# Output and Capital Stock: a Spatial Study

*William Y. N. Suzuki*

**Este trabalho foi apresentado na disciplina de Estatística Espacial ministrada por Márcio P. Laurini no segundo semestre de 2018. Programa de pós-graduação em economia da FEA RP USP.**

The objective of this work is to analyze the relationship between economic output and factors of production in a spatial econometric framework. We use data of Brazilian municipalities to explore the relationship between TFP, physical and human capital and economic output.

Hall and Jones (1999) argue that differences in output in coutries are due to what they call "social infrastructure", which one of the elements is institutions. Institutions, as proposed by North and Thomas (1973), are the cause of factors of production. And human capital, physical capital and technology are the cause of output. We are interested in exploring the last relationship, the link between factors of production and output.

One of the advantages of using intra-country data in the case of Brazil is that we have more security in saying that the macro *de jure* institutions (the official rules written in the law, it is in contraposition with the *de facto* institutions which are what happens in reality) are the same for all the territory (Pande and Udry 2005). This is an important advantange compared with cross-country analysis given that the institutional background is more homogeneous in the case of intra-country data which means that we can have a better picture of the impact of physical and human capital on output, holding the *de jure* characteristics the same for all the territory. Given that in Brazil we have a certain degree of cultural and linguistic homegeneity, which are informal institutions that can affect economic output. So some aspects of the informal institutions are hold the same for the whole sample of analysis.

Let us now build our data set:

```
#181120 mapa Br

library(sp)
library(spdep)
```

```
## Loading required package: Matrix
```

```
## Loading required package: spData
```

```
## To access larger datasets in this package, install the spDataLarge
## package with: `install.packages('spDataLarge',
## repos='https://nowosad.github.io/drat/', type='source'))`
```

```
library(maptools)
```

```
## Checking rgeos availability: TRUE
```

```r
library(readxl)
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
library(ggplot2)
library(ape)
```

```
##
## Attaching package: 'ape'
```

```
## The following object is masked from 'package:spdep':
##
##     plot.mst
```

```r
library(ggmap)
library(plyr)
library(rgdal)
```

```
## rgdal: version: 1.3-4, (SVN revision 766)
##  Geospatial Data Abstraction Library extensions to R successfully loaded
##  Loaded GDAL runtime: GDAL 2.2.3, released 2017/11/20
##  Path to GDAL shared files: C:/Users/willi/OneDrive/Documents/R/win-library/3.5/rgdal
##  GDAL binary built with GEOS: TRUE
##  Loaded PROJ.4 runtime: Rel. 4.9.3, 15 August 2016, [PJ_VERSION: 493]
##  Path to PROJ.4 shared files: C:/Users/willi/OneDrive/Documents/R/win-library/3.5/rgd
##  Linking to sp version: 1.3-1
```

```r
library(gridExtra)
library(grid)
library(tsoutliers)
library(tseries)

setwd("C:/Users/willi/Desktop/working/Projects/RAW_DATA")

#import data of municipalities' gdp
pib <- read_excel("ipeadata_munic_pib_2000.xlsx")
pib <- as.data.frame(pib[,c(2,4)])
names(pib) <- c("cod_ibge","pib2000" ) #gdp 2000 prices, thousands
summary(pib$pib2000) #without log
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
##      1537     13434     27416    214179     67717 160285568        89
```

```
pib$pib2000 <- log(pib$pib2000) #use log of pib
pib$cod_ibge <- as.character(pib$cod_ibge)

#import data about human capital
hc <- read_excel("ipeadata_2000_capital_humano.xlsx")
hc <- as.data.frame(hc[,c(2,4)])
names(hc) <- c("cod_ibge","hc2000" )
summary(hc$hc2000) #without log
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
##     18792    115161    232119   1001518    498816 463768520        89
```

```
hc$hc2000 <- log(hc$hc2000) #log of human capital

#import data about physical capital and create names for rows in data frame
pc <- read_excel("ipeadata_2000_capital_fisico.xlsx")
pc <- as.data.frame(pc)
names(pc) <- c("UF","cod_ibge","nome","pc2000" )
summary(pc$pc2000) #without log
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.      NA's
##       906     13378     27896    222989     72375 156891685        89
```

```
pc$pc2000 <- log(pc$pc2000)

#set back the working directory to project file
setwd("C:/Users/willi/Desktop/working/Projects/181120 trabalho espacial")

#merge of pib, hc, pc
mybase <- merge(pib, hc, by='cod_ibge', all=TRUE)
mybase <- merge(mybase, pc, by='cod_ibge', all=TRUE)

#build a variable for region
region <- substr(mybase[,1],1,1)
for (i in 1:dim(mybase)[1]) {
  if (region[i] == '1') {region[i] <- "Norte"}
  if (region[i] == '2') {region[i] <- "Nordeste"}
  if (region[i] == '3') {region[i] <- "Sudeste"}
  if (region[i] == '4') {region[i] <- "Sul"}
  if (region[i] == '5') {region[i] <- "Centro-Oeste"}
}
mybase$region <- region
```

```
#load the spatial polygon data frame
#munic <- readShapePoly(
#  "C:/Users/willi/Desktop/working/Projects/RAW_DATA/map_brasil/RG2017_regioesgeografi
#  proj4string = CRS("+proj=longlat +ellps=WGS84"))
munic <- readShapePoly(
  "C:/Users/willi/Desktop/working/Projects/RAW_DATA/map_brasil/RG2017_regioesgeograficas
  proj4string = CRS("+proj=longlat +ellps=GRS80 +no_defs"))
```

```
## Warning: readShapePoly is deprecated; use rgdal::readOGR or sf::st_read
```

```
munic <- munic[,3] #use only the cod_ibge
names(munic) <- c('cod_ibge')

mybase <- merge(munic, mybase, by='cod_ibge') #merge the data with map

#delete some objects that will not be used
rm(hc,pc,pib,munic)

#chunk to drop the missings
dim(mybase) #how many were
```

```
## [1] 5570    7
```

```
mybase <- mybase[!is.na(mybase$pib2000),]
mybase <- mybase[!is.na(mybase$pc2000),]
mybase <- mybase[!is.na(mybase$hc2000),]
dim(mybase) #how many rows after
```

```
## [1] 5507    7
```

```
#chunk for analysis of neighborhood matrices
hold.nb.br <- poly2nb(mybase, queen=T)  #nb object for the whole country
summary(hold.nb.br) #let us look at  "Link number distribution"
```

```
## Neighbour list object:
## Number of regions: 5507
## Number of nonzero links: 32300
## Percentage nonzero weights: 0.1065056
## Average number of links: 5.865262
## 5 regions with no links:
## 56 419 2070 3347 4043
## Link number distribution:
##
##    0    1    2    3    4    5    6    7    8    9   10   11   12   13   14
##    5    6   87  338  926 1219 1181  764  468  259  122   68   34   17    8
##   15   16   17   18   23
##    1    1    1    1    1
```

```
## 6 least connected regions:
## 569 857 927 1265 2341 4963 with 1 link
## 1 most connected region:
## 1740 with 23 links
```

```
#from the summary we know that there are 5 municipalities that are islands
#5 of them have 0 links

#we know that hold.nb.br is a list of vectors, the vectors tell the code
#of the neighbors, if a region have no neighbors then the only code
#that appears is 0

#chunk to find out which cities are islands, i.e. which do not have any neighbor
find.islands <-  #create this function to find which municipalities are islands
  function(nb.data){   #nb.data is of class "nb"
    n <- length(nb.data) #object to hold the length
    matrix.h1 <-  matrix(rep(0,2*n),ncol=2,nrow=n) #matrix to store information
    for (i in 1:n) {
      matrix.h1[i,1] <- i #build index column, to find municipality
      matrix.h1[i,2]<-sum(nb.data[[i]]) #sum the codes of neighbor regions
    }
    matrix.h2 <- as.data.frame(matrix.h1)
    matrix.h3 <- matrix.h2[order(matrix.h2[,2]),] #order according the sum of codes
    matrix.h3[1:10,] #give the 10 first municipalities, 5 of which are islands
}

find.islands(hold.nb.br) #use the function with "nb" of Brazil
```

```
##        V1  V2
## 57     57   0
## 400   400   0
## 2018 2018   0
## 3293 3293   0
## 3986 3986   0
## 73     73 108
## 81     81 156
## 52     52 225
## 229   229 256
## 117   117 260
```

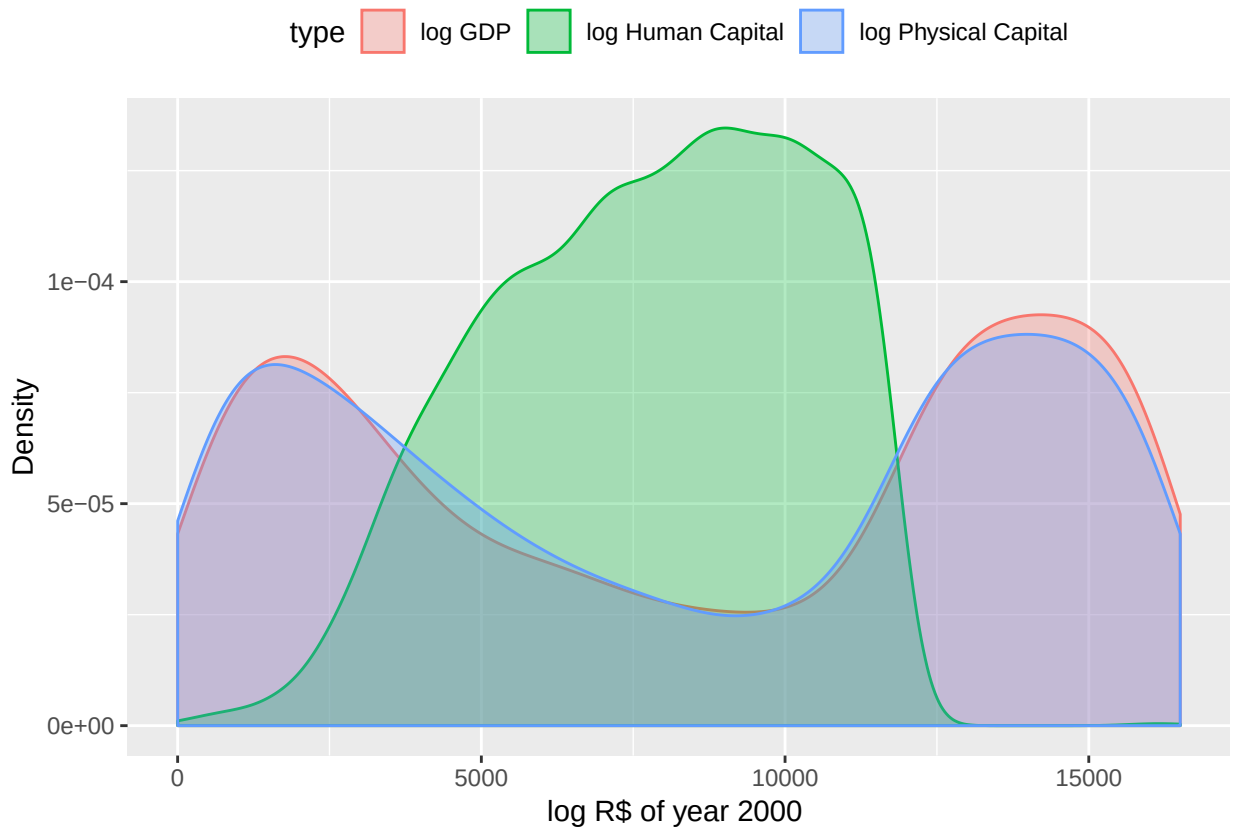
```
#the output is
#       V1  V2
#57     57   0
#400   400   0
#2018 2018   0
#3293 3293   0
```

```r
#3986 3986   0
#those are the municipalities that are islands
#drop those municipalities

#dropping municipalities that are islands
#reason: so that queen is usable

mybase <- mybase[-c(57,400,2018,3293,3986),] #drop the islands
#check if islands are gone
#summary(poly2nb(mybase, queen=T) ) #look in Link number distribution:
#there are no more regions with 0 neighbors

#chunk for visualization of data
#density graphs
density.viz <-  #build this function with the objective to visualize
  function(mybase){ #all three density plots in one graph
    df1 <- slot(mybase,'data')
    df_pib <- cbind(df1$pib2000,"log GDP")
    df_hc <- cbind(df1$hc2000,"log Human Capital")
    df_pc <- cbind(df1$pc2000,"log Physical Capital")
    df2 <- as.data.frame(rbind(df_pib,df_pc,df_hc))
    df2$V1 <- as.numeric(df2$V1)
    names(df2) <- c('value','type')
    ggplot(df2, aes(x=value, color=type, fill=type)) +
      geom_density(alpha=0.3) + theme(legend.position="top")+
      labs(x="log R$ of year 2000 ", y = "Density")
  }
density.viz(mybase)
```




The distribution of physical capital and GDP tend to follow the same format, it seems that they accumulate in convergence clubs, but human capital have a more normal-like distribution.

```
#plot of maps of physical, human capital and gdp
Brasil_uf <- readOGR(dsn="C:/Users/willi/Desktop/working/Projects/RAW_DATA/Brasil_nereus
```

```
## OGR data source with driver: ESRI Shapefile
## Source: "C:\Users\willi\Desktop\working\Projects\RAW_DATA\Brasil_nereus_uf", layer: "
## with 27 features
## It has 4 fields
## Integer64 fields read as strings:  ID
```

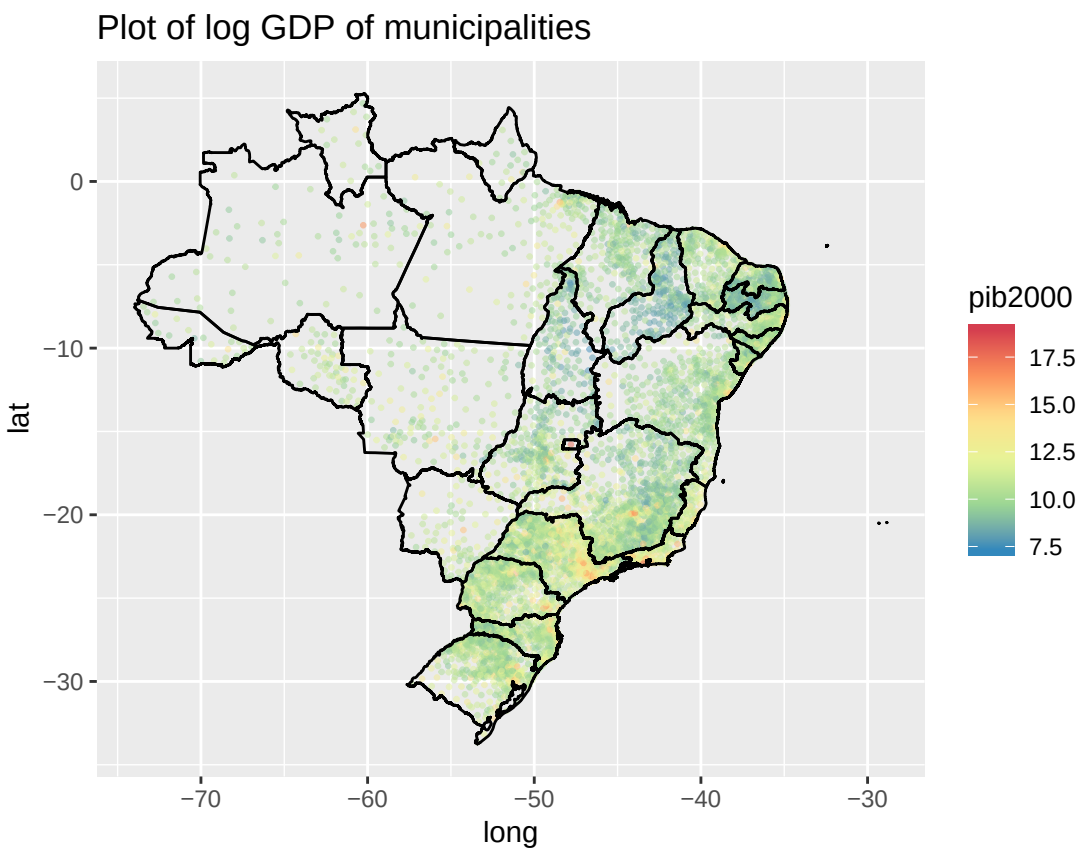
```
h1 <- Brasil_uf
h1@data$id <- rownames(h1@data)
mapa.p <- fortify(h1,region = "id")

h1.df <- cbind(coordinates(mybase),mybase@data)
names(h1.df)[c(1,2)] <- c('long','lat')

#visualization of data on maps:

gg1pib2000 <- #plot of log GDP of municipality
```

