Vienna University of Economics and Business

Assignment 2

Spatial Economics

—

Summer 2024

Students:

Elia Di Gregorio
Sophia Ludescher
Hannes Wilkovits

Lecturers/Tutors:

Nikolas Kuschnig
Lukas Vashold
Martin Prinz

Date: 15 April 2024

Table of Contents

1 Exercise A

In exercise A we want to learn more about the spatial dimension of the growth rate of productivity. In the following analysis we focus on 6 European countries on NUTS 2 level - Austria, Germany, Italy, France, Spain and Portugal. The variable of interest is the growth rate of productivity, therefore, we simple calculate the respective rate for each region by utilizing the initial productivity level in the year 1980 and 2013.

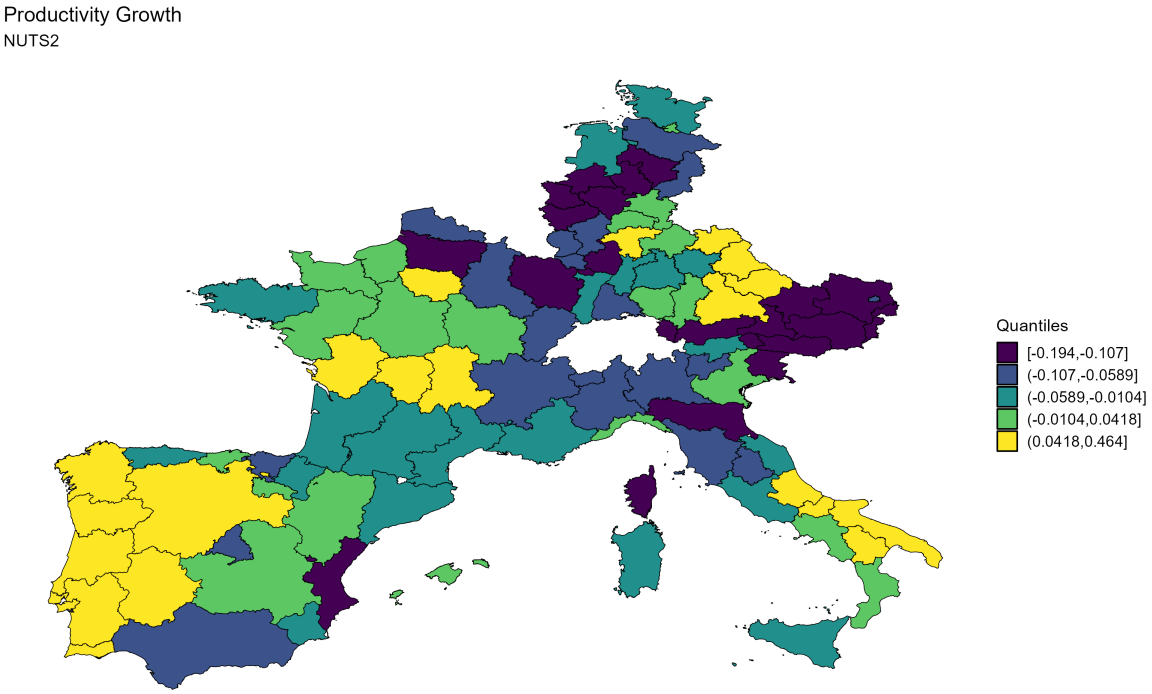

Figure 1: Growth Rate of Productivity from 1980 to 2013

1.1 Spatial Weight Matrices

For our spatial analysis we generate 3 different weight matrices, a queen contiguity measure, a distance threshold and distance decay matrix. In summary, the distance based measures are quiet similar, however, regarding the degree of contiguity-based measure there is some difference. Due to such definition (contiguity refers to observations having a common border), the regions Balearic Islands, Corsica, Sicily and Sardinia have no links connected to them in this setup. In contrast, with our two distance based measures we get at least one neighbor for each region.

