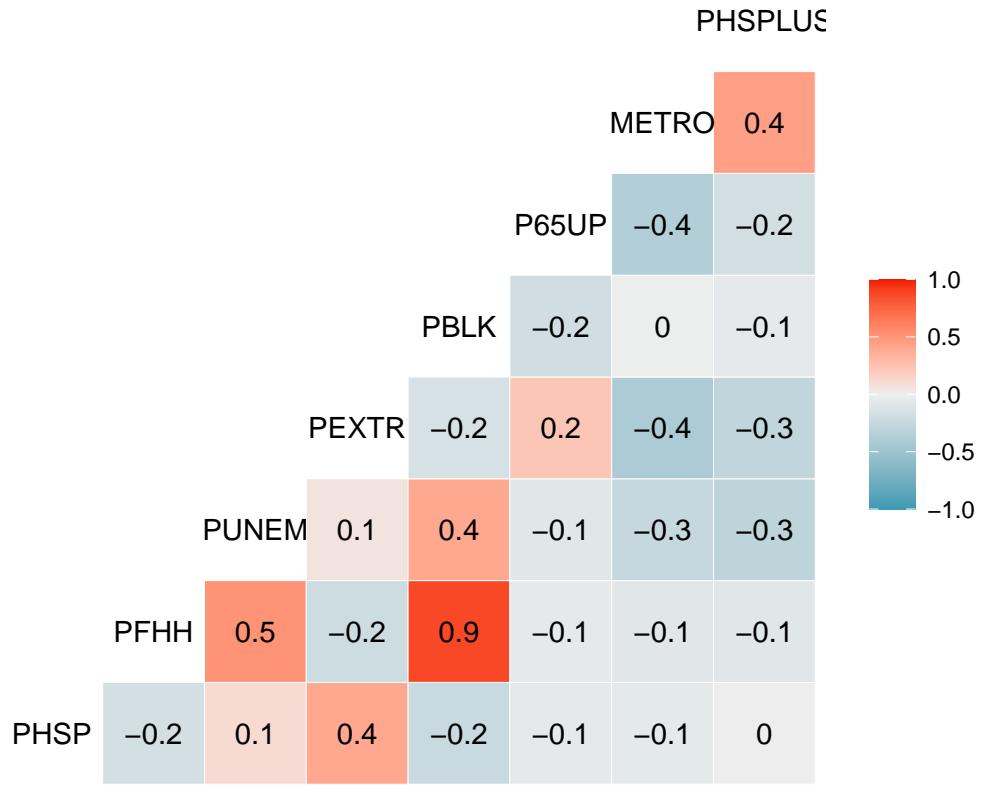
Slots des comtés significatifs avec la méthode Holm



Dans les comtés rouges, nous avons une forte association spatiale positive. Dans ces comtés, il y a une forte part d'enfants pauvres tout comme dans les comtés voisins. Deuxièmement les comtés bleus ont une association spatiale négative. Ils ont une faible part d'enfants pauvres tout comme leurs voisins. Il y a deux comtés qui ont des proportions d'enfants pauvres différentes de leurs voisins. Pour le comté jaune, cette proportion est faible tandis qu'elle est élevée pour le comté vert.

5. Corrélation des variables explicatives

```
df %>%
  as_tibble() %>%
  select(PHSP, PFHH, PUNEM, PEXTR, PBLK, P65UP, METRO, PHSPLUS) %>%
  ggc当地 = TRUE)
```



La corrélation la plus élevée se situe entre les proportions de femmes cheffes de foyers et d'afro-américains. Elle est positive avec une valeur de 0,9. La valeur absolue des autres coefficients de corrélation est inférieure à 0,6.