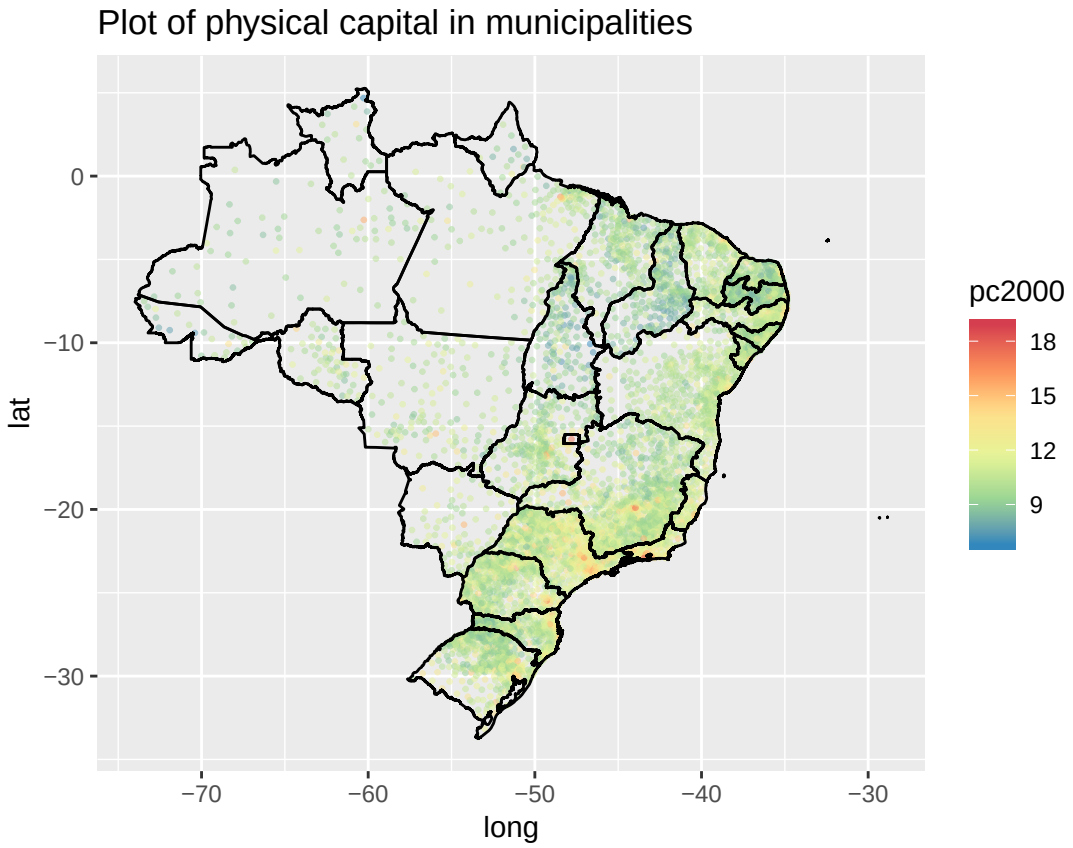
```
ggplot() +
  geom_point(data=h1.df, aes(y=lat, x=long, color=pib2000),size=.5,alpha=0.5)+
  geom_polygon(data=mapa.p,aes(long,lat,group=group), fill = NA, color = "black") +
  coord_equal()+
  scale_colour_distiller(palette = "Spectral")+
      ggtitle("Plot of log GDP of municipalities")
gg1pib2000
```



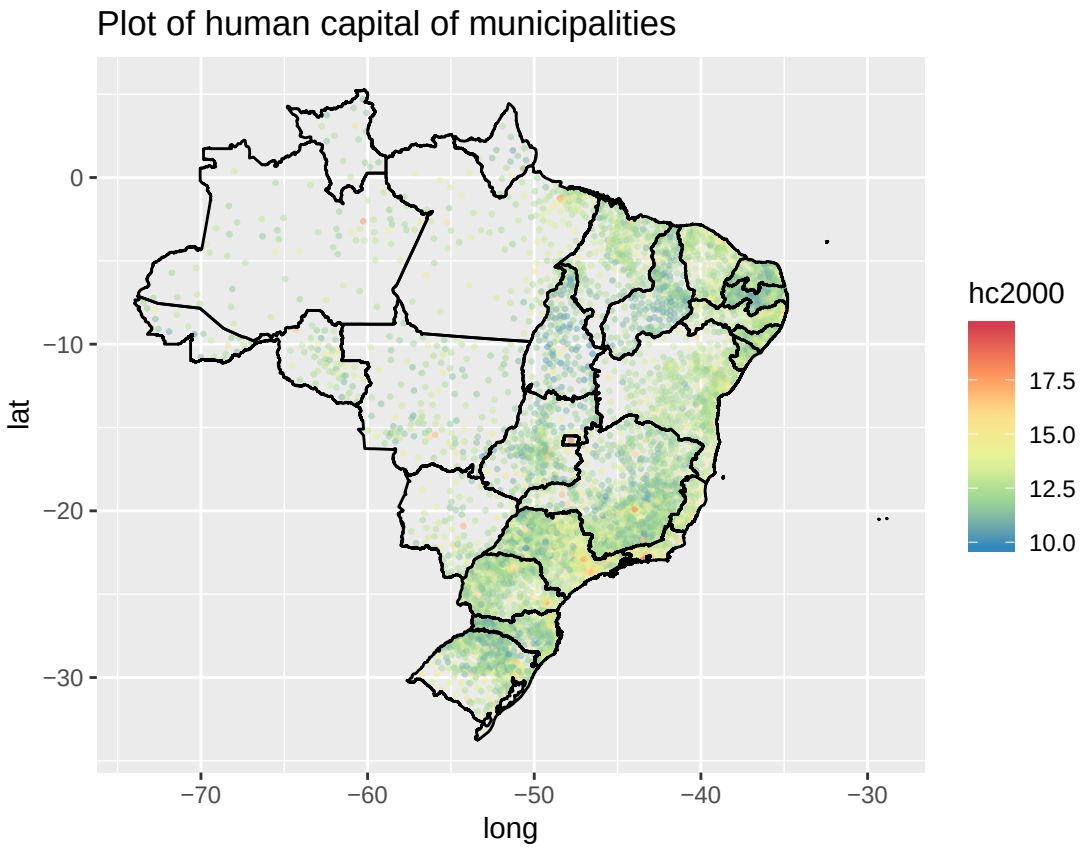

Warmer colors (red, yellow) refer to higher values for GDP and cooler colors (green, blue) refer to low levels of GDP. Note that there is a clear concentration of blue points in Piauí, Tocantins and north of Minas Gerais. Note that red points are seen in capitals of states. Around the capital os São Paulo and parts of Rio de Janeiro there is concentration of yellow points. This map shows the size of the economy in municipalities, it does not have necessarily relation with well being in municipalities.

```
gg1pc2000 <- #plot log physical capital of municipalities
ggplot() +
  geom_point(data=h1.df, aes(y=lat, x=long, color=pc2000),size=.5,alpha=0.5)+
  geom_polygon(data=mapa.p,aes(long,lat,group=group), fill = NA, color = "black") +
  coord_equal()+
  scale_colour_distiller(palette = "Spectral")+
      ggtitle("Plot of physical capital in municipalities")
```

```
gg1pc2000
```




```
gg1hc2000 <- #plot log human capital of municipalities
ggplot() +
  geom_point(data=h1.df, aes(y=lat, x=long, color=hc2000),size=.5,alpha=0.5)+
  geom_polygon(data=mapa.p,aes(long,lat,group=group), fill = NA, color = "black") +
  coord_equal()+
  scale_colour_distiller(palette = "Spectral")+
      ggtitle("Plot of human capital of municipalities")
gg1hc2000
```

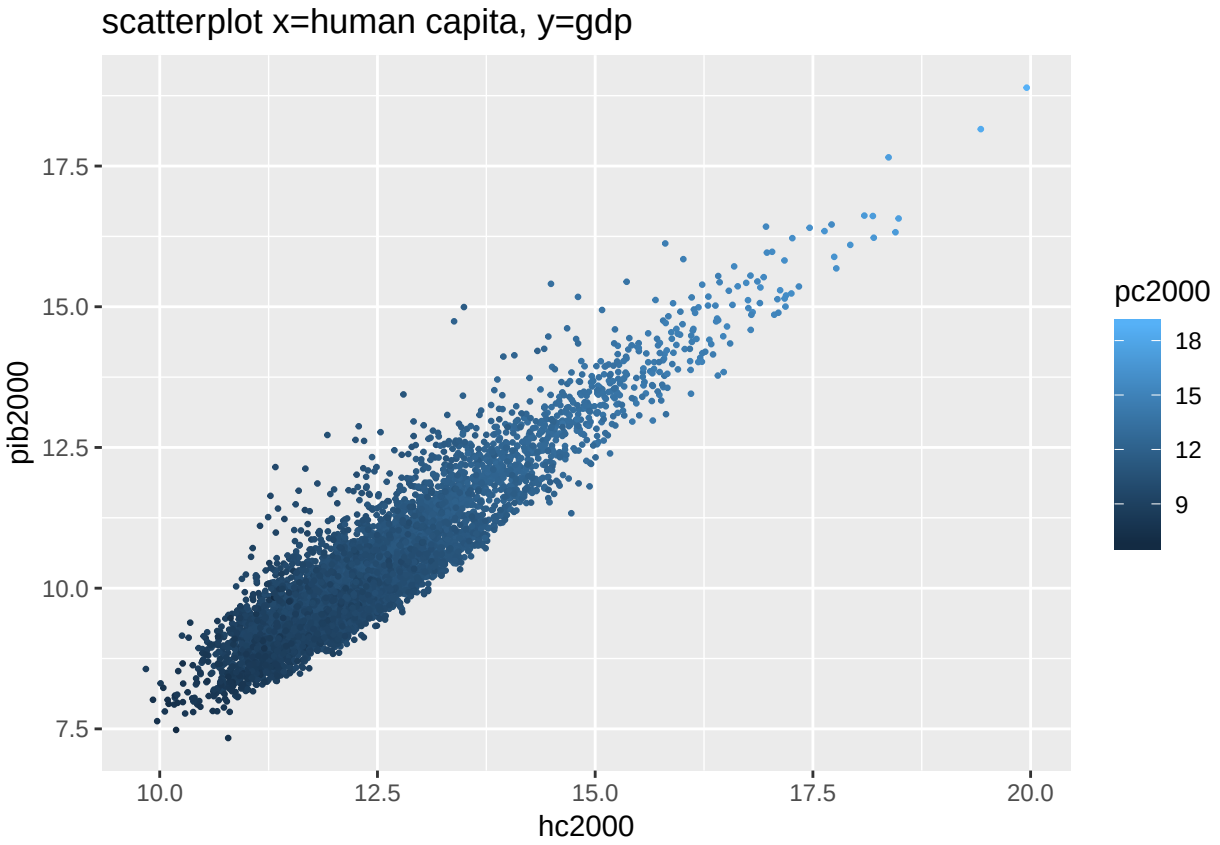



The variables `pib2000`, `pc2000` and `hc2000` capture the size of the economy inside the administrative borders of municipalities so it is hard to capture some kind welfare characteristic, we cannot make assertions about the quality of life of a municipality with lower levels of e.g. GDP, because there is the possibility that a municipality have a high GDP per person, but low aggregate GDP. Hence, the objective here is to capture spatial patterns of capital stock and output across the smallest political units. Another alternative is to plot the stock of capital and GDP per person, so that we can have a better idea of welfare measures.

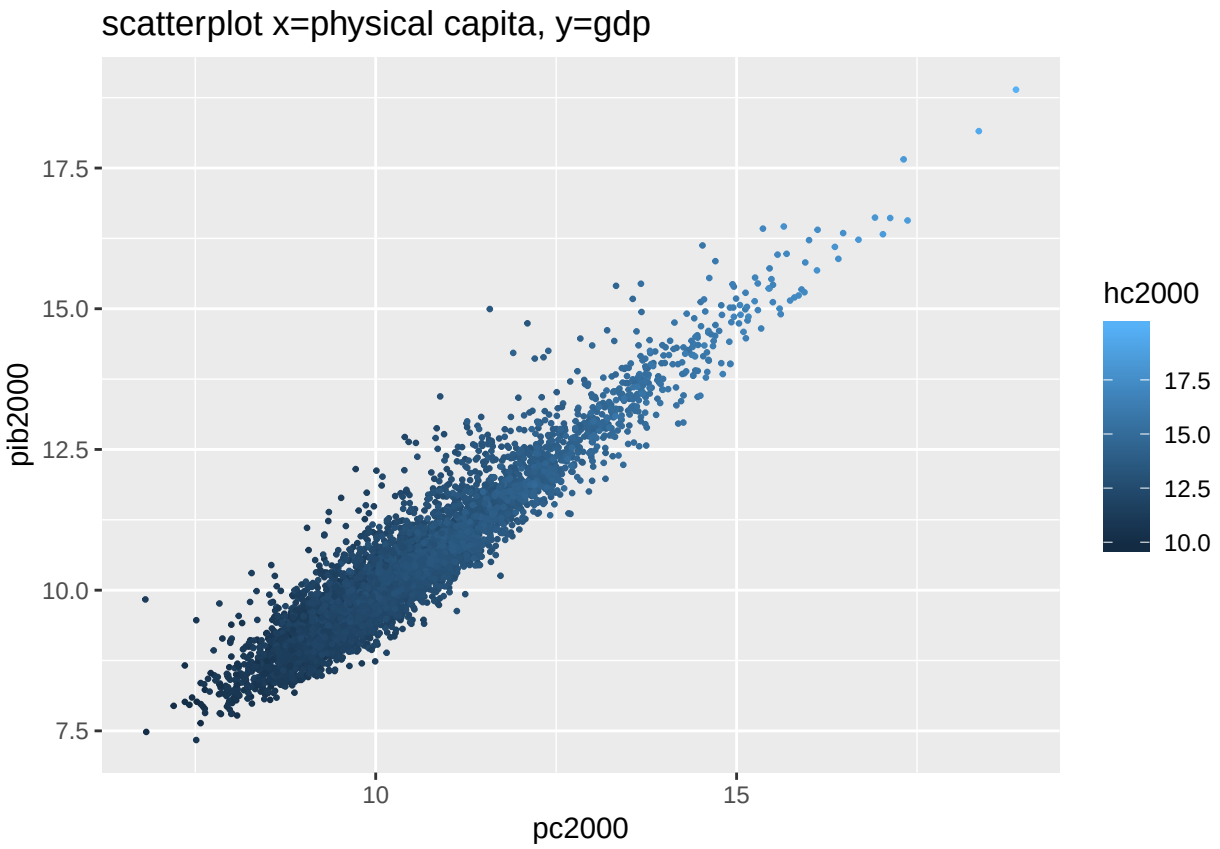
As expected in the states capitals the three variables tend to be higher. But for some parts of Nordeste we consistently have lower levels of capital stock and output. And around the region comprising the city of São Paulo, Rio de Janeiro, Curitiba and Belo Horizonte we have higher levels of capital stock and GDP.