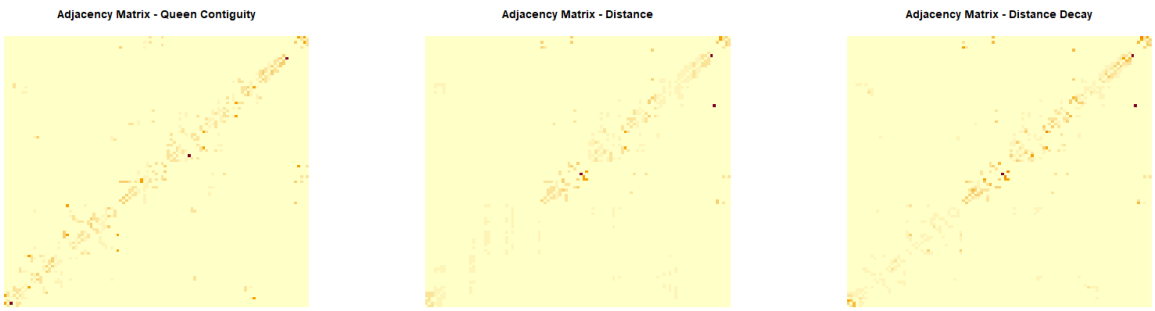

Figure 2: Spatial Weight Matrices

To visualize the differences in the spatial weight matrices, we first plot the image of the respective matrix (2) and try to showcase the network they represent using graph theory (3). One of our trials included a setup where we row-standardized the spatial weight matrices to create proportional weights and account for varying number of neighbors. In the standardized setting slight differences occur within our test-statistics for spatial autocorrelation (presented in upcoming subsections) but still the same levels of magnitude and significance.

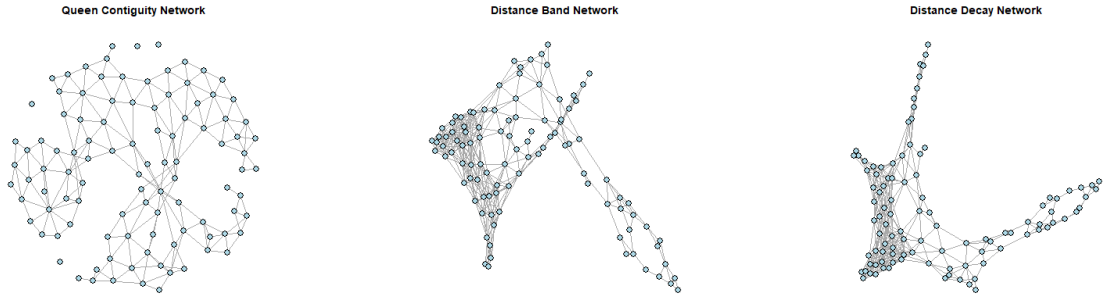

Figure 3: Visualized Networks of respective Adjacency Matrices

1.2 Spatial Autocorrelation

We calculate the Global Moran's I to test for spatial correlation utilizing the 3 different W setups.

Variable	**Moran I Statistic**	**p-value**
Queen Contiguity	0.5379	$<$ 2.2e-16
Distance	0.5748	$<$ 2.2e-16
Distance Decay	0.6124	$<$ 2.2e-16

Table 1: Moran's I Statistics

We have evidence to reject to Null-Hypothesis in all of our 3 setups, thus spatial dependence is present. On average, neighboring observations are somehow similar.

1.3 OLS estimation & Spatially Autocorrelated Errors

Finally, we estimate a linear regression model using OLS. We regress the growth of productivity on the initial productivity in 1980, the (log of) investment and the (log of) employment variable.

	Dependent variable:
	gr
pr80b	-0.253^{***}
	(0.025)
lninv1b	0.032^{***}
	(0.008)
lndens.empb	0.007
	(0.009)
Constant	0.314^{***}
	(0.061)
Observations	103
R^2	0.528
Adjusted R^2	0.514
Residual Std. Error	0.080 (df = 99)
F Statistic	36.983^{***} (df = 3; 99)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 2: Linear Model of Productivity Growth

We check the errors of the OLS model for spatial dependence by once more using the Moran's I test-statistics.

Variable	**Moran I Statistic**	**p-value**
Queen Contiguity	0.3225	2.543e-07
Distance	0.2931	6.257e-09
Distance Decay	0.3298	6.139e-10

Table 3: Spatial Dependence in OLS-Errors

The tests are significant and thus we can reject the H0 of uncorrelated errors for all 3 spatial weight matrix setups, there exists evidence for spatial dependence between the regions. Last but not least, we plot the fitted values and errors to to showcase the differences in comparison to the actual observations (1).