6. Modèle linéaire

```
mod <- lm(PPOV ~ PHSP + PFHH + PUNEM + PEXTR + PBLK + P65UP + METRO + PHSPLUS, data = df)
summary(mod)
```

```
##
## Call:
## lm(formula = PPOV ~ PHSP + PFHH + PUNEM + PEXTR + PBLK + P65UP +
##     METRO + PHSPLUS, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.194206 -0.025234 -0.001538  0.023895  0.246325 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 0.053335  0.009651  5.527 3.90e-08 ***
## PHSP        0.048178  0.009625  5.005 6.30e-07 ***
## PFHH        0.748316  0.038598 19.387 < 2e-16 ***
## PUNEM       1.408049  0.059929 23.495 < 2e-16 ***
```

```

## PEXTR      0.391433  0.024612 15.904 < 2e-16 ***
## PBLK       -0.110581  0.014623 -7.562 7.21e-14 ***
## P65UP      0.030057  0.035549  0.846   0.3980
## METRO     -0.007764  0.002998 -2.590   0.0097 **
## PHSPLUS    -0.243284  0.012841 -18.946 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04261 on 1378 degrees of freedom
## Multiple R-squared:  0.7832, Adjusted R-squared:  0.7819
## F-statistic: 622.3 on 8 and 1378 DF,  p-value: < 2.2e-16

```

Nous avons un modèle qui explique 78% de la variabilité totale ($R^2_{ajusté} = 0,7819$). Toutes nos variables sont significatives exceptée la proportion de personnes âgées de 65 ans et plus. La proportion des sans emploi a un effet plus prononcé ; une augmentation de cette proportion d'un point de pourcentage entraîne une augmentation de la proportion d'enfants pauvres de 1,4 points de pourcentage.

Les comtés métropolitains, la proportion d'afro-américains et la proportion de diplômés du secondaire ou plus ont un effet négatif sur la proportion d'enfants vivant dans la pauvreté. Toutefois la proportion d'afro-américains a un effet incohérent. En effet, elle est fortement et positivement corrélée avec la proportion de femmes cheffes de famille donc elle devrait avoir une influence qui va dans le même sens que cette dernière.

7. Remarque sur la variable

La remarque sur l'effet controversé de la proportion d'afro-américains (PBLK) dans le modèle pourrait être un éventuel problème de multi-collinearité entre elle et la proportion de femmes cheffes de famille.

8. Etude de la multicolinéarité des variables explicatives

```
car::vif(mod = mod)
```

```

##      PHSP      PFHH      PUNEM      PEXTR      PBLK      P65UP      METRO      PHSPLUS
## 1.413508 5.867381 1.710764 1.699723 5.224766 1.393145 1.615034 1.368849

```

Comme nous l'avons mentionné plus haut, le *VIF* confirme la multicolinéarité entre la proportion d'afro-américains (PBLK) et la proportion de femmes cheffes de famille (PFHH). Pour résoudre ce problème, on peut soit créer un indicateur synthétique à partir d'une analyse en composante principale (ACP) sur les régresseurs corrélés, soit exclure un de ces régresseurs.

9. Mise en oeuvre du deuxième modèle linéaire.

Dans ce modèle nous excluons la proportion d'afro-américains (PBLK) des régresseurs.

```
mod2 <- lm(PPOV ~ PHSP + PFHH + PUNEM + PEXTR + P65UP + METRO + PHSPLUS, data = df)
summary(mod2)
```

```

##
## Call:
## lm(formula = PPOV ~ PHSP + PFHH + PUNEM + PEXTR + P65UP + METRO +
## 
```

```

##      PHSPLUS, data = df)
##
## Residuals:
##       Min        1Q     Median        3Q       Max
## -0.205612 -0.024960 -0.001279  0.024781  0.262327
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.075849  0.009365   8.099 1.21e-15 ***
## PHSP        0.067341  0.009473   7.109 1.87e-12 ***
## PFHH        0.497657  0.020175  24.667 < 2e-16 ***
## PUNEM       1.465323  0.060647  24.162 < 2e-16 ***
## PEXTR        0.343934  0.024277  14.167 < 2e-16 ***
## P65UP        0.119585  0.034195   3.497 0.000485 ***
## METRO       -0.008858  0.003054  -2.900 0.003792 **
## PHSPLUS     -0.243540  0.013100 -18.591 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04347 on 1379 degrees of freedom
## Multiple R-squared:  0.7742, Adjusted R-squared:  0.7731
## F-statistic: 675.5 on 7 and 1379 DF,  p-value: < 2.2e-16

```

Le modèle est assez stable. Toutes les variables explicatives sont significatives et conservent leur effet (positif ou négatif) sur la variable dépendante qu'est la proportion d'enfants vivant dans la pauvreté.