The following are scatterplots of the values for human and physical capital and GDP in the data:

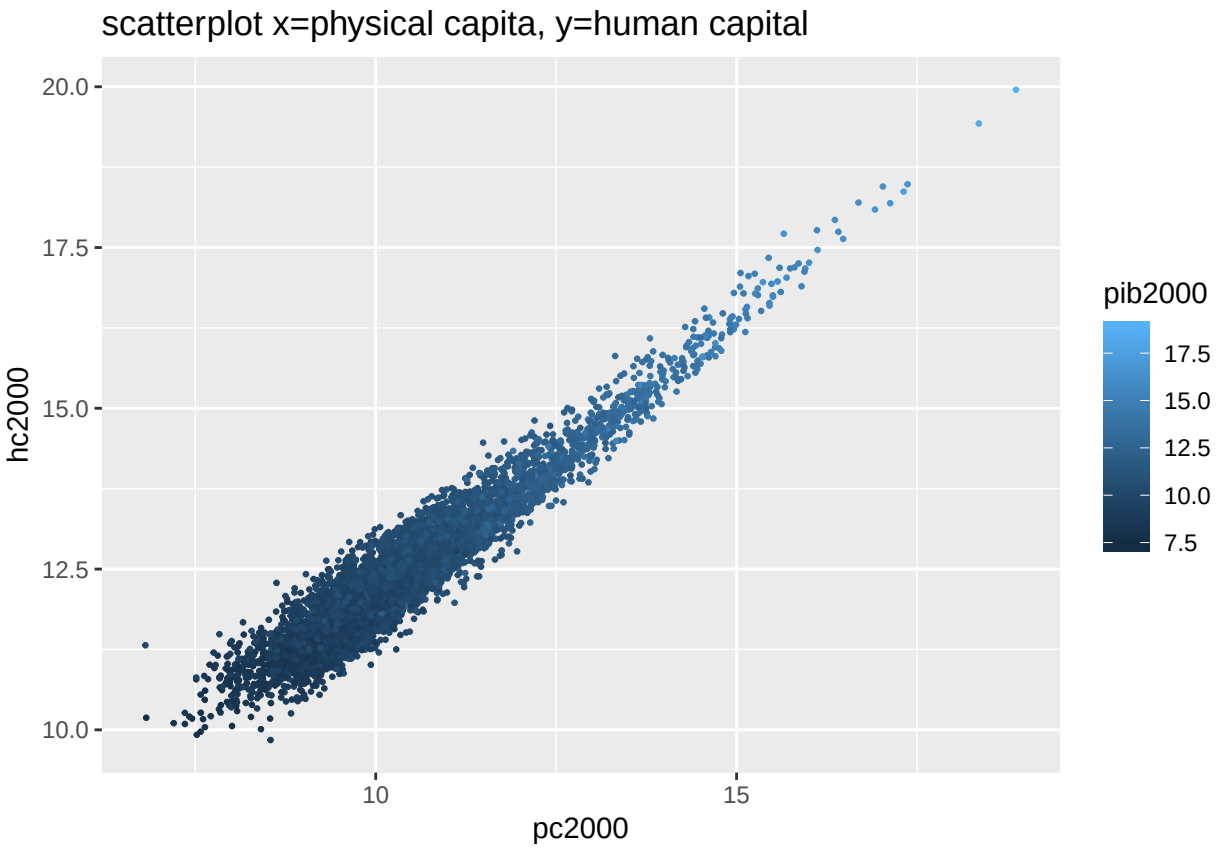
```
gg1.pib.hc <- #scatterplot x=log human capita, y=log gdp, color=physical capital
ggplot(data=slot(mybase,'data'), aes(hc2000, pib2000, colour = pc2000))+
  geom_point(size=.5)+
  ggtitle("scatterplot x=human capita, y=gdp")
gg1.pib.hc
```




```
gg1.pib.pc <- #scatterplot x=physical capita, y=gdp, color=human capital
ggplot(data=slot(mybase,'data'), aes(pc2000, pib2000, colour = hc2000))+
  geom_point(size=.5)+
  ggtitle("scatterplot x=physical capita, y=gdp")
gg1.pib.pc
```




```
gg1.hc.pc <- #scatterplot x=physical capita, y=human capital, color=GDP
ggplot(data=slot(mybase,'data'), aes(pc2000, hc2000, colour = pib2000))+
  geom_point(size=.5)+
  ggtitle("scatterplot x=physical capita, y=human capital")
gg1.hc.pc
```




The scatterplots of the relation between human and physical capital and GDP are well behaved, the log of the variables follow linear relationships.

As expected the relation between capital stocks and GDP follow a well behaved linear relationship in the scatterplots. It is important though to see any spatial patterns that arise in the relationship, which is difficult to capture through scatterplots.

## Motivation and Non-Spatial Analysis

To motivate the use of spatial econometrics we begin with the non-spatial approach. We assume a constant returns to scale Cobb-Douglas production function:

$$Y = AK^{\alpha}H^{1-\alpha}$$

in which for each region we have $Y$, $K$, $H$ and $A$, the output, physical and human capital stock and total factor productivity.

We apply log on the production function to obtain:

$$\ln Y = \alpha \ln K + (1-\alpha) \ln H + \ln A \tag{1}$$

From this specification we should observe that the estimator for human capital is the unit complement of physical capital, i.e. the coefficients are $\alpha$ and $1-\alpha$. And notice that $\ln A$

will act as the error term, capturing everything else that affect output, that is not included in $K$ and $H$.

Making the OLS regression with the whole sample of Brazil:

```
#analsis for Brasil
basetemp <- mybase
m1 <- lm(pib2000 ~ -1 + pc2000 + hc2000, data=basetemp)
summary(m1)
```

```
##
## Call:
## lm(formula = pib2000 ~ -1 + pc2000 + hc2000, data = basetemp)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3839 -0.3185 -0.0650  0.2476  3.4924
##
## Coefficients:
##        Estimate Std. Error t value Pr(>|t|)
## pc2000 0.762642   0.011341   67.25   <2e-16 ***
## hc2000 0.197871   0.009531   20.76   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4585 on 5500 degrees of freedom
## Multiple R-squared:  0.9981, Adjusted R-squared:  0.9981
## F-statistic: 1.456e+06 on 2 and 5500 DF,  p-value: < 2.2e-16
```

```
coef(m1)[1]+coef(m1)[2]
```

```
##   pc2000
## 0.960513
```

```
e <- residuals(m1)
```

Note that `pib2000` is the log of municipality's GDP in year 2000, `pc2000` is the physical capital stock of municipalities in 2000, and `hc2000` is the human capital stock of municipalities in 2000.

Note that the regression results are remarkbly compatible with the adopted theoretical framework. The result for log of physical capital is 0.762642 and for the log of human capital is 0.197871, their sum should be equal to unity, doing the sum of the estimators we find 0.960513. The estimators are all statistically significant. And notice that according to equation 1 the specification have no intercept, and that the average of residuals is different from zero, it is -0.0013115.

The values of coefficients found in this regression are elasticities hence 1% increase in capital stock according to this OLS will increase in 0.76% GDP, in the same way 1% increase in

human capital increase in 0.198% GDP. Further analysis will show that those elasticities change drastically depending on the spatial specification. And regional analysis not shown here present also strong heterogeneity in the values for those coefficients.

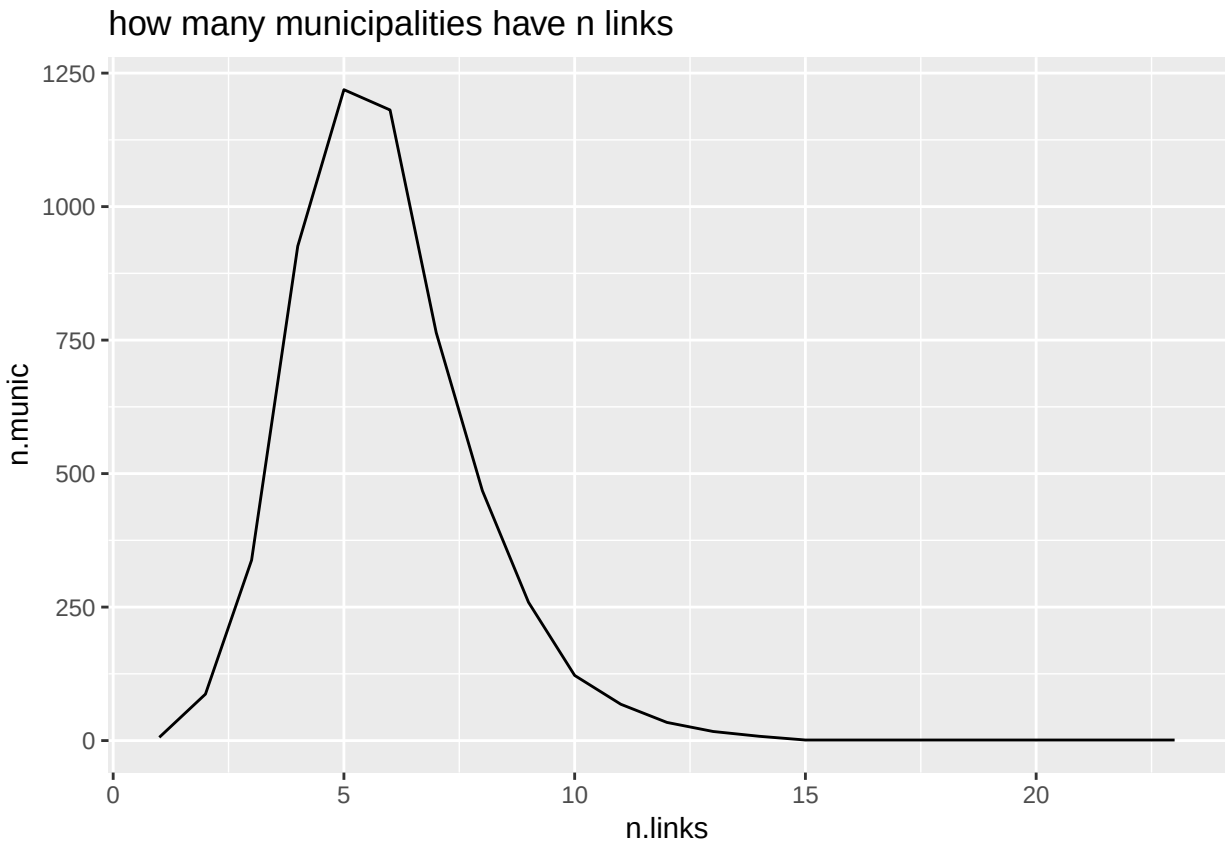
The $R^2$ is quite high, 0.9981, but we find that the residuals have spatial correlation as the following test shows:

```
#test for spatial dependence on redisuals
h1.nb <- poly2nb(basetemp, queen=T)  #using queen criterion
summary(h1.nb)
```

```
## Neighbour list object:
## Number of regions: 5502
## Number of nonzero links: 32300
## Percentage nonzero weights: 0.1066992
## Average number of links: 5.870593
## Link number distribution:
##
##    1    2    3    4    5    6    7    8    9   10   11   12   13   14   15
##    6   87  338  926 1219 1181  764  468  259  122   68   34   17    8    1
##   16   17   18   23
##    1    1    1    1
## 6 least connected regions:
## 569 857 927 1265 2341 4963 with 1 link
## 1 most connected region:
## 1740 with 23 links
```

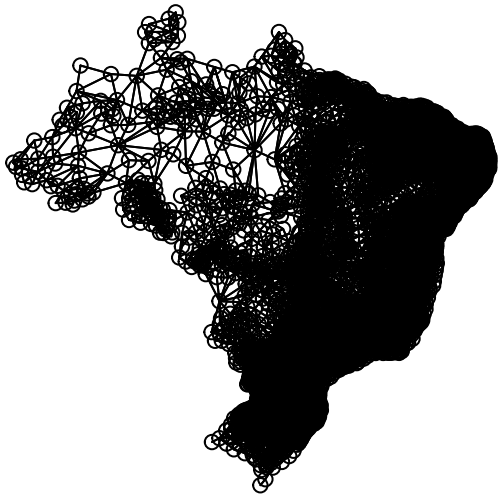
```
n.links.queen.br <-as.data.frame(  #number of links data frame
cbind(c(1,  2,  3,  4,  5,  6,  7,  8,  9,  10, 11, 12,
        13, 14, 15, 16, 17, 18, 23),
c(6,    87, 338,    926,    1219,   1181,   764,    468,    259,    122,
  68,   34, 17, 8,  1,  1,  1,  1,  1)))
names(n.links.queen.br) <-c("n.links","n.munic")

#make graph of how many municipalities have n number of links
gg1.n.links.queen.br <- #plot of relation number
  ggplot()+  # of links and quantity of municipalities
  geom_line(data=n.links.queen.br, aes(n.links, n.munic))+
  ggtitle("how many municipalities have n links")
gg1.n.links.queen.br
```




```
#we can see that the majority of municipalities gather around 4~7 links each.
#hold.nb <- dnearneigh(coordinates(basetemp), 0, 100, longlat=T)

#plot the links built in queen criterion
plot.nb(h1.nb,coordinates(basetemp))
```



```r
#those are the links that are used in the weight matrices in the queen
#criterion,  it is hard to see anything
W1<-nb2listw(h1.nb, glist=NULL, style ="W") #normlize weights to 1
W1$neighbours
```

```
## Neighbour list object:
## Number of regions: 5502
## Number of nonzero links: 32300
## Percentage nonzero weights: 0.1066992
## Average number of links: 5.870593
```

```r
lm.morantest(m1, W1) #Moran's I test for residuals
```

```
##
##  Global Moran I for regression residuals
##
## data:
## model: lm(formula = pib2000 ~ -1 + pc2000 + hc2000, data =
## basetemp)
## weights: W1
##
## Moran I statistic standard deviate = 49.763, p-value < 2.2e-16
```

```
## alternative hypothesis: greater
## sample estimates:
## Observed Moran I      Expectation         Variance
##     4.019006e-01    -3.060278e-04     6.532637e-05
```

```
#the p-value shows that there is spatial dependence.
```

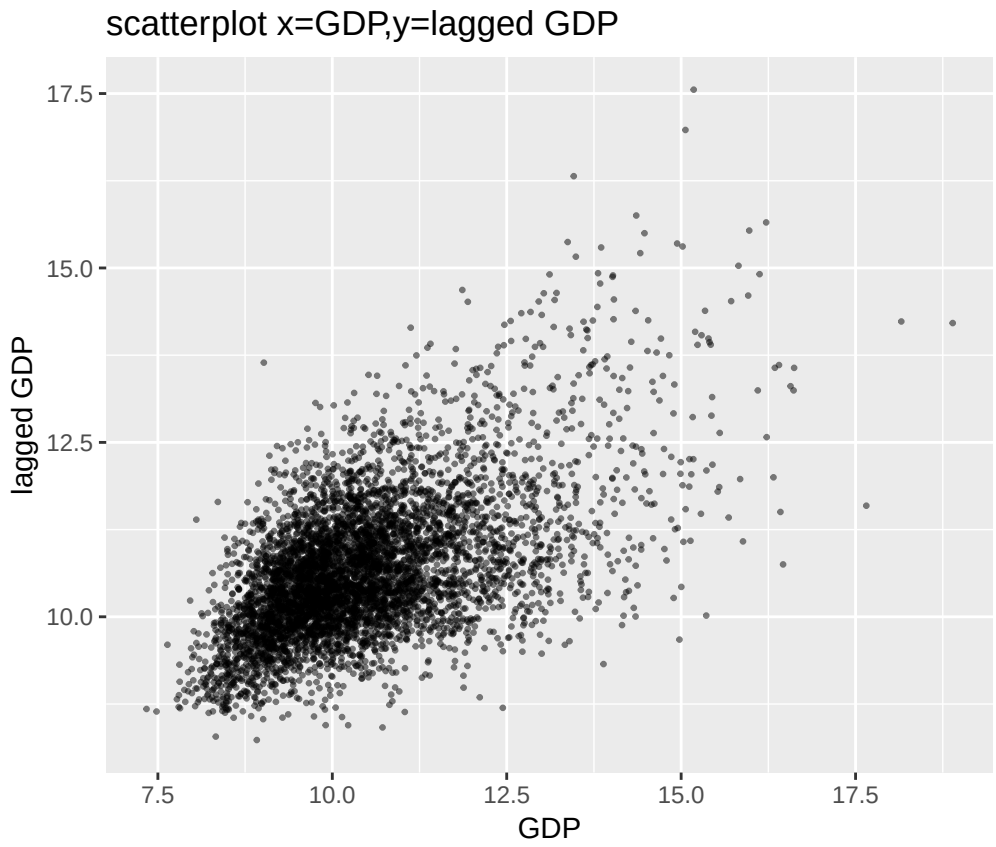
We can see that the residuals have spatial dependence. This is indication that it is needed to use a spatial specification to deal with the relationship output, human and physical capital.