Comparing the 3 figures showcases that OLS estimation seems to systematic over-predicts neighboring regions with higher growth rates and systematic under-predicts neighboring regions with low levels of productivity growth rates.

Spatial dependence often leads to autocorrelation, where nearby observations tend to be more similar to each other than those further apart. This violates the assumption of independence, which

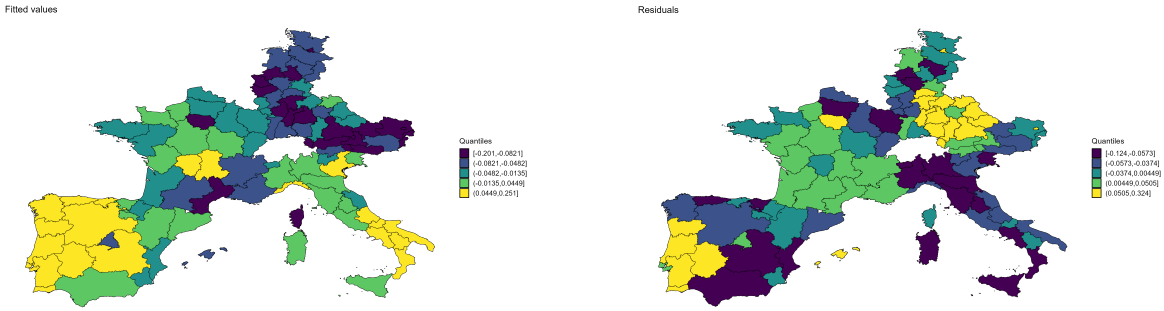

Figure 4: Fitted Values and Errors of OLS estimation

can lead to biased parameter estimates and incorrect inference. Addressing spatial dependence requires appropriate statistical techniques such as spatial regression models.

Disclaimer: the corresponding R-script for this exercise can be found in the Appendix of this document.

2 Exercise B

2.1 Visualizations

We made several representations of the relevant regions under study.
Figure (5) displays the map of the former Guarani Jesuit Mission in present day South America.

- The regions of Itapua and Misiones (PRY) belong to Paraguay,
- The regions Misiones and Corrientes to Argentina,
- The region of Rio Grande do Sur is located in Brazil.

Because of the same name, Guarani Misiones will always be identified with the country specification `(PRY)`.

