Spatial Economics – Assignment 3

Elia Di Gregorio (h12039313@s.wu.ac.at) Max Heinze (h11742049@s.wu.ac.at)
Sophia Ludescher (h11801892@s.wu.ac.at)

May 3, 2024

Contents

The executable code that was used in compiling the assignment is available on GitHub at https://github.com/maxmheinze/spatial.

Task A

Naïve Panel Model and Specification Search

We begin by fitting the spatially naïve panel model using the `plm` function. The output is printed below.

```r
cm1 <- plm(logc ~ logp + logy, data = cigs, effect = "twoway", model = "within",
    index = c("state", "year"))
```

	Dependent variable:
	logc
logp	-1.035^{***}
	(0.042)
logy	0.529^{***}
	(0.047)
Observations	1,380
R^2	0.394
Adjusted R^2	0.359
F Statistic	424.344^{***} (df = 2; 1303)
Note:	*p<0.1; **p<0.05; ***p<0.01

As the true spatial econometricians we are, we now ignore everything that is reminiscent of economic theory and venture on a standard specification testing path. We begin by performing a Lagrange Multiplier test for spatial lags in both the error and the dependent. The results are printed below.

```r
slmtest(cm1, cigm, test = "lme")
```

```
##
##  LM test for spatial error dependence
##
## data:  formula (within transformation)
## LM = 54.655, df = 1, p-value = 1.437e-13
## alternative hypothesis: spatial error dependence
```

```r
slmtest(cm1, cigm, test = "lml")
```

```
##
##  LM test for spatial lag dependence
##
## data:  formula (within transformation)
## LM = 46.901, df = 1, p-value = 7.468e-12
## alternative hypothesis: spatial lag dependence
```

As we can see, both tests return significant p-values, meaning that we reject the null hypotheses of spatial dependence being present in the errors or the dependent, respectively. We thus continue by performing robust LM tests for spatial dependence in both the errors and the dependent which account for the respectively other type of spatial dependence potentially being present. The results of these two tests are printed below.

```r
slmtest(cm1, cigm, test = "rlme")
```

```
##
##  Locally robust LM test for spatial error dependence sub spatial lag
##
## data:  formula (within transformation)
## LM = 8.9106, df = 1, p-value = 0.002835
## alternative hypothesis: spatial error dependence
```

```r
slmtest(cm1, cigm, test = "rlml")
```

```
##
##  Locally robust LM test for spatial lag dependence sub spatial error
```

```
##
## data:  formula (within transformation)
## LM = 1.1563, df = 1, p-value = 0.2822
## alternative hypothesis: spatial lag dependence
```

Since this time, testing for spatial dependence in the error given spatial dependence in the dependent returns a significant result, and testing for spatial dependence in the dependent given spatial dependence in the errors does not, we conclude that based on these test results, a **SEM model** is the most appropriate specification.

Estimating a SEM Model and Comparing it to an SLX Model

We use the `spml` function from the `splm` package to estimate a spatial panel model using Maximum Likelihood. Consistent with our previous answer, we estimate a **SEM model**, and we do this by setting `lag = FALSE` as well as `spatial.error = "b"`. Written down, the model equation is identical to the one presented in the assignment question, except for the important distinction that

$$\varepsilon = \rho \boldsymbol{W} \varepsilon + \boldsymbol{u}, \quad \boldsymbol{u} \sim \mathrm{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}).$$

```
cm2 <- spml(logc ~ logp + logy, data = cigs, listw = cigm, effect = "twoways", model =
"within",
    index = c("state", "year"), lag = FALSE, spatial.error = "b")
```

The results of this estimation are printed in the following.