But we can also verify the spatial dependence of the variables (output, capital stock) in our data. We calculate the Moran's I for physical and human capital and output:

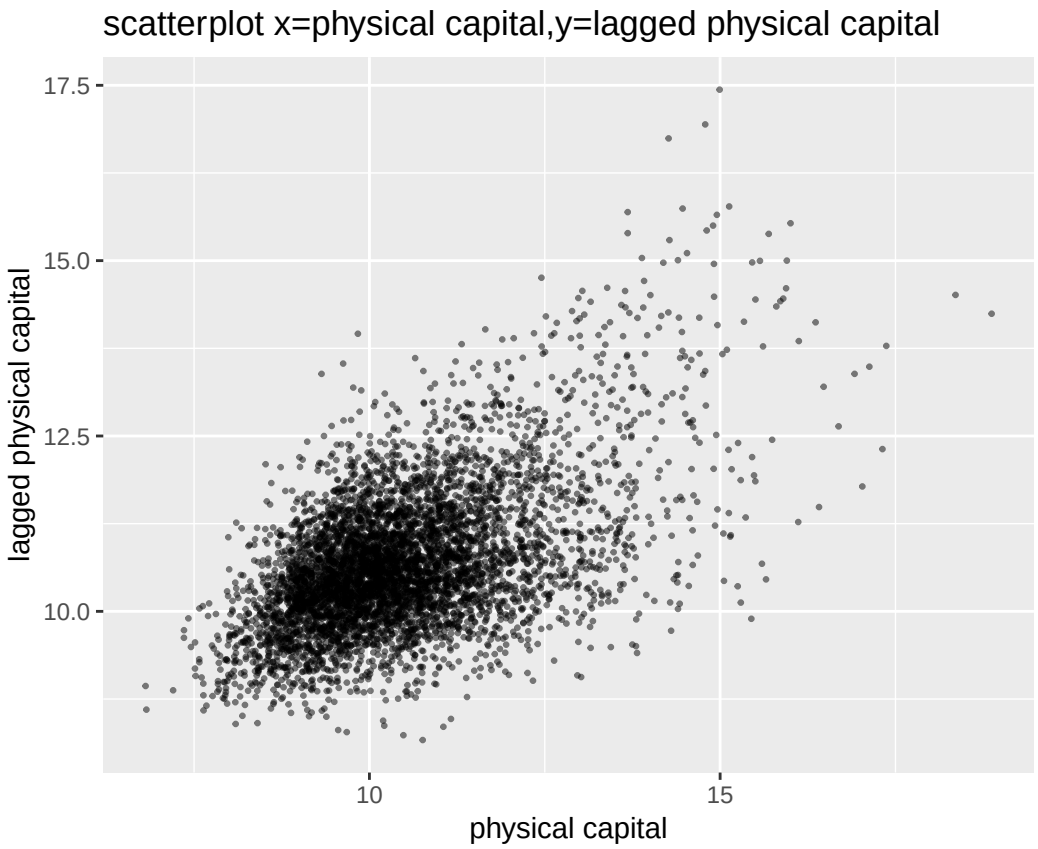
```
#Moran's I test for GDP
moran.test(mybase@data$pib2000,W1) #W1 is on queen criterion
```

```
##
##  Moran I test under randomisation
##
## data:  mybase@data$pib2000
## weights: W1
##
## Moran I statistic standard deviate = 50.517, p-value < 2.2e-16
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##      4.081864e-01     -1.817851e-04      6.534815e-05
```

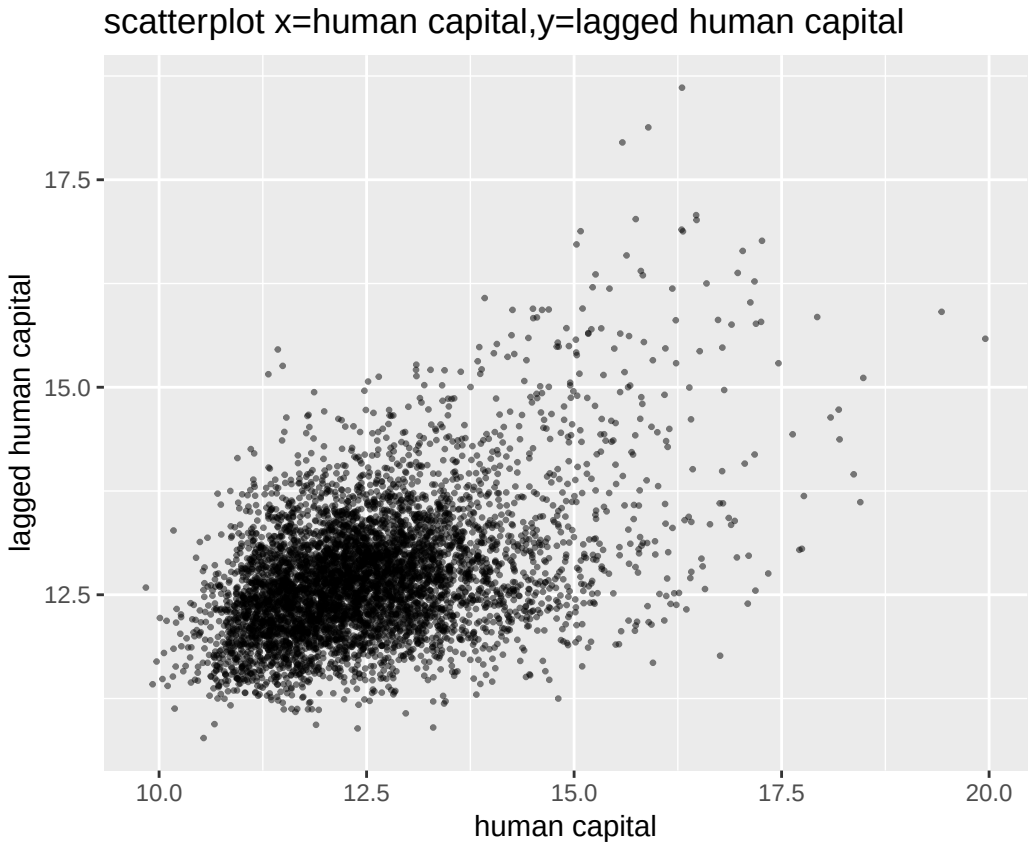
```
#plot for GDP and spatially lagged GDP
scatter.plot.spatial.lag <-
  function(vec1,W){
    vec1lag <- lag.listw(W,vec1)
    h2.df <- as.data.frame(cbind(vec1,vec1lag))
    gg1.gdp.lagged.gdp <- #scatterplot x=vec, y=lagged vec1
      ggplot(data=h2.df, aes(y=vec1lag, x=vec1)) +
      geom_point(size=.5,alpha=0.5)+
      coord_equal()+
      ggtitle("scatterplot x=GDP,y=lagged GDP ")+
      xlab("GDP") + ylab("lagged GDP")
    gg1.gdp.lagged.gdp
  }
scatter.plot.spatial.lag(mybase@data$pib2000,W1)
```




```
#plot for physical capital and spatially lagged physical capital
scatter.plot.spatial.lag <-
  function(vec1,W){
    vec1lag <- lag.listw(W,vec1)
    h2.df <- as.data.frame(cbind(vec1,vec1lag))
    gg1.gdp.lagged.gdp <- #scatterplot x=vec, y=lagged vec1
      ggplot(data=h2.df, aes(y=vec1lag, x=vec1)) +
      geom_point(size=.5,alpha=0.5)+
      coord_equal()+
      ggtitle("scatterplot x=physical capital,y=lagged physical capital")+
      xlab("physical capital") + ylab("lagged physical capital")
    gg1.gdp.lagged.gdp
  }
scatter.plot.spatial.lag(mybase@data$pc2000,W1)
```




```
#plot for human capital and spatially lagged human capital
scatter.plot.spatial.lag <-
  function(vec1,W){
    vec1lag <- lag.listw(W,vec1)
    h2.df <- as.data.frame(cbind(vec1,vec1lag))
    gg1.gdp.lagged.gdp <- #scatterplot x=vec, y=lagged vec1
      ggplot(data=h2.df, aes(y=vec1lag, x=vec1)) +
      geom_point(size=.5,alpha=0.5)+
      coord_equal()+
      ggtitle("scatterplot x=human capital,y=lagged human capital")+
      xlab("human capital") + ylab("lagged human capital")
    gg1.gdp.lagged.gdp
  }
scatter.plot.spatial.lag(mybase@data$hc2000,W1)
```




The scatterplots show that there is a correlation pattern between the variable and its lag.

```
#a test for robustness of spatial dependence
moran.mc(mybase@data$pib2000,W1,nsim=1000) #with 1000 simulations
```

```
##
##  Monte-Carlo simulation of Moran I
##
## data:  mybase@data$pib2000
## weights: W1
## number of simulations + 1: 1001
##
## statistic = 0.40819, observed rank = 1001, p-value = 0.000999
## alternative hypothesis: greater
```

```
#the test above randomizes the weights given the
#spatial weighting criterion in the matrix and test for
#Moran's I. We test 1000 simulations. The result is that there is
#spatial dependence.

#Moran's I test for physical capital
moran.test(mybase@data$pc2000,W1) #W1 is on queen criterion
```

```
##
```

```
##  Moran I test under randomisation
##
## data:  mybase@data$pc2000
## weights: W1
##
## Moran I statistic standard deviate = 47.142, p-value < 2.2e-16
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##      3.809106e-01     -1.817851e-04      6.534931e-05
```

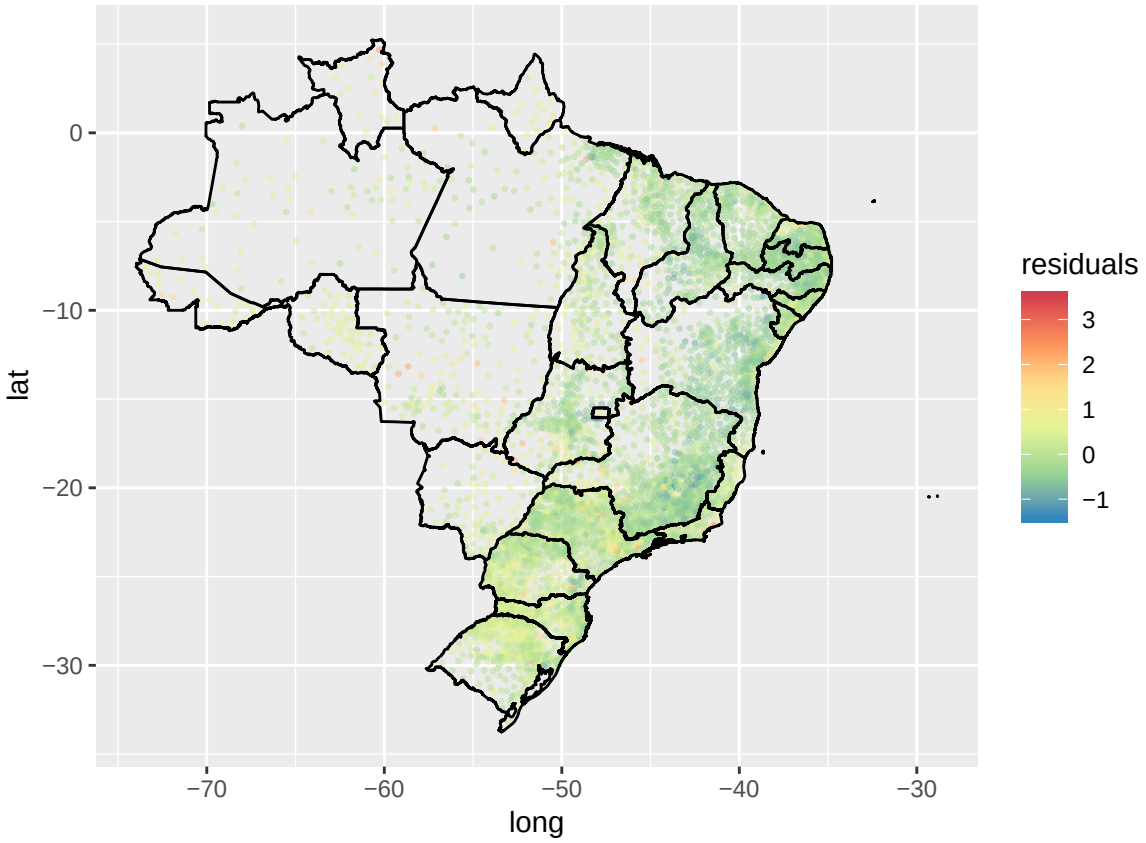
```
#Moran's I test for human capital
moran.test(mybase@data$hc2000,W1) #W1 is on queen criterion
```

```
##
##  Moran I test under randomisation
##
## data:  mybase@data$hc2000
## weights: W1
##
## Moran I statistic standard deviate = 36.228, p-value < 2.2e-16
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##      2.926706e-01     -1.817851e-04      6.534525e-05
```

Tests for physical and human capital with `moran.mc` (simulation randomization of weights) gave the same result of spatial dependence. The tests show that all variables in our data have spatial dependence. We can cofirm what we see on the maps: the presence of spatial clusters on physical and human capital and output.

Now we show the map of residuals of the non-spatial linear model of Brasil:

```
#plot map of residuals of Brasil non spatial regression
basetemp$residuals <- e #the residuals of linear model of Brasil
basetemp.df<- cbind(coordinates(basetemp), basetemp@data)
names(basetemp.df)[c(1,2)] <- c('long','lat')
gg1.res.lm.br <- #plot map residual linear model of Brasil
  ggplot() +
  geom_point(data=basetemp.df, aes(y=lat, x=long, color=residuals),size=.5,alpha=0.5)+
  geom_polygon(data=mapa.p,aes(long,lat,group=group), fill = NA, color = "black") +
  coord_equal()+
  scale_colour_distiller(palette = "Spectral")
gg1.res.lm.br
```




This map captures the term $\ln A$ in equation 1, the TFP or productivity. Warmer colors (red and yellow) mean that the region have a relatively high level of productivity, cooler colors mean that the region have low productivity, given the amount of capital stock that it has.

It is noticible a more bluish tone in parts of Minas Gerais and Nordeste. Region Norte and parts of Centro-Oeste tend to have more yellow greenish tones. The Sul and Sudeste have regions with blue tones but we can notice a more yellow-greenish tone. Red points are scattered throughout the country.

Residual average per region, and t-statistic:

```
#we can test if the residuals have different averages depending on the region

#average of residuals Centro-Oeste
mean(basetemp.df[basetemp.df$region=='Centro-Oeste',]$residuals)
```

```
## [1] 0.1788969
```

```
#average of residuals Sul
mean(basetemp.df[basetemp.df$region=='Sul',]$residuals)
```

```
## [1] 0.2097141
```

```r
#average of residuals Nordeste
mean(basetemp.df[basetemp.df$region=='Nordeste',]$residuals)
```

```
## [1] -0.1949075
```

```r
#average of residuals Norte
mean(basetemp.df[basetemp.df$region=='Norte',]$residuals)
```

```
## [1] 0.2564695
```

```r
#average of residuals Sudeste
mean(basetemp.df[basetemp.df$region=='Sudeste',]$residuals)
```

```
## [1] -0.058146
```

We can see that in Nordeste the residual mean is lower than the other regions as expected, but Sudeste also have negative average which was not expected. But the t-statistic (not presented here) shows that all of those values can be considered statistically zero.

Next are presented tests for normality and homoskestacity in the residuals:

```r
#test for normality and heteroskedasticity of residuals

bptest(m1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  m1
## BP = 15.699, df = 1, p-value = 7.427e-05
```

```r
#the residuals shows heteroskedasticity

jarque.bera.test(m1$residuals)
```

```
##
##  Jarque Bera Test
##
## data:  m1$residuals
## X-squared = 3352.6, df = 2, p-value < 2.2e-16
```

```r
#the residuals are not normally distributed
```

## Data Source

Let us understand the source of data used here. The data about human and physical capital and GDP in municipalities were downloaded in the site of IPEA (Institute for Applied

Economic Research). All of them refer to the year of 2000, and are measured as reais (R$) os 2000.