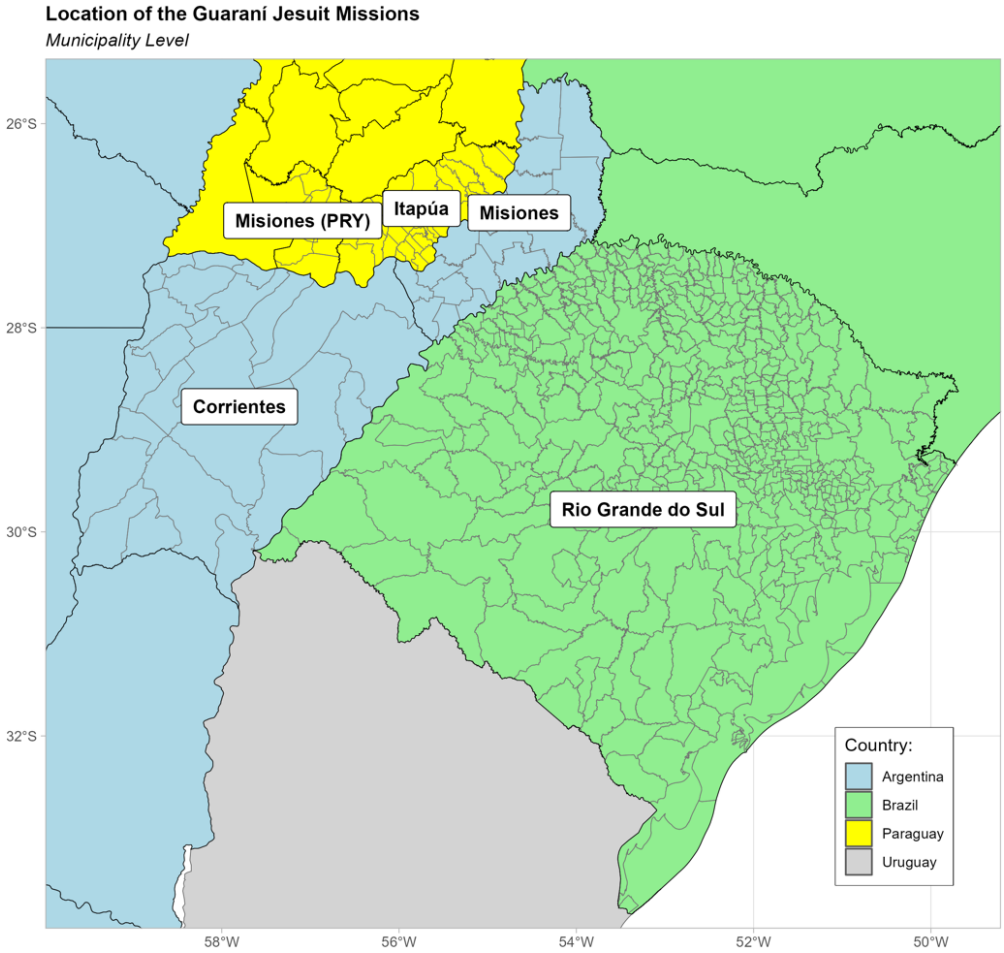

Figure 5: Modern day borders of the Guarani Jesuit Missions

The distribution of current illiteracy rates across municipalities within the former Guarani Jesuit Mission region is depicted in Figure (6). Notably, the overall illiteracy rate in the region does not exceed 25%, indicating a relatively moderate level of illiteracy. However, upon closer inspection, clusters of high illiteracy levels emerge, particularly in the more remote and secluded municipalities situated towards the interior of the region.