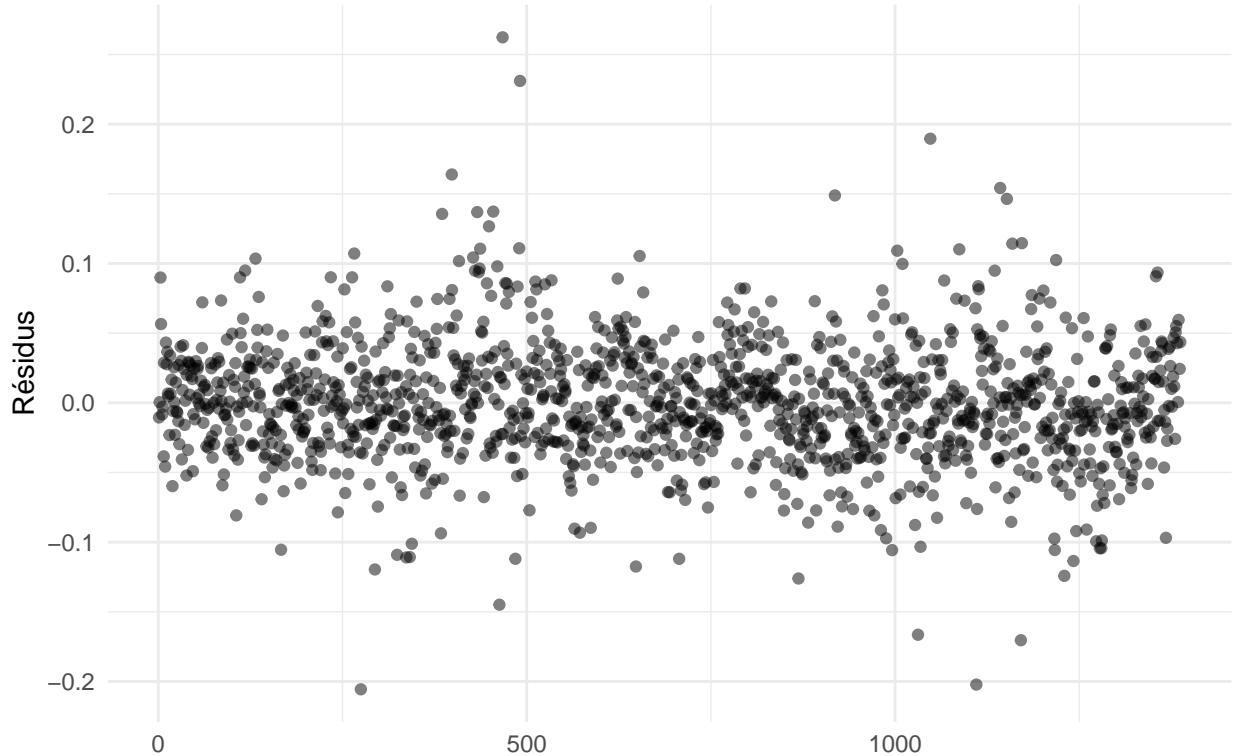
Résidus du modèle.

```

ggplot() +
  aes(x = 1:length(mod2$residuals), y = mod2$residuals) +
  geom_jitter(color = "black", alpha = 0.5) +
  labs(x = "", y = "Résidus", title = "Distribution des résidus") +
  theme_minimal()

```

Distribution des résidus



```
# Test de Shapiro-Wilk sur les résidus
shapiro.test(mod2$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data: mod2$residuals
## W = 0.97188, p-value = 8.403e-16
```

Le test de Shapiro-Wilk rejette l'hypothèse de normalité des résidus.