Human capital is measured in monetary terms. It is the expected value of present retuns, assumed 10% rate of return, for years of schooling. Accounting also for the age and number of the working population (16~65 age). The return for education is calculated estimating the difference between wages of someone with education and someone without formal education, the data used to calulate those estimates are from the demographic Census and PNAD.

The physical capital data is calculated as the present value of infinite returns of 0.75% per month on proporty value. Rent values of building is simulated using hedonic models in which building characteristics are taken into account, those characteristics are location and average income. The data used by IPEA comes from the national Census survey.

## Spatial Analysis

### Spatial Durbin Model

Let us test the Spatial Durbin Model (SDM) with the data of the whole country:

$$y = X\beta_1 + WX\beta_2 + u$$

in which there is no estimation issues when applying OLS (Arbia 2014). To use the SDM we need to propose a production function as:

$$Y = AK^{\alpha_1}(WK)^{\alpha_2}H^{\beta_1}(WH)^{\beta_2}$$

Notice that the neighboring capital stock enter in a product form to determine the output. The drawback of this kind of production function is that if we have zero capital stock in the neighbor regions (either human or physical) then we have zero production in ours, which is not a reasonable outcome. Given that a region could be isolated and still be able to produce output.

But let us explore what this specification tell us. Taking log:

$$\ln Y = \alpha_1 \ln K + \alpha_2 \ln WK + \beta_1 \ln H + \beta_2 \ln WH + \ln A$$

$\ln A$ behaves here as the error term.

```
#no issues involved in applying OLS in the spatial Durbin
WK<- log(lag.listw(W1, exp(mybase@data$pc2000))) #first apply the neighbor matrix and
WH<-log(lag.listw(W1, exp(mybase@data$hc2000))) #then take the log
m2 <- lm(pib2000~-1+pc2000+hc2000+WK+WH,data=mybase@data)
#exclude the intercept according to the model
summary(m2)
```

```
##
## Call:
## lm(formula = pib2000 ~ -1 + pc2000 + hc2000 + WK + WH, data = mybase@data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.3635 -0.3021 -0.0602  0.2294  3.3099
##
## Coefficients:
##        Estimate Std. Error t value Pr(>|t|)
## pc2000  0.33276    0.02975   11.18   <2e-16 ***
## hc2000  0.66201    0.03317   19.96   <2e-16 ***
## WK      0.55216    0.03403   16.23   <2e-16 ***
## WH     -0.57053    0.03686  -15.48   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4472 on 5498 degrees of freedom
## Multiple R-squared:  0.9982, Adjusted R-squared:  0.9982
## F-statistic: 7.651e+05 on 4 and 5498 DF,  p-value: < 2.2e-16
```

It is plausible to say that neighboring capital stock benefit output in a region through spillover effects. That is, the hypothesis here is that `WK` and `WH` should have positive coefficients.

The result is not promising, the sign of neigboring physical capital is in conformity with our hypothesis but neighboring human capital is not, it is negative. In the case of pure spatial Durbin interpretation is straightforward like the usual OLS, there is no need to use `impacts()`. The following table shows the comparion with OLS and pure SDM:

```
table.OLS.SDM <- matrix(c( 0.763, 0.333,
                           0.198, 0.662,
                            NA  , 0.552,
                            NA  ,-0.571 ),nrow = 4,ncol=2,byrow = TRUE)
table.OLS.SDM <- as.data.frame(table.OLS.SDM)
rownames(table.OLS.SDM) <- c("pc2000","hc2000","WK","WH")
colnames(table.OLS.SDM) <- c("OLS","SDM")

knitr::kable(
  table.OLS.SDM, caption = 'results of OLS and SDM'
)
```

Table 1: results of OLS and SDM

| | OLS | SDM |
|---|---|---|
| pc2000 | 0.763 | 0.333 |
| hc2000 | 0.198 | 0.662 |

| | OLS | SDM |
|---|---|---|
| WK | NA | 0.552 |
| WH | NA | -0.571 |

Note that there is a huge shift in the coefficients depending on the inclusion of the neighboring covariates. Physical capital had an effect of 0.76 and now it have 0.33, but it is important also to see the indirect and direct effects how they work in this case. Again it is unexpected that human capital have a negative impact in neighboring regions.

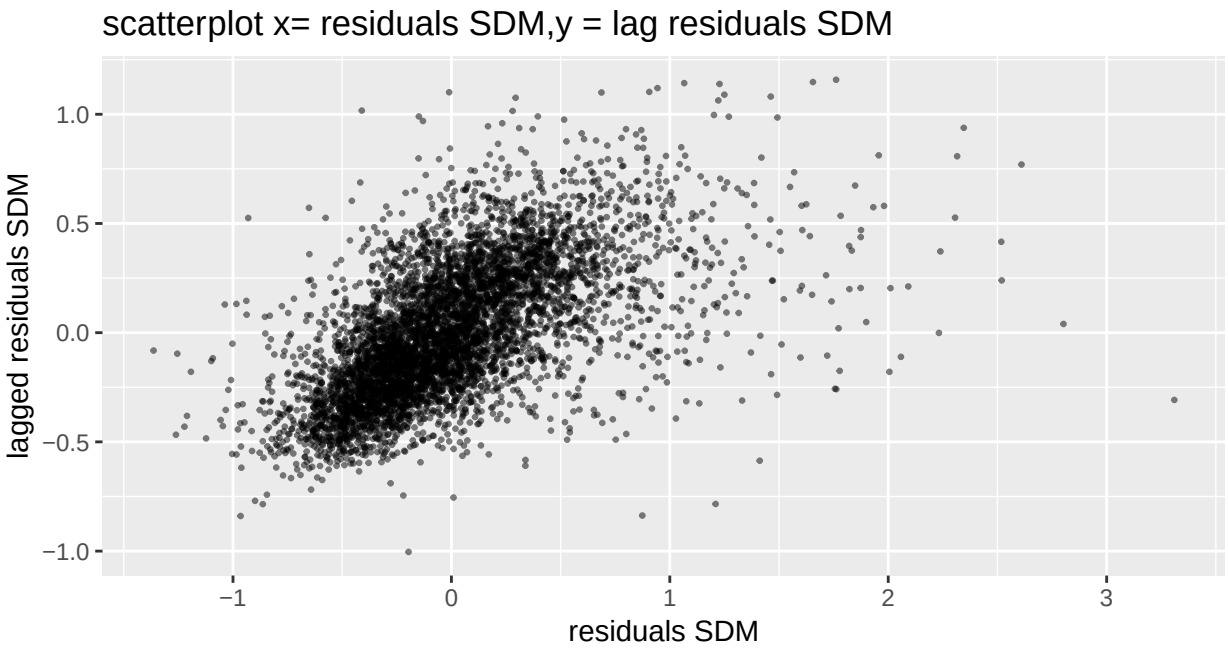
Make analysis with the residuals of this regression:

```
e.SDM <- residuals(m2)
lm.morantest(m2, W1) #Moran's I test for residuals
```

```
##
##  Global Moran I for regression residuals
##
## data:
## model: lm(formula = pib2000 ~ -1 + pc2000 + hc2000 + WK + WH, data
## = mybase@data)
## weights: W1
##
## Moran I statistic standard deviate = 51.061, p-value < 2.2e-16
## alternative hypothesis: greater
## sample estimates:
## Observed Moran I      Expectation         Variance
##     4.124027e-01    -2.996182e-04     6.532715e-05
```

```
#there is still strong spatial correlation, even applying SDM
```

```
scatter.plot.spatial.lag.residuals.SDM <-
  function(vec1,W){
    vec1lag <- lag.listw(W,vec1)
    h2.df <- as.data.frame(cbind(vec1,vec1lag))
    gg1.gdp.lagged.gdp <- #scatterplot x=vec, y=lagged vec1
      ggplot(data=h2.df, aes(y=vec1lag, x=vec1)) +
      geom_point(size=.5,alpha=0.5)+
      coord_equal()+
      ggtitle("scatterplot x= residuals SDM,y = lag residuals SDM")+
      xlab("residuals SDM") + ylab("lagged residuals SDM")
    gg1.gdp.lagged.gdp
  }
scatter.plot.spatial.lag.residuals.SDM(e.SDM,W1)
```




Clearly the scatterplot and the Moran's I statistic show the presence of spatial correlation on the residuals.

We then experiment with the Spatial Error Model (SEM):

### Spatial Error Model

The SEM specification follows:

$$y = X\beta_1 + WX\beta_2 + u \qquad\qquad u = \lambda Wu + \varepsilon$$

In our case of interest, where $\ln A = u$ , the autoregressive error structure tells that the productivity of neighboring regions have an effect on the productivity of the region, it is a spill-over effect of TFP.

First we begin by applying SEM without neighbor covariates:

```
#SEM non Durbin ML
m.SEM1.nDurbin <- errorsarlm(pib2000~ -1 + pc2000 + hc2000,
                      data = mybase@data, listw = W1)
summary(m.SEM1.nDurbin)
```

```
##
```

```
## Call:
## errorsarlm(formula = pib2000 ~ -1 + pc2000 + hc2000, data = mybase@data,
##     listw = W1)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -1.322780 -0.232806 -0.061771  0.154814  3.636587
##
## Type: error
## Coefficients: (asymptotic standard errors)
##        Estimate Std. Error z value  Pr(>|z|)
## pc2000 0.587626   0.014097  41.684 < 2.2e-16
## hc2000 0.344322   0.011977  28.747 < 2.2e-16
##
## Lambda: 0.65701, LR test value: 1786.1, p-value: < 2.22e-16
## Asymptotic standard error: 0.013364
##     z-value: 49.161, p-value: < 2.22e-16
## Wald statistic: 2416.8, p-value: < 2.22e-16
##
## Log likelihood: -2622.313 for error model
## ML residual variance (sigma squared): 0.13822, (sigma: 0.37179)
## Number of observations: 5502
## Number of parameters estimated: 4
## AIC: 5252.6, (AIC for lm: 7036.7)
```

```
#SEM non Durbin FGLS
m.SEM1.FGLS.nDurbin <- GMerrorsar(pib2000~ -1 + pc2000 + hc2000,
                          data = mybase@data, listw = W1)
summary(m.SEM1.FGLS.nDurbin)
```

```
##
## Call:
## GMerrorsar(formula = pib2000 ~ -1 + pc2000 + hc2000, data = mybase@data,
##     listw = W1)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -1.318855 -0.329687 -0.063774  0.254994  3.537179
##
## Type: GM SAR estimator
## Coefficients: (GM standard errors)
##        Estimate Std. Error z value  Pr(>|z|)
## pc2000 0.616491   0.013832  44.570 < 2.2e-16
## hc2000 0.319993   0.011687  27.381 < 2.2e-16
##
```

```
## Lambda: 0.56859 (standard error): 0.03011 (z-value): 18.884
## Residual variance (sigma squared): 0.14628, (sigma: 0.38247)
## GM argmin sigma squared: 0.14703
## Number of observations: 5502
## Number of parameters estimated: 4
```

```
e.SEM1.nDurbin <- residuals(m.SEM1.nDurbin) #SEM non Durbin with ML
moran.test(e.SEM1.nDurbin,W1) #W1 is on queen criterion
```

```
##
##  Moran I test under randomisation
##
## data:  e.SEM1.nDurbin
## weights: W1
##
## Moran I statistic standard deviate = -7.0375, p-value = 1
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##     -5.704709e-02     -1.817851e-04      6.529115e-05
```

```
#NO spatial correlation

e.SEM1.FGLS.nDurbin <- residuals(m.SEM1.FGLS.nDurbin) #SEM non Durbin with FGLS
moran.test(e.SEM1.FGLS.nDurbin,W1) #W1 is on queen criterion
```

```
##
##  Moran I test under randomisation
##
## data:  e.SEM1.FGLS.nDurbin
## weights: W1
##
## Moran I statistic standard deviate = 54.787, p-value < 2.2e-16
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##      4.426573e-01     -1.817851e-04      6.533432e-05
```

```
#there is spatial correlation
```

All of those results are summarized in a table below. Notice that ML method gives no spatial correlation, but FGLS shows spatial correaltion, although their estimates are similar.

Estimating SEM using Maximum Likelihood method:

```
m.SEM1 <- errorsarlm(pib2000~ -1 + pc2000 + hc2000 + WK + WH,
                     data = mybase@data, listw = W1)
summary(m.SEM1)
```

```
##
## Call:errorsarlm(formula = pib2000 ~ -1 + pc2000 + hc2000 + WK + WH,
##     data = mybase@data, listw = W1)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -1.274274 -0.222341 -0.053924  0.150819  3.564417
##
## Type: error
## Coefficients: (asymptotic standard errors)
##         Estimate Std. Error z value  Pr(>|z|)
## pc2000  0.304144   0.023421  12.986 < 2.2e-16
## hc2000  0.684713   0.025631  26.715 < 2.2e-16
## WK      0.457701   0.029991  15.261 < 2.2e-16
## WH     -0.489073   0.030885 -15.835 < 2.2e-16
##
## Lambda: 0.64423, LR test value: 1761.6, p-value: < 2.22e-16
## Asymptotic standard error: 0.013645
##     z-value: 47.213, p-value: < 2.22e-16
## Wald statistic: 2229, p-value: < 2.22e-16
##
## Log likelihood: -2496.228 for error model
## ML residual variance (sigma squared): 0.13262, (sigma: 0.36416)
## Number of observations: 5502
## Number of parameters estimated: 6
## AIC: 5004.5, (AIC for lm: 6764)
```

The results are that the lagged covariates of human and physical capital stock are significant. The $\lambda$ of the autoregressive error is significant. The magnitude for direct effect of physical and human capital are similar to SDM, see table below. And the sign of indirect effect of human capital is also negative here in SEM.

Also in SEM there is no need to analyze the coefficient value using `impacts()`. The interpretation is as usual.

We can estimate a SEM using FGLS, it is another method and we can verify if the results are robust for changing the method.

Estimating SEM using FGLS:

```
m.SEM1.FGLS <- GMerrorsar(pib2000~ -1 + pc2000 + hc2000 + WK + WH,
                          data = mybase@data, listw = W1)
summary(m.SEM1.FGLS)
```

```
##
## Call:GMerrorsar(formula = pib2000 ~ -1 + pc2000 + hc2000 + WK + WH,
##     data = mybase@data, listw = W1)
```

```
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -1.303135 -0.307439 -0.058646  0.235286  3.392499
##
## Type: GM SAR estimator
## Coefficients: (GM standard errors)
##         Estimate Std. Error z value  Pr(>|z|)
## pc2000  0.311617   0.023760  13.115 < 2.2e-16
## hc2000  0.678834   0.026109  26.000 < 2.2e-16
## WK      0.478421   0.029748  16.082 < 2.2e-16
## WH     -0.507024   0.030937 -16.389 < 2.2e-16
##
## Lambda: 0.57983 (standard error): 0.029841 (z-value): 19.431
## Residual variance (sigma squared): 0.13638, (sigma: 0.3693)
## GM argmin sigma squared: 0.13705
## Number of observations: 5502
## Number of parameters estimated: 6
```

The FGLS estimation confirms the results of ML, the values for all coefficients are close to one and another. Neighbor human capital have a negative effect, and the direct physical capital have lower coefficient than direct human capital.