	Dependent variable:
	logc
logp	-1.004^{***}
	(0.040)
logy	0.554^{***}
	(0.049)
Spatial error rho	0.240^{***}
	(0.033)
Observations	1,380
Note:	*p<0.1; **p<0.05; ***p<0.01

Next, we estimate an **SLX model** using the `plm` function in combination with the `slag` function to create spatial lags. Since there neither is a lag of the dependent nor are there spatial errors, there is no need to use a function for a dedicatedly spatial panel model. The model we estimate below amounts to

$$\boldsymbol{y}_t = \boldsymbol{X}_t \boldsymbol{\beta} + \boldsymbol{W} \boldsymbol{X}_t \boldsymbol{\gamma} + \boldsymbol{\mu} + \phi_t \boldsymbol{\iota} + \varepsilon_t,$$

where, diverting from the original notation for ease of reading, $\boldsymbol{y}_t$ is a $n \times 1$ vector of stacked $\log(C_{it})$, $\boldsymbol{X}_t$ is a $n \times 2$ matrix consisting of stacked $\log(P_{it})$ and $\log(I_{it})$ as columns, $\boldsymbol{W}$ is the provided weights matrix, $\boldsymbol{\beta} = (\beta_1, \beta_2)'$, $\boldsymbol{\gamma} = (\gamma_1, \gamma_2)'$, and both $\boldsymbol{\mu}$ and ε are stacked over individuals and thus $n \times 1$ vectors.

```
cm3 <- plm(logc ~ logp + logy + slag(logp, listw = cigm) + slag(logy, listw = cigm),
    data = cigs, effect = "twoways", model = "within", index = c("state", "year"))
```

Again, the results are printed below.

	Dependent variable:
	logc
logp	-1.017^{***}
	(0.042)
logy	0.608^{***}
	(0.060)
slag(logp, listw = cigm)	-0.220^{***}
	(0.077)
slag(logy, listw = cigm)	-0.219^{***}
	(0.080)
Observations	1,380
R^2	0.400
Adjusted R^2	0.364
F Statistic	217.105^{***} (df = 4; 1301)
Note:	*p<0.1; **p<0.05; ***p<0.01

We can see that there is positive spatial autocorrelation of errors in the SEM model, and a negative spillover effect of price changes in the SLX model. The **positive spatial autocorrelation of errors** can be interpreted as that states that are contiguous are likely to experience similar shocks. This can be explained as spillover effects of an unobserved variable affecting cigarette demand; however, it is not possible to deduce information about the spillover effect of price changes, one of the explanatory variables in the model, from this error autocorrelation.

In contrast, the SLX model allows for directly examining and interpreting spillover effects, which could be the rationale for choosing it over the SEM (or a SAR) model. The coefficient we receive for **the effect of adjacent states' price changes**, γ_1, is $-$0.220. This can directly be interpreted as that given a one-unit increase in log price of cigarettes, log demand in adjacent states decreases by an average of 0.22 (or, if you want to be imprecise, the price of cigarettes increasing by one percent is associated with demand in adjacent states decreasing by an average of 0.22 percent), ceteris paribus.

Of course, this is really weird, or should at least strike us as such if we hold any belief in our original bootlegging hypothesis of people traveling to other states if cigarettes are cheaper there. This is because the above mentioned result indicates the exact opposite, i.e., if the price of cigarettes in State A increases, *less* people will demand cigarettes in neighboring State B, where the price has *not* increased.

Distance Decay SLX

Since we know that results depend quite strongly on our choice of $\boldsymbol{W}$, it makes sense that Halleck Vega and Elhorst (2015) try a different $\boldsymbol{W}$ and we are asked to follow after them. So, in the following, we estimate an **SLX model with a distance decay specification**, where $w_{ij} = d_{ij}^{-\gamma}$ and $\gamma = 3$.

```
cigd <- read_excel("./assignment3/data/cigarettes/cigar_states.xls") %>%
    select(longitude, latitude) %>%
    as.matrix() %>%
    `rownames<-`(cign)

cigi <- distm(cigd, fun = distVincentyEllipsoid)

cigj <- (cigi/1e+06)^(-3) %>%
    `diag<-`(0) %>%
    mat2listw(style = "B")
```

```
cm4 <- plm(
  logc ~ logp + logy + slag(logp, listw = cigj) + slag(logy, listw = cigj), #SLX w/ distance
decay
  data = cigs,
  effect = "twoways",
```

```
  model = "within",
  index = c("state", "year")
)
```