Test de Moran sur les résidus

```
lm.morantest(mod2, listw = poids_queen, zero.policy = TRUE)
```

```
##
##  Global Moran I for regression residuals
##
## data:
## model: lm(formula = PPOV ~ PHSP + PFHH + PUNEM + PEXTR + P65UP + METRO
## + PHSPLUS, data = df)
## weights: poids_queen
##
```

```

## Moran I statistic standard deviate = 19.154, p-value < 2.2e-16
## alternative hypothesis: greater
## sample estimates:
## Observed Moran I      Expectation      Variance
##          0.3044980977    -0.0030597865    0.0002578173

```

Le test de Moran sur les résidus montre qu'il existe une autocorrélation spatiale positive entre les résidus des comtés.

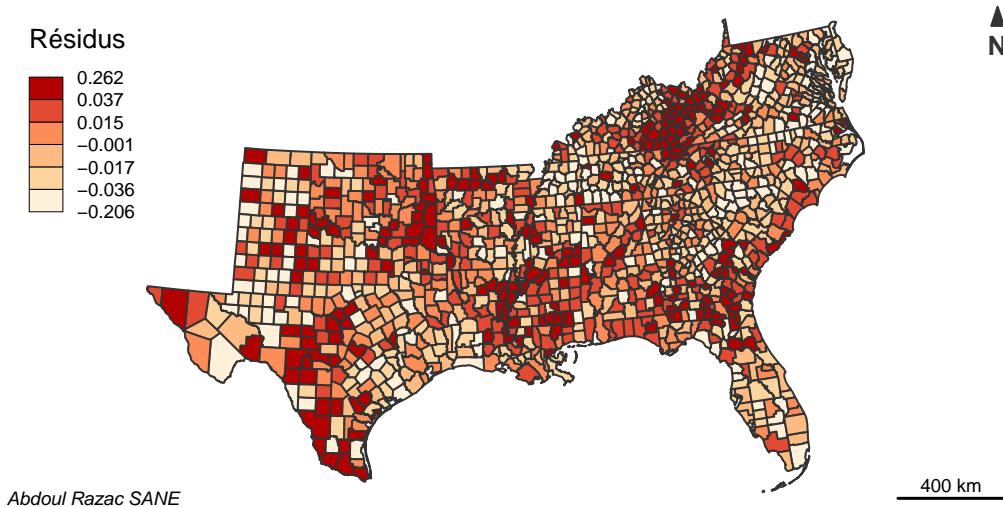
Cartographie de résidus

```

# Cartographie des résidus
df$residus_mod2 <- mod2$residuals
plot_sf(df)
choroLayer(df, var = "residus_mod2",
           legend.title.txt = "Résidus",
           nclass = 6,
           legend.values.rnd = 3,
           col = brewer.pal(6, 'OrRd'),
           legend.title.cex = .8, legend.pos = "topleft",
           add = TRUE)
title("Répartition spatiale des résidus du modèle 2")
layoutLayer( title = "", author = "Abdoul Razac SANE", north = TRUE, frame = F)

```

Répartition spatiale des résidus du modèle 2



La carte des résidus représente l'écart de proportion d'enfants pauvres à la moyenne globale. On remarque que les résidus ne sont pas répartis de manière régulière dans l'espace.

Test d'hétéroscélasticité des résidus

```
lmtest::bgtest(mod2)

##
## Breusch-Godfrey test for serial correlation of order up to 1
##
## data: mod2
## LM test = 7.2402, df = 1, p-value = 0.007129
```

Le test d'hétéroscélasticité des résidus est significatif au seuil de 1%. Donc le paramètre de variance des résidus est aléatoire. Ce qui confirme l'interprétation précédante de la carte.

Justification du modèle spatial.

Les résidus du modèle linéaire ne sont pas gaussiens. Ils sont spatialement corrélés et on a une hétéroscélasticité des résidus dans l'espace. De ce fait le modèle linéaire n'est pas approprié pour notre phénomène. Un modèle spatial serait plus approprié pour la modélisation de notre phénomène car ce dernier prend en compte l'autocorrélation spatiale.