Let us test for spatial dependence of residuals for both methods of estimation:

```
e.SEM1 <- residuals(m.SEM1) #SEM with ML
moran.test(e.SEM1,W1) #W1 is on queen criterion
```

```
##
##  Moran I test under randomisation
##
## data:  e.SEM1
## weights: W1
##
## Moran I statistic standard deviate = -6.7435, p-value = 1
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##     -5.466931e-02     -1.817851e-04      6.528567e-05
```

```
#there is NO spatial correlation

e.SEM1.FGLS <- residuals(m.SEM1.FGLS) #SEM with FGLS
moran.test(e.SEM1.FGLS,W1) #W1 is on queen criterion
```

```
##
##  Moran I test under randomisation
```

```
##
## data:  e.SEM1.FGLS
## weights: W1
##
## Moran I statistic standard deviate = 52.437, p-value < 2.2e-16
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##      4.236500e-01     -1.817851e-04      6.533043e-05
```

```
#there is spatial correlation
```

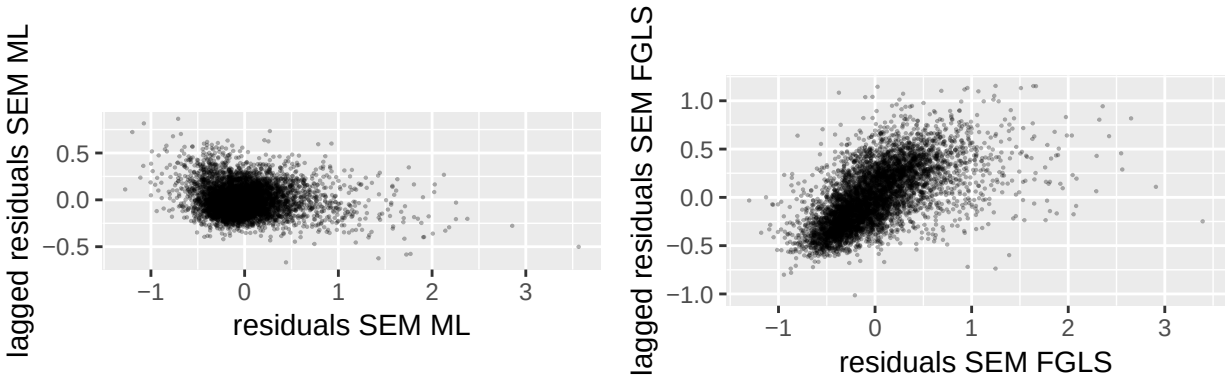
It is strange the fact that with ML, SEM does not have spatial correlation in the residuals. But using FGLS, SEM does have spatial correlation on the residuals.

Let us further analyze using the scatterplot of residuals and lagged residuals:

```
scatter.plot.spatial.lag.residuals.SEM.ML <-
  function(vec1,W){
    vec1lag <- lag.listw(W,vec1)
    h2.df <- as.data.frame(cbind(vec1,vec1lag))
    gg1.gdp.lagged.gdp <- #scatterplot x=vec, y=lagged vec1
      ggplot(data=h2.df, aes(y=vec1lag, x=vec1)) +
      geom_point(size=.1,alpha=.3)+
      coord_equal()+
      xlab("residuals SEM ML") + ylab("lagged residuals SEM ML")
    gg1.gdp.lagged.gdp
  }

scatter.plot.spatial.lag.residuals.SEM.FGLS <-
  function(vec1,W){
    vec1lag <- lag.listw(W,vec1)
    h2.df <- as.data.frame(cbind(vec1,vec1lag))
    gg1.gdp.lagged.gdp <- #scatterplot x=vec, y=lagged vec1
      ggplot(data=h2.df, aes(y=vec1lag, x=vec1)) +
      geom_point(size=.1,alpha=.3)+
      coord_equal()+
      xlab("residuals SEM FGLS") + ylab("lagged residuals SEM FGLS")
    gg1.gdp.lagged.gdp
  }

grid.arrange(scatter.plot.spatial.lag.residuals.SEM.ML(e.SEM1,W1),
             scatter.plot.spatial.lag.residuals.SEM.FGLS(e.SEM1.FGLS,W1),
             nrow = 1)
```




The scatterplots confirm the results in `moran.test()`, surprisingly using ML we already eliminated spatial correaltion in the residuals, although the values for both methods, ML and FGLS, are very close to each other. This shows that even though the estimates of both models are pretty close according to the summary table below, the implication for spatial correlation can be very different.

```
table.OLS.SDM.SEM <- matrix(c( 0.763, 0.333, 0.588, 0.616, 0.304, 0.312,
                               0.198, 0.662, 0.344, 0.320, 0.685, 0.678,
                                NA  , 0.552,  NA  ,  NA  , 0.458, 0.478,
                                NA  ,-0.571,  NA  ,  NA  ,-0.489,-0.507,
                                NA  ,   NA , 0.657, 0.571, 0.644, 0.580,
                               "yes","yes" , "no" , "yes", "no" , "yes" ),
                             nrow = 6,ncol=6,byrow = TRUE)
table.OLS.SDM.SEM <- as.data.frame(table.OLS.SDM.SEM)
rownames(table.OLS.SDM.SEM) <- c("pc2000","hc2000","WK","WH","lambda","sp.corr.")
colnames(table.OLS.SDM.SEM) <- c("OLS(1)","SDM(2)","SEM ML(3)","SEM FGLS(4)","SEM ML(5)"

knitr::kable(
  table.OLS.SDM.SEM, caption = 'results of OLS and SDM and SEM'
)
```

Table 2: results of OLS and SDM and SEM

| | OLS(1) | SDM(2) | SEM ML(3) | SEM FGLS(4) | SEM ML(5) | SEM FGLS(6) |
|---|---|---|---|---|---|---|
| pc2000 | 0.763 | 0.333 | 0.588 | 0.616 | 0.304 | 0.312 |
| hc2000 | 0.198 | 0.662 | 0.344 | 0.32 | 0.685 | 0.678 |
| WK | NA | 0.552 | NA | NA | 0.458 | 0.478 |
| WH | NA | -0.571 | NA | NA | -0.489 | -0.507 |
| lambda | NA | NA | 0.657 | 0.571 | 0.644 | 0.58 |
| sp.corr. | yes | yes | no | yes | no | yes |

The acceptance of this model, SDEM - spatial Durbin error model, is that the underlying economic processes must have spatial interaction of technology through the term of spatial error and that the neighboring capital stock influence its neighboring output results.

Using SEM without lagged covariates it is observed that the ML method, as in the Durbin case, gives the result that there is no spatial correlation. Maybe this is indication that the lagged covariates are not necessary. In the table above always the application of Durbin (lagged covariates) makes the size of `pc2000` shrink and `hc2000` increase, and in all cases `WH` is negative and `WK` positive.

This is a pattern, nevertheless strange, because we thought that neighboring human capital should improve neighboring production. But there are some possible explanations for this phenomenon, we propose two hypotheses: i) the increase in neighboring human capital attracts human capital from neighboring cities, hence decreasing output of neighboring cities; ii) the increase in human capital attracts physical capital from neighboring municipalities. In both cases there is competition between cities, in which one can attract more capital stock through the increase of human capital. Those hypotheses can be tested, using capital stock as explaneid variables in SLM - Durbin models.

In a perfect situation, both ML and FGLS should give us the same results. That is of indication that SEM is the adequate model through the absence of spatial correaltion in the residuals. But this is not the case, hence we should take this conclusion with caution. Next we experiment using the pure spatial autoregressive model, that is without any covariates.

**Pure Spatial Autoregressive - SAR**

This model have the following structure:

$$y = \lambda W y + u$$

Our variable of interest only depends on neighboring values of itself.

The pure autoregression implies a production function of the following format:

$$Y = A(WY)^{\lambda}$$

which structurally speaking is not clarifying because it is hard to believe that the production function does not depend on the factors of production. But in a reduced form is acceptable, in the term $WY$ there is a lot of information about the factors that determine output in the region of interest. Which possibly can substantially explain $Y$.

Testing for pure autoregressive model with data of Brasil:

```
model.pure.autoreg <- spautolm(pib2000~ 1 , data=mybase@data,list=W1)
summary(model.pure.autoreg)
```

```
##
## Call: spautolm(formula = pib2000 ~ 1, data = mybase@data, listw = W1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.19320 -0.81518 -0.22689  0.60059  6.64973
##
## Coefficients:
##              Estimate Std. Error z value  Pr(>|z|)
## (Intercept) 10.162656   0.038175  266.21 < 2.2e-16
##
## Lambda: 0.58888 LR test value: 1563.1 p-value: < 2.22e-16
## Numerical Hessian standard error of lambda: 0.012773
##
## Log likelihood: -8842.728
## ML residual variance (sigma squared): 1.3553, (sigma: 1.1642)
## Number of observations: 5502
## Number of parameters estimated: 3
## AIC: 17691
```

The value for $\lambda$ is positive which agrees with the visual analysis of clusters of income on the map.

```
e.pure.autoreg <- residuals(model.pure.autoreg)

moran.test(e.pure.autoreg,W1) #W1 is on queen criterion
```

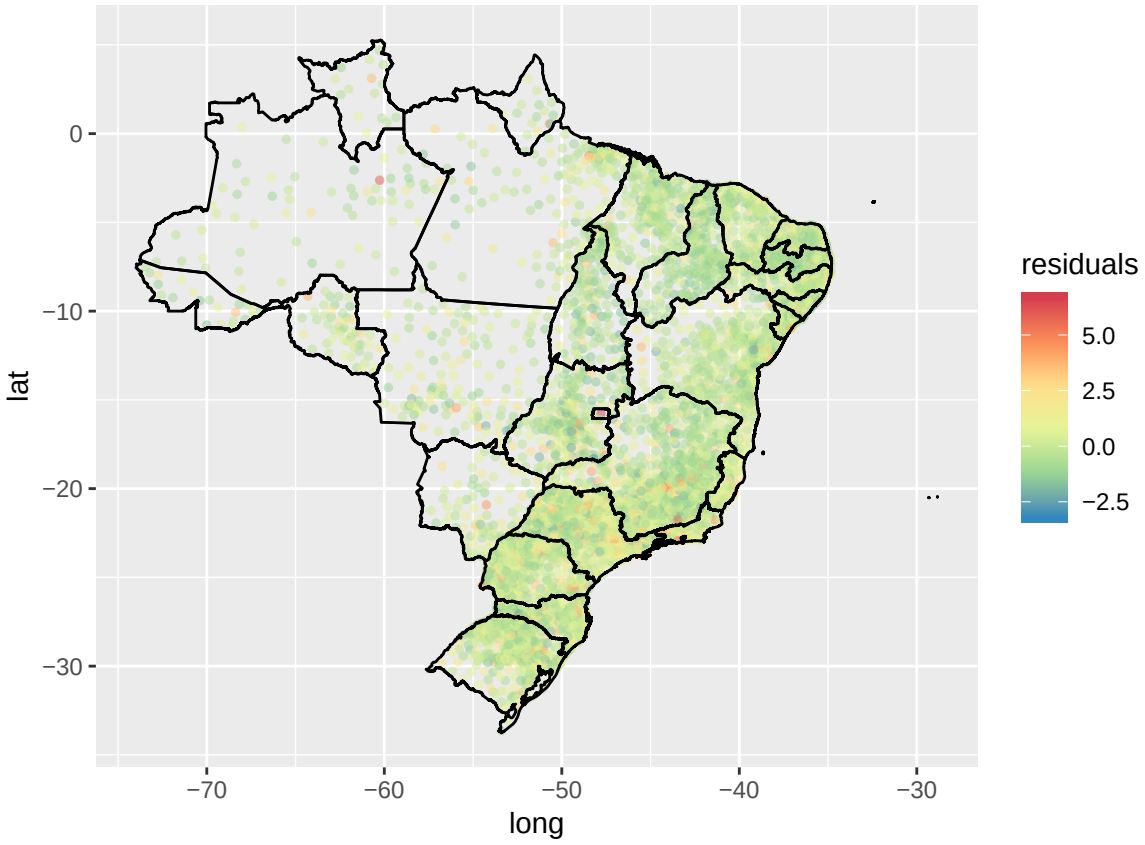
```
##
##  Moran I test under randomisation
##
## data:  e.pure.autoreg
## weights: W1
##
## Moran I statistic standard deviate = -6.2322, p-value = 1
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
```

```
##      -5.056199e-02      -1.817851e-04       6.534924e-05
```

It seems that there is no spatial correlation in the residuals of the pure autoregression model for output.

Make map of residuals, to visualize how there is no spatial correlation.

```
#plot map of residuals of pure spatial model
basetemp$residuals <- e.pure.autoreg  #the residuals of pure autoregression
basetemp.df<- cbind(coordinates(basetemp), basetemp@data)
names(basetemp.df)[c(1,2)] <- c('long','lat')
gg1.res.pure.br <- #plot map residual pure autoregression of Brasil
  ggplot() +
  geom_point(data=basetemp.df, aes(y=lat, x=long, color=residuals),size=1,alpha=0.5)+
  geom_polygon(data=mapa.p,aes(long,lat,group=group), fill = NA, color = "black") +
  coord_equal()+
  scale_colour_distiller(palette = "Spectral")
gg1.res.pure.br
```

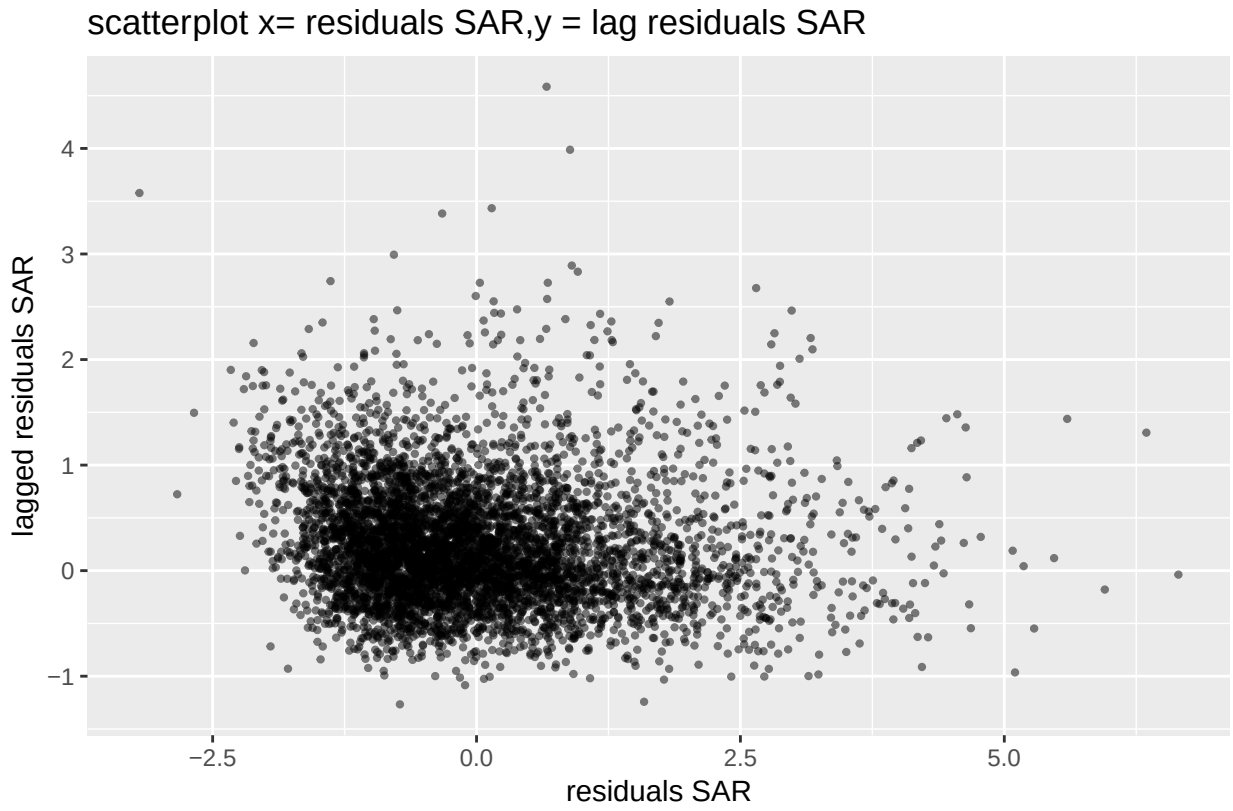



Even with the map it is hard to see any patterns of spatial clusters, we can compare with earlier ones, the predominant characteristic of this maps is greenish tone. Compared with the map of residuals of simple OLS regression e.g., this one have much less concentration of blue points in the regions of Minas Gerais and Nordeste. Now it is clear that bigger cities

like states capitals have a strong positive productivity on them.