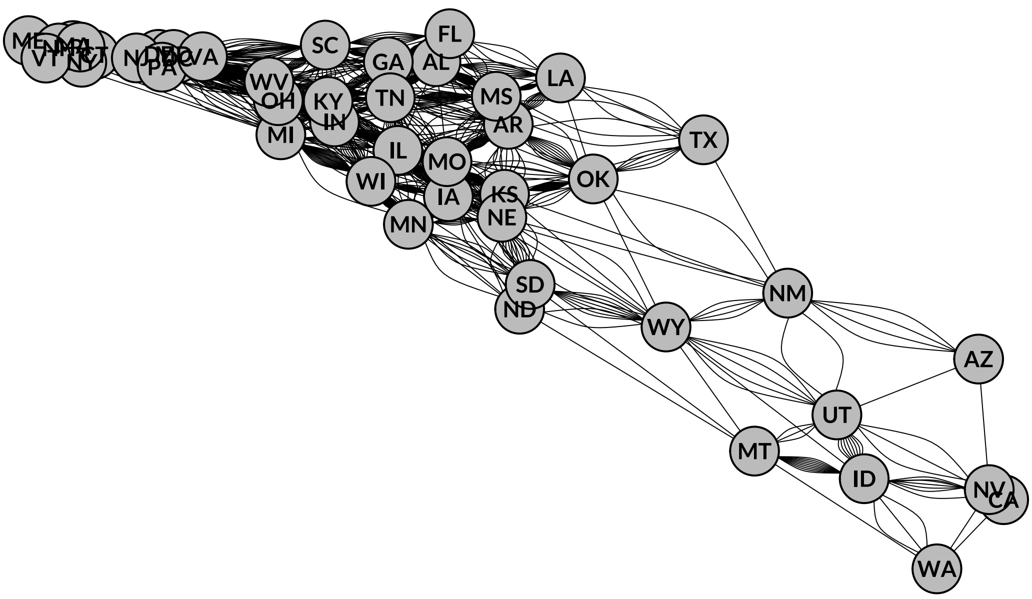
Again, the model results are printed below.

	Dependent variable:
	logc
logp	-0.900^{***}
	(0.038)
logy	0.642^{***}
	(0.042)
slag(logp, listw = cigj)	0.00004^{***}
	(0.00001)
slag(logy, listw = cigj)	-0.0001^{***}
	(0.00001)
Observations	1,380
R^2	0.520
Adjusted R^2	0.491
F Statistic	351.699^{***} (df = 4; 1301)
Note:	*p<0.1; **p<0.05; ***p<0.01

We can see that this time, we actually get the suspected bootlegging effect, as γ_1 is positive. This may indicate that ~~you get your desired results if you just try a sufficiently large number of specifications~~ the distance decay matrix fits the true DGP better.

From a network perspective, this distance decay matrix has interesting implications. Our idea was that if there is a large number of edges, and these edges are weighted by the inverse of a power of the distance, then plotting these network must yield something that vaguely looks like the U.S. And it actually did:

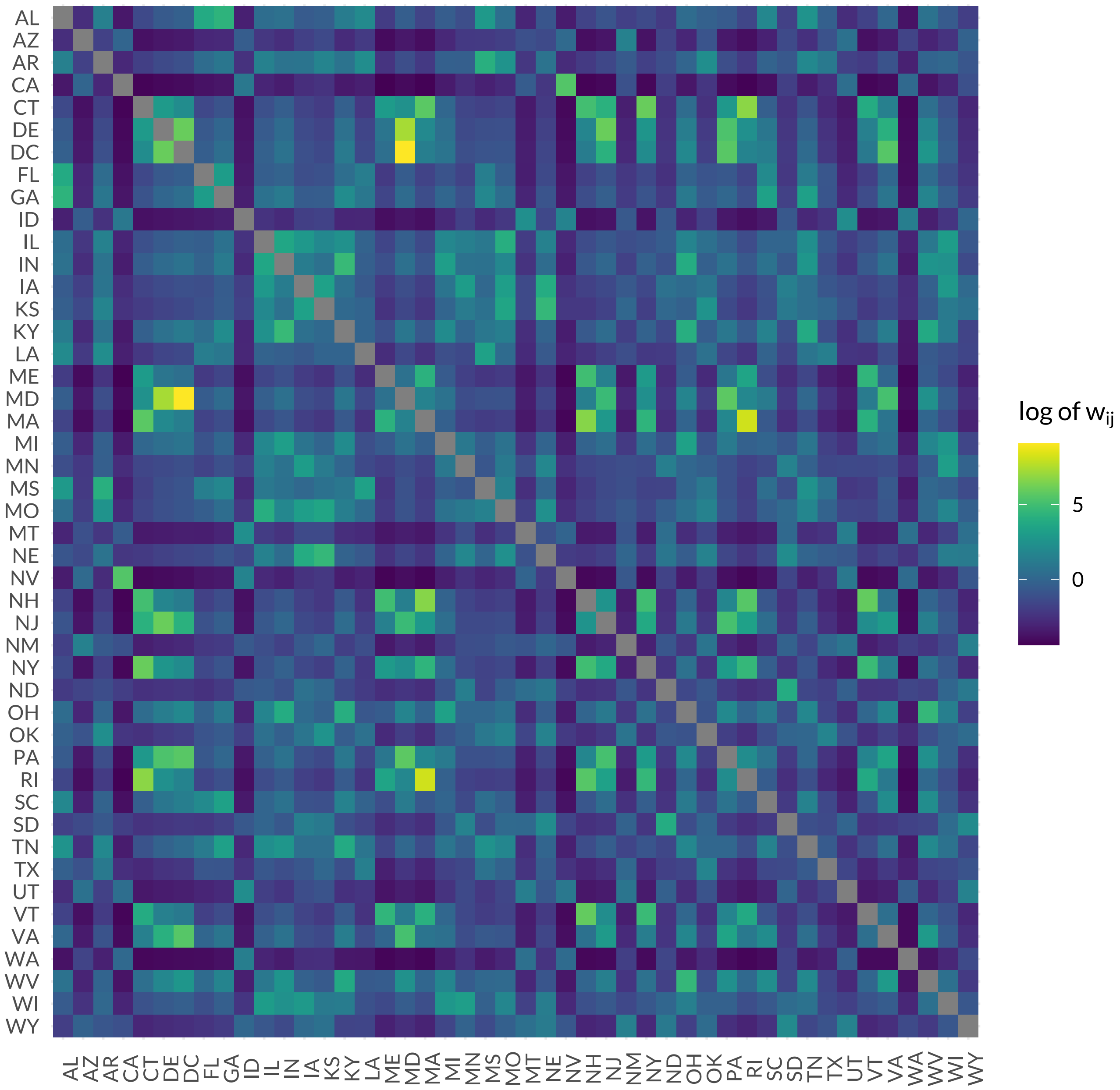
```
plot(cigg, vertex.size = 10, vertex.color = "#BBBBBB", vertex.label.cex = 0.7,
vertex.label.font = 2,
    vertex.label.family = "Lato", vertex.label.color = "black", edge.color = "black",
    edge.width = 0.5, asp = 0)
```



Since the plot is constructed anew every time we compile the document, we do not know how exactly it looks in our final submission. However, it should always resemble the true spatial position of the states recognizably—it may be flipped, it may be mirrored, it may not have north at the top, but the result should resemble the continental United States.

Then we thought about printing the weights matrix, only to find out that this is all but impossible on an A4 sheet of paper. So instead, here is a heatmap of the matrix (with logged values so that the scale appears more clearly):

Spatial Weights Matrix Heatmap



Regarding centrality, we calculated two well-known measures: row sums (since our $\boldsymbol{W}$ is not row-stochastic) and eigenvector centrality. The results are printed below.

```
rowSums(cigk) %>%
    sort(decreasing = TRUE) %>%
    knitr::kable(col.names = "Row Sum")
```

	Row Sum
MD	11026.67198
DC	9945.46534
RI	4756.35048
MA	4728.87454
DE	2663.12139
CT	1913.64955