10. 11. 12. Choix entre modèle LAD et SEM

Nous faisons tout d'abord les différents tests du Multiplicateur de Lagrange.

```
test_ML <- lm.LMtests(mod2, listw = poids_queen, test = c("LMerr", "RLMerr", "LMlag", "RLMlag", "SARMA"), ze
plyr::ldply(test_ML, function(x) c("Statistic" = as.numeric(x$statistic), "P.valeur" = x$p.value))

##      .id Statistic P.valeur
## 1  LMerr  354.07875     0
## 2 RLMerr  125.34062     0
## 3  LMlag  300.88294     0
## 4 RLMlag   72.14481     0
## 5  SARMA  426.22355     0
```

Les tests de Multiplicateur de Lagrange et les tests robustes ne permettent pas de choisir entre le modèle SEM et LAG puisque toutes les p-valeurs sont significatives et tendent vers 0. Les termes d'interraction ρ et λ sont non nuls. Donc nous implémentons ces deux modèles et nous déterminons le meilleur à l'aide du test de rapport de vraisemblance.

```
# Modèle LAD et SEM
mod_lag <- lagsarlm(mod2$call$formula, data = df, poids_queen)
mod_sem <- errorsarlm(mod2$call$formula, data = df, poids_queen)

# AIC des modèles LAG et SEM
glue::glue("AIC modèle LAG = {round(AIC(mod_lag), 2)}\n AIC modèle SEM = {round(AIC(mod_sem), 2)}")
```

```
## AIC modèle LAG = -5022.74
## AIC modèle SEM = -5096.33
```

L'AIC du modèle SEM est le plus faible, on choisit donc celui-là pour la modélisation spatiale.

```
summary(mod_sem)
```

```
##
## Call:errorsarlm(formula = mod2$call$formula, data = df, listw = poids_queen)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.23731349 -0.02157244 -0.00087547  0.01974822  0.19201350
##
## Type: error
## Coefficients: (asymptotic standard errors)
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept) 0.0732903 0.0102571 7.1454 8.977e-13
## PHSP        0.0850482 0.0153964 5.5239 3.316e-08
## PFHH        0.5947810 0.0220961 26.9180 < 2.2e-16
## PUNEM       0.9373402 0.0600132 15.6189 < 2.2e-16
## PEXTR        0.3130227 0.0264418 11.8382 < 2.2e-16
## P65UP        0.1384429 0.0371617 3.7254 0.000195
## METRO       -0.0034557 0.0027368 -1.2627 0.206708
## PHSPLUS     -0.2194900 0.0124924 -17.5699 < 2.2e-16
##
## Lambda: 0.64842, LR test value: 345.75, p-value: < 2.22e-16
## Asymptotic standard error: 0.026464
##      z-value: 24.501, p-value: < 2.22e-16
## Wald statistic: 600.32, p-value: < 2.22e-16
##
## Log likelihood: 2558.163 for error model
## ML residual variance (sigma squared): 0.0013299, (sigma: 0.036467)
## Number of observations: 1387
## Number of parameters estimated: 10
## AIC: -5096.3, (AIC for lm: -4752.6)
```

Le modèle SEM s'interprète de la même manière que le modèle linéaire. La prise en compte de la dimension spatiale dans ce modèle annule l'effet de la variable « comté métropolitain » et atténue l'estimation des autres variables. Tous les coefficients des variables explicatives conservent leur signe.

```
# Test du ratio de vraisemblance
LR.Sarlm(mod2, mod_sem)
```

```
##
## Likelihood ratio for spatial linear models
##
## data:
## Likelihood ratio = -345.75, df = 1, p-value < 2.2e-16
## sample estimates:
##      Log likelihood of mod2 Log likelihood of mod_sem
##                      2385.288                      2558.163
```

Le test de log ratio nous confirme que le modèle SEM nous donne de meilleurs résultats que la régression linéaire.