Next is presented the scatterplot of residuals vs spatial lag of residuals, another way to see the non-correlation in the residuals:

```
scatter.plot.spatial.lag.residuals.SAR <-
  function(vec1,W){
    vec1lag <- lag.listw(W,vec1)
    h2.df <- as.data.frame(cbind(vec1,vec1lag))
    gg1.gdp.lagged.gdp <- #scatterplot x=vec, y=lagged vec1
      ggplot(data=h2.df, aes(y=vec1lag, x=vec1)) +
      geom_point(size=.8,alpha=0.5)+
      coord_equal()+
      ggtitle("scatterplot x= residuals SAR,y = lag residuals SAR")+
      xlab("residuals SAR") + ylab("lagged residuals SAR")
    gg1.gdp.lagged.gdp
  }
scatter.plot.spatial.lag.residuals.SAR(e.pure.autoreg,W1)
```




The scatterplot shows clear sign that the SAR model eliminates spatial correlation. The pure autoregressive model can capture quite well the process but it is a reduced form equation with lack of economic interpretation.

### Spatial Lag Model (SLM)

The SLM have the following structure:

$$y = \lambda W y + X\beta_1 + WX\beta_2 + u$$

This model have endogeneity problems when applying OLS, hence the literature have two estimation methods in this case: i) Maximum Likelihood (ML), ii) two stage least squares (2SLS) (Arbia 2014).

The usage of SLM implies the following production function:

$$Y = A(WY)^{\lambda} K^{\alpha_1} (WK)^{\alpha_2} H^{\beta_1} (WH)^{\beta_2}$$

which taking log gives us:

$$\ln Y = \lambda WY + \alpha_1 \ln K + \alpha_2 \ln WK + \beta_1 \ln H + \beta_2 \ln WH + \ln A$$

The interpretation here is that output in neighboring regions have an impact on the output of the region of interest. But the drawback of this production function is that it does not give the possibility that a region can have production even without any neighboring regions production. That is, if output of neighbor regions is zero then the model imply that the production in our region is also zero.

The spatial lag model is an extension of SAR model, but in SLM we include covariates.

```
#SLM with ML

#no lagged covariates
m.SLM.ML.1 <- lagsarlm(pib2000~ -1 + pc2000 + hc2000,
                       data = mybase@data, listw = W1)
summary(m.SLM.ML.1)
```

```
##
## Call:
## lagsarlm(formula = pib2000 ~ -1 + pc2000 + hc2000, data = mybase@data,
##     listw = W1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.46754 -0.30930 -0.05762  0.23688  3.41692
##
## Type: lag
## Coefficients: (asymptotic standard errors)
##        Estimate Std. Error z value  Pr(>|z|)
## pc2000 0.772209   0.011049  69.887 < 2.2e-16
## hc2000 0.109603   0.010699  10.244 < 2.2e-16
##
```

```
## Rho: 0.094607, LR test value: 295.72, p-value: < 2.22e-16
## Asymptotic standard error: 0.0056582
##     z-value: 16.72, p-value: < 2.22e-16
## Wald statistic: 279.57, p-value: < 2.22e-16
##
## Log likelihood: -3367.513 for lag model
## ML residual variance (sigma squared): 0.19883, (sigma: 0.4459)
## Number of observations: 5502
## Number of parameters estimated: 4
## AIC: 6743, (AIC for lm: 7036.7)
## LM test for residual autocorrelation
## test value: 1687.6, p-value: < 2.22e-16
```

```
impacts(m.SLM.ML.1, listw = W1)
```

```
## Impact measures (lag, exact):
##           Direct   Indirect    Total
## pc2000 0.7734142 0.07948585 0.852900
## hc2000 0.1097742 0.01128179 0.121056
```

```
#with lagged covariates
m.SLM.ML.2 <- lagsarlm(pib2000~ -1 + pc2000 + hc2000,
                       data = mybase@data, listw = W1, type = "mixed")
summary(m.SLM.ML.2)
```

```
##
## Call:
## lagsarlm(formula = pib2000 ~ -1 + pc2000 + hc2000, data = mybase@data,
##     listw = W1, type = "mixed")
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -1.267516 -0.215174 -0.053688  0.144143  3.606722
##
## Type: mixed
## Coefficients: (asymptotic standard errors)
##             Estimate Std. Error  z value  Pr(>|z|)
## pc2000      0.165383   0.026925   6.1423 8.134e-10
## hc2000      0.833792   0.029740  28.0359 < 2.2e-16
## lag.pc2000  0.148904   0.030757   4.8413 1.290e-06
## lag.hc2000 -0.795481   0.031459 -25.2861 < 2.2e-16
##
## Rho: 0.63913, LR test value: 1808.1, p-value: < 2.22e-16
## Asymptotic standard error: 0.013667
##     z-value: 46.765, p-value: < 2.22e-16
## Wald statistic: 2187, p-value: < 2.22e-16
```

```
##
## Log likelihood: -2447.104 for mixed model
## ML residual variance (sigma squared): 0.13049, (sigma: 0.36124)
## Number of observations: 5502
## Number of parameters estimated: 6
## AIC: 4906.2, (AIC for lm: 6712.3)
## LM test for residual autocorrelation
## test value: 360.55, p-value: < 2.22e-16
```

```
impacts(m.SLM.ML.2, listw = W1)
```

```
## Impact measures (mixed, exact):
##           Direct   Indirect     Total
## pc2000 0.2078377  0.6630838 0.8709215
## hc2000 0.7900075 -0.6838450 0.1061625
```

Note that comparing from SLM ML with and without lagged covariantes there is a significant difference in $\rho$, without lagged covariates $\rho$ is 0.0946, with lagged covariates 0.6391, both of them are significant.

Those tests already show residual spatial correlation shown in " `LM test for residual autocorrelation`", both them show that the SLM ML have spatial dependence.

```
#SLM with 2SLS with lagged covariates
m.SLM.2SLS.nDurbin <- stsls(pib2000~ -1 + pc2000 + hc2000,
                    data = mybase@data, listw = W1)

summary(m.SLM.2SLS.nDurbin)
```

```
##
## Call:stsls(formula = pib2000 ~ -1 + pc2000 + hc2000, data = mybase@data,
##     listw = W1)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -1.410802 -0.316456 -0.062507  0.244518  3.468137
##
## Coefficients:
##         Estimate Std. Error t value  Pr(>|t|)
## Rho    0.0304423  0.0058146  5.2354 1.646e-07
## pc2000 0.7657206  0.0111952 68.3971 < 2.2e-16
## hc2000 0.1694686  0.0108493 15.6202 < 2.2e-16
##
## Residual variance (sigma squared): 0.20427, (sigma: 0.45197)
```

```
impacts(m.SLM.2SLS.nDurbin, listw = W1)
```

```
## Impact measures (lag, exact):
```

```
##           Direct    Indirect     Total
## pc2000 0.7658413 0.023921393 0.7897627
## hc2000 0.1694954 0.005294262 0.1747896
```

```
#SLM with 2SLS with lagged covariates
m.SLM.2SLS <- stsls(pib2000~ -1 + pc2000 + hc2000 + WK + WH,
                    data = mybase@data, listw = W1)
summary(m.SLM.2SLS)
```

```
##
## Call:
## stsls(formula = pib2000 ~ -1 + pc2000 + hc2000 + WK + WH, data = mybase@data,
##     listw = W1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.31225 -0.29339 -0.06339  0.21587  3.38224
##
## Coefficients:
##         Estimate Std. Error t value  Pr(>|t|)
## Rho     0.052014   0.015897   3.272  0.001068
## pc2000  0.312531   0.029760  10.502 < 2.2e-16
## hc2000  0.677941   0.032812  20.661 < 2.2e-16
## WK      0.548779   0.033309  16.475 < 2.2e-16
## WH     -0.608650   0.037901 -16.059 < 2.2e-16
##
## Residual variance (sigma squared): 0.19143, (sigma: 0.43752)
```

```
impacts(m.SLM.2SLS, listw = W1)
```

```
## Impact measures (lag, exact):
##            Direct    Indirect      Total
## pc2000  0.3126757  0.01700288  0.3296786
## hc2000  0.6782559  0.03688264  0.7151385
## WK      0.5490339  0.02985572  0.5788896
## WH     -0.6089324 -0.03311292 -0.6420454
```

Test for spatial correlation of residuals:

```
#SLM - ML - without lagged covariates
e.SLM.ML.1 <- residuals(m.SLM.ML.1)
moran.test(e.SLM.ML.1, W1)
```

```
##
##  Moran I test under randomisation
##
## data:  e.SLM.ML.1
```

```
## weights: W1
##
## Moran I statistic standard deviate = 39.799, p-value < 2.2e-16
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##      3.215113e-01     -1.817851e-04      6.533419e-05
```

```
#there is spatial correlation

#SLM - ML - with lagged covariates
e.SLM.ML.2 <- residuals(m.SLM.ML.2)
moran.test(e.SLM.ML.2, W1)
```

```
##
##  Moran I test under randomisation
##
## data:  e.SLM.ML.2
## weights: W1
##
## Moran I statistic standard deviate = -6.4296, p-value = 1
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##     -5.213161e-02     -1.817851e-04      6.528219e-05
```

```
#there is no spatial correlation

#SLM - 2SLS
e.SLM.2SLS.nDurbin <- residuals(m.SLM.2SLS.nDurbin)
moran.test(e.SLM.2SLS.nDurbin, W1)
```

```
##
##  Moran I test under randomisation
##
## data:  e.SLM.2SLS.nDurbin
## weights: W1
##
## Moran I statistic standard deviate = 47.177, p-value < 2.2e-16
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##      3.811488e-01     -1.817851e-04      6.533335e-05
```

```
#there is spatial correlation
```

```
#SLM - 2SLS
e.SLM.2SLS <- residuals(m.SLM.2SLS)
moran.test(e.SLM.2SLS, W1)
```

```
##
##  Moran I test under randomisation
##
## data:  e.SLM.2SLS
## weights: W1
##
## Moran I statistic standard deviate = 47.335, p-value < 2.2e-16
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##      3.824032e-01     -1.817851e-04      6.532812e-05
```

```
#there is spatial correlation
```

Overall the results agree in the sense that there is spatial autocorrelation, even using SLM, which is not a good sign for the usefulness of this model, given that the objective is to find a model that eliminate the presence of spatial correlation in the residuals. And we found that when using the `moran.test()` in the SLM ML without lagged covariates there is spatial correlation, but SLM ML with lagged covariates there is no spatial correlation (which contradicts the result found in `lagsarlm()` previously). Using SLM 2SLS with lagged covariates there is spatial correlation, according to `moran.test()` which agrees with the results in `lagsarlm()`.

The spatial correlation calculated directly from `lagsarlm()` and from `moran.test()` do not agree when applied to SLM ML with lagged covariates. This is the only aspect that the two do not concur. But it is important to investigate further given that there is still strong evidence that SLM and SEM do not hinder spatial correlation on residuals. Next we investigate how the SARAR model behaves with the data.

```
#1# m.SLM.ML.1
## Impact measures (lag, exact):
##          Direct    Indirect   Total
## pc2000 0.7734142 0.07948585 0.852900
## hc2000 0.1097742 0.01128179 0.121056

#2# m.SLM.ML.2 with lagged covariates
## Impact measures (mixed, exact):
##          Direct    Indirect   Total
## pc2000 0.2078377  0.6630838 0.8709215
## hc2000 0.7900075 -0.6838450 0.1061625

#3# m.SLM.2SLS.nDurbin (no lagged covariates)
```

```
## Impact measures (lag, exact):
##          Direct     Indirect    Total
## pc2000 0.7658413 0.023921393 0.7897627
## hc2000 0.1694954 0.005294262 0.1747896

#4# m.SLM.2SLS
## Impact measures (lag, exact):
##         Direct     Indirect   Total
## pc2000 0.3126757  0.01700288  0.3296786
## hc2000 0.6782559  0.03688264  0.7151385
## WK     0.5490339  0.02985572  0.5788896
## WH    -0.6089324 -0.03311292 -0.6420454
```

Table 1 shows results for ML without lagged covariates, table 2 present results for ML with lagged covariates, table 3 show results for 2SLS without lagged covariates and table 3 shows results for 2SLS with lagged covariates. The inclusion of lagged covariates make the relative size of physical capital shrink relative to human capital, and the neighboring human capital coefficient is negative in all tables.

### SARAR

Spatial autoregressive and error autoregressive model.

According to Arbia (2014) there are three ways to estimate a SARAR model: i) ML, ii) Generalized Spatial 2SLS, iii) Lee's Instrumental Variable estimator (LIV).

SARAR regressions and `impacts()`

```
#SARAR with ML
m.SARAR.ML <- sacsarlm(pib2000~ -1 + pc2000 + hc2000,
                       data = mybase@data, listw = W1)
summary(m.SARAR.ML)
```

```
##
## Call:
## sacsarlm(formula = pib2000 ~ -1 + pc2000 + hc2000, data = mybase@data,
##     listw = W1)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -1.247745 -0.223790 -0.059354  0.147427  3.663676
##
## Type: sac
## Coefficients: (asymptotic standard errors)
##        Estimate Std. Error z value  Pr(>|z|)
## pc2000 0.421655   0.020716  20.354 < 2.2e-16
```

```
## hc2000 0.532953   0.021405  24.898 < 2.2e-16
##
## Rho: -0.058813
## Asymptotic standard error: 0.0059854
##     z-value: -9.8262, p-value: < 2.22e-16
## Lambda: 0.73692
## Asymptotic standard error: 0.011784
##     z-value: 62.536, p-value: < 2.22e-16
##
## LR test value: 1850.9, p-value: < 2.22e-16
##
## Log likelihood: -2589.93 for sac model
## ML residual variance (sigma squared): 0.13227, (sigma: 0.36369)
## Number of observations: 5502
## Number of parameters estimated: 5
## AIC: 5189.9, (AIC for lm: 7036.7)
```

Rho in `sacsarlm()` is associated with the lagged explained variable, and `Lambda` is the coefficient of the lagged error.

```
impacts(m.SARAR.ML,listw = W1)
```

```
## Impact measures (sac, exact):
##           Direct    Indirect     Total
## pc2000 0.4218960 -0.02366282 0.3982332
## hc2000 0.5332581 -0.02990877 0.5033493
```

```
#SARAR with ML Durbin (with lagged covariates)
m.SARAR.ML.D <- sacsarlm(pib2000~ -1 + pc2000 + hc2000 + WK + WH,
                         data = mybase@data, listw = W1)
summary(m.SARAR.ML.D)
```

```
##
## Call:sacsarlm(formula = pib2000 ~ -1 + pc2000 + hc2000 + WK + WH,
##     data = mybase@data, listw = W1)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -1.275708 -0.221699 -0.054135  0.150965  3.558213
##
## Type: sac
## Coefficients: (asymptotic standard errors)
##         Estimate Std. Error z value  Pr(>|z|)
## pc2000  0.304500   0.023454  12.983 < 2.2e-16
## hc2000  0.684343   0.025639  26.691 < 2.2e-16
## WK      0.453335   0.030673  14.780 < 2.2e-16
## WH     -0.478961   0.035647 -13.436 < 2.2e-16
```

```
##
## Rho: -0.0078194
## Asymptotic standard error: 0.015251
##     z-value: -0.5127, p-value: 0.60816
## Lambda: 0.64952
## Asymptotic standard error: 0.016054
##     z-value: 40.46, p-value: < 2.22e-16
##
## LR test value: 1761.8, p-value: < 2.22e-16
##
## Log likelihood: -2496.126 for sac model
## ML residual variance (sigma squared): 0.13237, (sigma: 0.36383)
## Number of observations: 5502
## Number of parameters estimated: 7
## AIC: 5006.3, (AIC for lm: 6764)
```

```
impacts(m.SARAR.ML.D,listw = W1)
```

```
## Impact measures (sac, exact):
##            Direct     Indirect      Total
## pc2000  0.3045035 -0.002365664  0.3021378
## hc2000  0.6843504 -0.005316666  0.6790337
## WK      0.4533398 -0.003521963  0.4498179
## WH     -0.4789656  0.003721047 -0.4752446
```

```
#SARAR with GS2SLS
m.SARAR.GS2SLS <- gstsls(pib2000~ -1 + pc2000 + hc2000,
                         data = mybase@data, listw = W1)
summary(m.SARAR.GS2SLS)
```

```
##
## Call:gstsls(formula = pib2000 ~ -1 + pc2000 + hc2000, data = mybase@data,
##     listw = W1)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -1.316490 -0.242248 -0.060778  0.165354  3.626758
##
## Type: GM SARAR estimator
## Coefficients: (GM standard errors)
##          Estimate Std. Error z value Pr(>|z|)
## Rho_Wy -0.0070057  0.0053338 -1.3135    0.189
## pc2000  0.6107984  0.0162577 37.5698   <2e-16
## hc2000  0.3306992  0.0165944 19.9283   <2e-16
##
## Lambda: 0.55049
```

```
## Residual variance (sigma squared): 0.14457, (sigma: 0.38023)
## GM argmin sigma squared: 0.14767
## Number of observations: 5502
## Number of parameters estimated: 5
```

```
impacts(m.SARAR.GS2SLS,listw = W1)
```

```
## Impact measures (lag, exact):
##           Direct     Indirect     Total
## pc2000 0.6108034 -0.004254356 0.6065491
## hc2000 0.3307019 -0.002303399 0.3283985
```

```
#SARAR with GS2SLS Durbin (with lagged covariates)
m.SARAR.GS2SLS.D <- gstsls(pib2000~ -1 + pc2000 + hc2000 + WK + WH,
                           data = mybase@data, listw = W1)
summary(m.SARAR.GS2SLS.D)
```

```
##
## Call:
## gstsls(formula = pib2000 ~ -1 + pc2000 + hc2000 + WK + WH, data = mybase@data,
##     listw = W1)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -1.259472 -0.230799 -0.054364  0.155765  3.606975
##
## Type: GM SARAR estimator
## Coefficients: (GM standard errors)
##         Estimate Std. Error z value  Pr(>|z|)
## Rho_Wy  0.070442   0.015175   4.642 3.451e-06
## pc2000  0.299601   0.024046  12.460 < 2.2e-16
## hc2000  0.689158   0.026368  26.136 < 2.2e-16
## WK      0.497654   0.029700  16.756 < 2.2e-16
## WH     -0.580623   0.034102 -17.026 < 2.2e-16
##
## Lambda: 0.55214
## Residual variance (sigma squared): 0.13708, (sigma: 0.37024)
## GM argmin sigma squared: 0.13795
## Number of observations: 5502
## Number of parameters estimated: 7
```

```
impacts(m.SARAR.GS2SLS.D,listw = W1)
```

```
## Impact measures (lag, exact):
##            Direct    Indirect      Total
## pc2000  0.2998581  0.02244726  0.3223053
## hc2000  0.6897483  0.05163429  0.7413826
```

```
## WK      0.4980805  0.03728611  0.5353666
## WH     -0.5811204 -0.04350245 -0.6246229
```