	Row Sum
NH	1878.45853
PA	1082.00099
NJ	983.52610
NY	968.47870
VT	746.96390
VA	615.20596
ME	386.75814
KY	333.66481
OH	308.21679
IN	303.99804
WV	271.93286
NV	253.46415
CA	248.41830
KS	215.91668
NE	209.03832
IL	207.81238
GA	201.84045
AL	193.97210
MO	191.43001
IA	185.93943
TN	176.79008
MS	160.89378
AR	140.08363
WI	127.73114
MI	123.99357
SC	100.11099
FL	98.10254
SD	90.44541
MN	85.93133
ND	71.63010
LA	70.08663
OK	60.44297
WY	38.15529
ID	35.44483
UT	32.25951
TX	25.43502
MT	25.15952
NM	24.47703
AZ	14.19491
WA	10.75984

	Eigenvector Centrality
MD	1.0000000
DC	0.9871728
DE	0.2112788
PA	0.0682453
VA	0.0486044
NJ	0.0310696
CT	0.0033553
WV	0.0030710
RI	0.0028691
NY	0.0025882
MA	0.0023642
OH	0.0013980
NH	0.0013860
VT	0.0008475
SC	0.0007672

	Eigenvector Centrality
KY	0.0006332
IN	0.0003951
MI	0.0003824
ME	0.0003685
GA	0.0002656
TN	0.0002394
AL	0.0000095
IL	0.0000043
WI	0.0000021
FL	0.0000017
MO	0.0000013
MS	0.0000012
IA	0.0000005
AR	0.0000005
MN	0.0000003
KS	0.0000002
LA	0.0000001
NE	0.0000001
OK	0.0000000
SD	0.0000000
TX	0.0000000
ND	0.0000000
WY	0.0000000
NM	0.0000000
MT	0.0000000
UT	0.0000000
AZ	0.0000000
ID	0.0000000
NV	0.0000000
CA	0.0000000
WA	0.0000000

Using both measures of centrality, Maryland is the most central state, followed by DC as the second-most central entity. The then following positions differ a bit between the two measures of centrality.

Regarding the question where spillover effects occur, let us take another look at the model equation of the SLX model:

$$\boldsymbol{y}_t = \boldsymbol{X}_t\boldsymbol{\beta} + \boldsymbol{W}\boldsymbol{X}_t\boldsymbol{\gamma} + \boldsymbol{\mu} + \phi_t\boldsymbol{\iota} + \boldsymbol{\varepsilon}_t,$$

The **spillover effect** we are interested in is described by γ_1, which in the equation is pre-multiplied by the first column of $\boldsymbol{W}\boldsymbol{X}_t$. In a sense, the first column of $\boldsymbol{W}\boldsymbol{X}_t$ determines where, or how strongly, spillover effects arise. Given a one-unit increase in every state's $\log(P_{it})$, this means that spillover effects of state i on state j will arise where w_{ij} is large. For an illustration, you can thus refer to the heatmap above.

The **average partial effect** is the effect that a (say, one-unit) change in an explanatory variable exerts on the dependent, averaged across all units. In a normal regression setting without spatial lags, this would be represented by the coefficient $\boldsymbol{\beta}$, which is constant over all i. However, in this case, changing the log income vector by one unit for all states i would exert influence over the demand for cigarettes via two channels: the "direct" channel, represented by the parameter β_2, and the "indirect" channel, represented by γ_2 times the i-th row of $\boldsymbol{W}$. The partial effect of a one-unit change in $\log(I_{it})$ for a certain i is thus

$$\mathrm{PE}_i = \beta_2 + \gamma_2 \sum_j w_{ij},$$

and the average partial effect of a one-unit change in $\log(I_{it})$ is then

$$\mathrm{APE} = \beta_2 + \gamma_2 n^{-1} \sum_i \sum_j w_{ij}.$$

Task B

Task B.1 - Unit of Observation

Describe the unit of observation: * subnational "cells" –> grid squares * used to conduct a geographically disaggregated analysis of civil conflict and weather shocks

What are their areas? * cells have dimensions of 110 km by 110 km (square)

Why and to which (relative) extent do they differ? * spatial resolution? * In the Online appendix they also mention " We also compute the area, in squared kilometers, of each cell corresponding to land (i.e. excluding sea or lakes)."

Task B.2 - Un-normalized weights matrix

References

Halleck Vega, S., & Elhorst, J. P. (2015). The slx model. *Journal of Regional Science*, *55*(3), 339–363. https://doi.org/10.1111/jors.12188