- Effet du modèle non choisi : le modèle LAG.

```

trMatc <- as(poids_queen, "CsparseMatrix") %>%
  trW(type="mult")

# Effet du modèle LAG
impacts(mod_lag, R = 200, tr = trMatc) %>% summary(zstats=TRUE, short=TRUE)

## Impact measures (lag, trace):
##          Direct   Indirect      Total
## PHSP     0.059817239  0.028996955  0.088814194
## PFHH     0.434755189  0.210751560  0.645506749
## PUNEM    1.097369013  0.531959681  1.629328695
## PEXTR    0.246901630  0.119687827  0.366589457
## P65UP    0.139179581  0.067468577  0.206648158
## METRO   -0.006330816 -0.003068921 -0.009399737
## PHSPLUS -0.212895872 -0.103203224 -0.316099096
## -----
## Simulation results ( variance matrix):
## -----
## Simulated standard errors
##          Direct   Indirect      Total
## PHSP     0.009392402  0.004883102  0.013802800
## PFHH     0.020864595  0.017513695  0.030441731
## PUNEM    0.056667796  0.047894311  0.087199575
## PEXTR    0.022308159  0.013447394  0.032832445
## P65UP    0.031246537  0.016101755  0.046616928
## METRO    0.002968919  0.001481799  0.004432736
## PHSPLUS 0.011989376  0.009425157  0.018047032
## 
## Simulated z-values:
##          Direct   Indirect      Total
## PHSP     6.351169  5.993634  6.442190
## PFHH     20.675627 12.098099 21.131222
## PUNEM    19.408477 11.283545 18.810335
## PEXTR    11.052678  9.004101 11.197661
## P65UP    4.506676  4.296666  4.504842
## METRO   -2.138023 -2.107129 -2.136369
## PHSPLUS -17.704490 -11.064226 -17.540162
## 
## Simulated p-values:
##          Direct   Indirect      Total
## PHSP     2.1369e-10 2.0520e-09 1.1776e-10
## PFHH     < 2.22e-16 < 2.22e-16 < 2.22e-16
## PUNEM    < 2.22e-16 < 2.22e-16 < 2.22e-16
## PEXTR    < 2.22e-16 < 2.22e-16 < 2.22e-16
## P65UP    6.5851e-06 1.7339e-05 6.6422e-06
## METRO    0.032515  0.035106  0.032649
## PHSPLUS < 2.22e-16 < 2.22e-16 < 2.22e-16

```

Nous avons ici les effets moyens directs et indirects des régresseurs sur la proportion moyenne d'enfants pauvres. Les effets directs concernent les régresseurs d'un comté observé sur lui-même et les effets indirects concernent les régresseurs des comtés voisins sur le comté observé. Par exemple, plus la proportion d'hispaniques (PHSP) est élevée dans le comté observé et ses voisins, plus la proportion d'enfants pauvres est élevée dans le comté observé.

- Test d'hétéroscédasticité sur notre modèle choisi, le modèle SEM.

```
bptest.Sarlm(mod_sem)
```

```
##  
## studentized Breusch-Pagan test  
##  
## data:  
## BP = 167.63, df = 7, p-value < 2.2e-16
```

L'hétéroscédasticité des résidus demeure toujours avec le modèle SEM (P-valeur < 0,05) bien qu'il donne de meilleurs résultats par rapport aux modèles linéaire et LAG. On essayera par la suite de voir s'il existe un modèle meilleur que le SEM.

14. Pour le modèle SDM au lieu du modèle SEM ou LAG

Le modèle SDM (Spatial Durbin Model) est plus général par rapport modèle SEM ou LAG. Ce modèle ajoute un autre paramètre ($\theta = -\rho\beta$) qui permet de prendre en compte l'hétéroscédasticité et la corrélation spatiale.

15. Mise en oeuvre du modèle SDM

```
mod_sdm <- lagsarlm(mod2$call$formula, data = df, listw = poids_queen, type="mixed")
```

16. Comparaison des modèles SEM et SDM

```
LR.Sarlm(mod_sem, mod_sdm)
```

```
##  
## Likelihood ratio for spatial linear models  
##  
## data:  
## Likelihood ratio = -108.5, df = 7, p-value < 2.2e-16  
## sample estimates:  
## Log likelihood of mod_sem Log likelihood of mod_sdm  
##           2558.163             2612.412
```

Le modèle SEM est un cas particulier du modèle SDM (modèles emboités). La statistique du test du rapport de vraisemblance nous donne la variation de la déviance du passage d'un modèle à l'autre. Elle suit une loi de χ^2 . Ce test nous dit que le modèle SDM est meilleur que le modèle SEM (P-valeur < 0,05). Donc l'ajout de l'interaction est justifié.