Above are four regressions they are the SARAR with ML with and without lagged covariates, and SARAR with GS2SLS with and without lagged covariates. The results are summarized in the following table:

```
#1# SARAR ML
## Impact measures (sac, exact):
##           Direct    Indirect   Total
## pc2000 0.4218960 -0.02366282 0.3982332
## hc2000 0.5332581 -0.02990877 0.5033493
##    Rho  = -0.058   Lambda = 0.767

#2#SARAR ML.D
## Impact measures (sac, exact):
##          Direct     Indirect    Total
## pc2000 0.3045035 -0.002365664  0.3021378
## hc2000 0.6843504 -0.005316666  0.6790337
## WK     0.4533398 -0.003521963  0.4498179
## WH    -0.4789656  0.003721047 -0.4752446
##    Rho*  = -0.007   Lambda = 0.649
## (*) not significant

#3#SARAR.GS2SLS
## Impact measures (lag, exact):
##           Direct     Indirect  Total
## pc2000 0.6108034 -0.004254356 0.6065491
## hc2000 0.3307019 -0.002303399 0.3283985
##    Rho*  = -0.007   Lambda = 0.550
## (*) not significant

#4#SARAR.GS2SLS.D
## Impact measures (lag, exact):
##          Direct     Indirect   Total
## pc2000 0.2998581  0.02244726  0.3223053
## hc2000 0.6897483  0.05163429  0.7413826
## WK     0.4980805  0.03728611  0.5353666
## WH    -0.5811204 -0.04350245 -0.6246229
##    Rho  = -0.070   Lambda = 0.552
```

Table 1 show the results calculated through ML without lagged covariates, table 2 is calculated using ML with lagged covariates, table 3 is calculated using GS2SLS without lagged covariates and table 4 is GS2SLS with lagged covariates. First we can see that the direct impact shown all in tables are very close to the estimated values for the parameters.

As it is seem below all of the specifications above have no signs of spatial correlation in the

residuals. It is interesting to notice that the values for indirec impact for almost all tables is negative, that means that the increase of capital stock in neighboring municipalities decrease the output in the municipality of interest, the values for the coefficients are small though.

Table 1 and 3 have the same covariates (no lagged) but use different methods of estimation. In a perfect situation those results should be close. But the relative magnitude of physical and human capital are reversed, in ML human capital have a higher impact than physical capital, in GS2SLS physical capital is higher. The indirect impact of both are negative but the order of magnitude are different. In GS2SLS in module the indirect impact is smaller.

Table 2 and 4 have the same specification but with different methods of estimation. Interestingly their impact analysis have similar results. In other specification we found that lagged human capital have a negative impact on output, this is a consistent pattern. The estimated `Lambda` for both are also similar, this coefficient capture the spill-over effect of productivity. But `Rho` estimates do not match in size, and their are both negative, this means that higher output in the neighbor cities is detrimental to output of the city. The indirect impact of both tables have reversed signs and different magnitudes, ten fold differences, in GS2SLS is bigger in module.

Moran's I tests:

```
#SARAR with ML
e.SARAR.ML <- residuals(m.SARAR.ML)
moran.test(e.SARAR.ML, W1)
```

```
##
##  Moran I test under randomisation
##
## data:  e.SARAR.ML
## weights: W1
##
## Moran I statistic standard deviate = -8.551, p-value = 1
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##     -6.927135e-02     -1.817851e-04      6.528156e-05
```

```
#SARAR with ML Durbin (with lagged covariates)
e.SARAR.ML.D <- residuals(m.SARAR.ML.D)
moran.test(e.SARAR.ML.D, W1)
```

```
##
##  Moran I test under randomisation
##
## data:  e.SARAR.ML.D
## weights: W1
##
## Moran I statistic standard deviate = -6.8114, p-value = 1
```

```
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##     -5.521748e-02     -1.817851e-04      6.528572e-05
```

```
#SARAR with GS2SLS
e.SARAR.GS2SLS <- residuals(m.SARAR.GS2SLS)
moran.test(e.SARAR.GS2SLS, W1)
```

```
##
##  Moran I test under randomisation
##
## data:  e.SARAR.GS2SLS
## weights: W1
##
## Moran I statistic standard deviate = 3.2184, p-value = 0.0006445
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##      2.582499e-02     -1.817851e-04      6.529706e-05
```

```
#SARAR with GS2SLS Durbin (with lagged covariates)
e.SARAR.GS2SLS.D <- residuals(m.SARAR.GS2SLS.D)
moran.test(e.SARAR.GS2SLS.D, W1)
```

```
##
##  Moran I test under randomisation
##
## data:  e.SARAR.GS2SLS.D
## weights: W1
##
## Moran I statistic standard deviate = -2.5163, p-value = 0.9941
## alternative hypothesis: greater
## sample estimates:
## Moran I statistic       Expectation          Variance
##     -2.051378e-02     -1.817851e-04      6.528802e-05
```

Excluding the SARAR G2SLS without lagged covariates all other specification show no spatial correlation in the residuals.

### A Proposed production function and estimation

We propose a new production function that accounts for spatial influence on economic outcomes:

$$Y = A(K + \beta WK)^{\alpha}(H + \gamma WH)^{1-\alpha}$$

This production function have the neighboring capital stock adding the region's capital stock with a weight coefficient $\beta$ and $\gamma$. This proposed production function allows the possibility that a region be isolated and even then be able to produce some output.

Applying log we find

$$\ln Y = \alpha \ln(K + \beta WK) + (1 - \alpha) \ln(H + \gamma WH) + \ln A$$

Note that this is a non-linear specification.

Hence we propose a `function()` in R that can estimate this kind of non-linearity. The basic idea is: we fixate the values for $\beta$ and $\gamma$ then we can use the usual `lm()` function. We make this evaluation for a discrete range of values, the criterion of fitting is the least sum of squares of residuals.

Let us take the partial derivative,

$$\frac{\partial \ln Y_i}{\partial K_i} = \frac{\alpha}{K_i + \beta(WK)_i} \qquad \frac{\partial \ln Y_i}{\partial (WK)_i} = \frac{\alpha}{K_i + \beta(WK)_i}\beta$$

It is reasonable to expect that the direct effect of capital is greater than the indirect effect, hence we believe that $\beta < 1$. But also that the increase in the neighboring capital stock will not decrease the output of its neighbors, then we expect the $\beta > 0$. Hence we believe that $0 < \beta < 1$. By the same reasoning with human capital we can expect that $0 < \gamma < 1$.

The following R function finds the most adequate fitting value for $\beta$ and $\gamma$:

```
# 181122 function file nonlinear estimator

prod.function.sr1.internal1 <-
  function(data,W,beta,gamma){
    K <- exp(data$pc2000)
    WK<-lag.listw(W,K)
    H <- exp(data$hc2000)
    WH<-lag.listw(W,H)

    physical.c <- log(K + beta*WK)
    human.c <- log(H + gamma*WH)

    if (sum(is.nan(physical.c))==0 & sum(is.nan(human.c))==0) {
      model1 <- lm( pib2000 ~ -1 + physical.c + human.c, data = data )
      e <- resid(model1)
      output <- sum((e-mean(e))^2)
      output
    }
    else {
      output <- NA
      output
```

```
    }

  }


prod.function.sr1<-
  function(data,W){
    accuracy<-.01
    seq.holder1 <- seq(-1,1,by = accuracy)
    length <- length(seq.holder1)
    matrix.holder1 <- matrix(rep(0,3*length^2),ncol=3,nrow=length^2)
    position.all <- 1 #initial position of data entry
    for (beta in seq.holder1) {
      for (gamma in seq.holder1) {

        SQR <- prod.function.sr1.internal1(data,W,beta,gamma)

        matrix.holder1[position.all,1] <- beta
        matrix.holder1[position.all,2] <- gamma
        matrix.holder1[position.all,3] <- SQR

        position.all <- position.all + 1
      }}
    df.holder1 <- as.data.frame(matrix.holder1)
    nm.holder1 <- as.matrix(df.holder1[order(df.holder1$V3),])
    vec.holder1 <- nm.holder1[1,]

    #increasing degree of accuracy
    accuracy<-.00001
    seq.holder1 <- seq(-0.001,0.001,by = accuracy)
    length <- length(seq.holder1)
    matrix.holder1 <- matrix(rep(0,3*length^2),ncol=3,nrow=length^2)
    position.all <- 1 #initial position of data entry
    for (beta in seq(vec.holder1[1]-.001,vec.holder1[1]+.001,by=accuracy)) {
      for (gamma in seq(vec.holder1[2]-.001,vec.holder1[2]+.001,by=accuracy)) {

        SQR <- prod.function.sr1.internal1(data,W,beta,gamma)

        matrix.holder1[position.all,1] <- beta
        matrix.holder1[position.all,2] <- gamma
        matrix.holder1[position.all,3] <- SQR

        position.all <- position.all + 1
      }}
```

```
    df.holder1 <- as.data.frame(matrix.holder1)
    df.holder2 <- df.holder1[order(df.holder1$V3),]
    df.holder3 <- df.holder2[1,]
    names(df.holder3) <- c('beta','gamma','min SSR')
    df.holder3
  }

#apply function
options(warn=-1) #turn warning messages off, there are too many warnings
answer1 <- prod.function.sr1(mybase@data,W1)
options(warn=0) #turn warning messages on
answer1
```

```
##           beta gamma  min SSR
## 12060 -0.00041 0.051 1149.726
```

Note that we found $\beta$ negative but very close to zero, this is sign that the neighboring physical capital does not make much difference in the output of the municipality. $\gamma$ on the other hand is between (0,1) and is positive, 0.051.

```
beta <- as.numeric(answer1[1])
gamma <- as.numeric(answer1[2])
K <- exp(mybase$pc2000)
WK<-lag.listw(W1,K)
H <- exp(mybase$hc2000)
WH<-lag.listw(W1,H)

physical.c <-log(K + beta*WK)
human.c <- log(H + gamma*WH)

model.p1 <- lm( mybase$pib2000 ~ -1 + physical.c + human.c)
summary(model.p1)
```

```
##
## Call:
## lm(formula = mybase$pib2000 ~ -1 + physical.c + human.c)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.4610 -0.3204 -0.0599  0.2494  3.3790
##
## Coefficients:
##            Estimate Std. Error t value Pr(>|t|)
## physical.c 0.771077   0.010385   74.25   <2e-16 ***
## human.c    0.188527   0.008615   21.88   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4572 on 5500 degrees of freedom
## Multiple R-squared:  0.9981, Adjusted R-squared:  0.9981
## F-statistic: 1.464e+06 on 2 and 5500 DF,  p-value: < 2.2e-16
```

Still the values for the direct effect of physical and human capital are very close to what they were before. Just to remember in the simple OLS they were 0.762 and 0.198.

Let us see if there is spatial correlation in the residuals

```
lm.morantest(model.p1, W1) #Moran's I test for residuals
```

```
##
##  Global Moran I for regression residuals
##
## data:
## model: lm(formula = mybase$pib2000 ~ -1 + physical.c + human.c)
## weights: W1
##
## Moran I statistic standard deviate = 49.279, p-value < 2.2e-16
## alternative hypothesis: greater
## sample estimates:
## Observed Moran I      Expectation         Variance
##     3.980403e-01    -2.691219e-04     6.532966e-05
```

```
#there is still strong spatial correlation
```

Apply the same model with spatial errors, in this context productivity and technology have spatial spill over effects.

### Concluding Remarks

In this work we analyze the Cobb-Douglas production function in a spatial econometrics framework using data of Brazilian municipalities of 2000. The results are not conclusive in many aspects. But there are some patterns of behavior in the data that are worth noting. The models tested are: SDM, SLM, SEM, SARAR.

In all those models tests were made with and without lagged covariates, human and physical capital. Usually without lagged covariates physical capital have higher impact on output than human capital. As expected lagged physical capital have a positive impact on output, that means that an increase in the neighbor physical capital stock have a positive impact in the municipality output. But it is a consistent result between models that human capital have a negative impact in neighboring municipalities, that was not expected. Including lagged human capital, the direct human capital impact becomes relatively higher than physical capital, this is consistent between the models.

The models that showed no spatial correlation are: i) SEM using ML (with and without lagged covariates); ii) the pure spatial autoregressive model; iii) SLM ML with caution because `lagsarlm()` LM test shows presence of spatial correlation; iv) SARAR all specifications.

In this work we propose a new equation of production function, this equation accounts for the hypothesis that neighboring capital stock enter additively with the city's capital stock. The econometric drawback of this approach is that the specification becomes non-linear, hence we also propose an algorithm to determine the best values for neighbor weights coefficients using SSR as our criterion of fit. The results are that neighbor physical capital have a slightly negative impact, and that neighboring human capital have a positive impact on output. And the coefficients of physical and human capital are as expected. Albeit the residuals have spatial correlation.

Further investigation: i) how those results change when using different weight matrices; ii) restricting the analysis with different regions or states and see if the coefficients are similar, see if there is shift in relative impact of physical and human capital depending on the inclusion of lagged covariates, see if lagged human capital have negative impact; iii) search for hypothesis that explain the negative impact of neighboring human capital on output iv) make analysis using measures of capital stock and output per capita, and include welfare measures.

## References


Arbia, Giuseppe. 2014. *A Primer for Spatial Econometrics with Applications in R*. Palgrave Macmillan.

Hall, Robert E., and Charles I. Jones. 1999. "Why Do Some Countries Produce so Much More Output Per Worker Than Others?" *The Quarterly Journal of Economics* 114 (1): 83–116.

North, Douglass C., and Robert P. Thomas. 1973. *The Rise of the Western World: A New Economic History*. Cambridge University Press, Cambridge UK.

Pande, Rohini, and Christopher Udry. 2005. "Institutions and Development: A View from Below." *Economic Growth Center Working Paper Yale University* 322 (10): 891–921.