17. Estimation des effets directs et indirects du modèle SDM

```
# Effet du modèle SDM
impacts(mod_sdm, R = 200, tr = trMatc) %>% summary(zstats=TRUE, short=TRUE)

## Impact measures (mixed, trace):
##          Direct   Indirect     Total
## PHSP    0.086739763 -0.04124511  0.04549465
## PFHH    0.588585317 -0.25300495  0.33558037
## PUNEM   0.932418704  1.29380480  2.22622350
## PEXTR   0.289721777  0.02254475  0.31226653
## P65UP   0.116169610 -0.04300684  0.07316277
## METRO   -0.004211458 -0.03131028 -0.03552174
## PHSPLUS -0.205788497 -0.01748015 -0.22326865
## =====
## Simulation results ( variance matrix):
## =====
## Simulated standard errors
##          Direct   Indirect     Total
## PHSP    0.021383469 0.025510290 0.01811637
## PFHH    0.023070864 0.050109330 0.04972400
## PUNEM   0.059578305 0.167420691 0.17557058
## PEXTR   0.026575615 0.059400849 0.05780442
## P65UP   0.038425952 0.091636312 0.09299414
## METRO   0.002785423 0.009604668 0.01073626
## PHSPLUS 0.012424711 0.037032733 0.04107106
##
## Simulated z-values:
##          Direct   Indirect     Total
## PHSP    4.029173 -1.5264714  2.6063161
## PFHH    25.440323 -4.9388222  6.8266659
## PUNEM   15.630142  7.6938865 12.6406894
## PEXTR   10.964043  0.3484116  5.3987586
## P65UP   3.044441 -0.4889558  0.7761719
## METRO   -1.599448 -3.2297346 -3.3042848
## PHSPLUS -16.561194 -0.5946186 -5.5462019
##
## Simulated p-values:
##          Direct   Indirect     Total
## PHSP    5.5973e-05 0.1268925  0.00915220
## PFHH    < 2.22e-16 7.8596e-07 8.6910e-12
## PUNEM   < 2.22e-16 1.4211e-14 < 2.22e-16
## PEXTR   < 2.22e-16 0.7275311  6.7104e-08
## P65UP   0.0023311  0.6248730  0.43764748
## METRO   0.1097211  0.0012391  0.00095219
## PHSPLUS < 2.22e-16 0.5520985  2.9194e-08
```

La proportion d'hispaniques (PHSP), la proportion d'employés dans l'industrie extractive (PEXTR), la proportion des 65 ans et plus (P65UP), la proportion des diplômés (PHSPLUS) n'ont pas d'impact indirect (P-valeur > 0,05) sur la proportion d'enfants vivant dans la pauvreté.

Le fait que le comté soit métropolitain ou pas n'a pas d'influence sur sa proportion d'enfants pauvres, cependant la nature des comtés voisins aura un effet sur celle-ci. Les effets de cette régression spatiale (modèle SDM) s'interprètent comme le modèle LAG.

18. Le modèle SDM est-il justifié ?

Comme nous l'avons mentionné à la question 16, le modèle SDM donne de meilleurs résultats par rapport aux modèles linéaire et LAG. Toutefois, il ne prend pas en compte la totalité de l'autocorrélation spatiale comme le montre le test suivant :

```
bptest.Sarlm(mod_sdm)
```

```
##  
## studentized Breusch-Pagan test  
##  
## data:  
## BP = 222.55, df = 14, p-value < 2.2e-16
```

Il existe d'autres modèles qui prennent mieux en compte l'hétéroscédasticité et l'autocorrélation spatiale (FLOCH et LE SAOUT, Insee, N° 131 Octobre 2018) tels que :

- les clusters spatiaux du type Newey-West
- la méthode paramétrique de type HAC