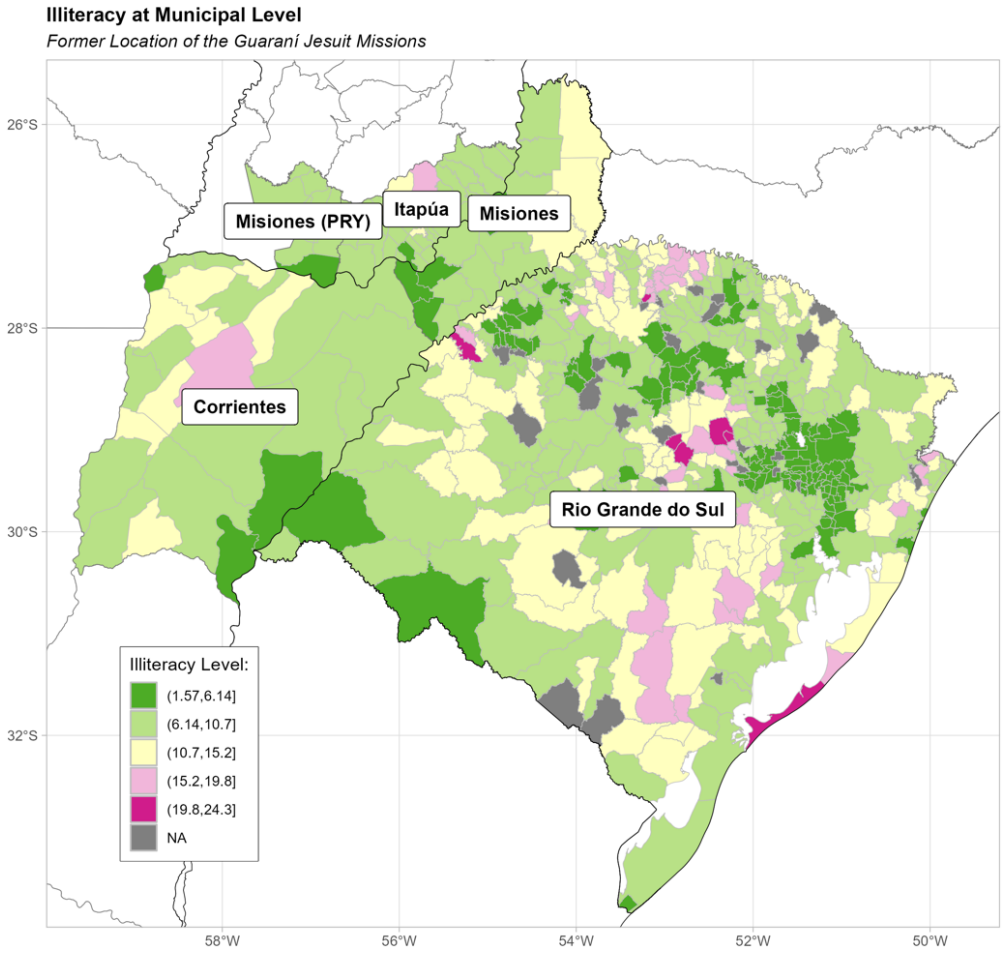

Figure 6: Illiteracy clusters

Figure (7) illustrates the geographical distribution of precipitation levels across the region. Darker shades of color indicate higher levels of rainfall, with notable concentrations observed in the northern part of the region, closer to the Amazonian area. This spatial pattern of precipitation aligns with broader geographical features, as regions nearer to the Amazon are likely to experience higher levels of rainfall due to the influence of the Amazon rainforest on local weather patterns.

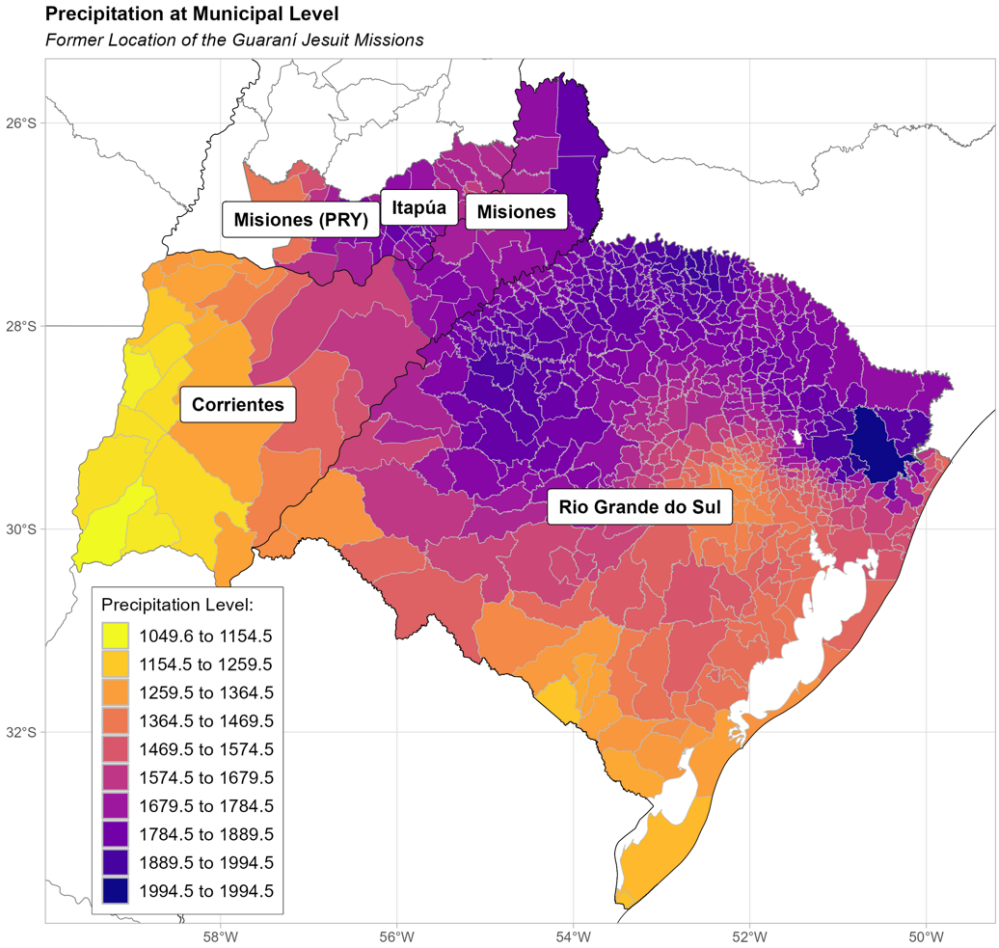

Figure 7: Precipitations levels in the former Guarani Jesuit Missions regions

In Figure (8), we present the distribution of population density across the region, utilizing the `tmap()` function with the `log10_pretty` style specification. This style employs a logarithmic base-10 transformation, wherein each class represents a value ten times larger than the beginning of the previous class. Consequently, each successive class represents the next order of magnitude, facilitating a clearer distinction between low, medium, and high population density areas.
The region with the highest population density is notably concentrated around Porto Alegre, the capital of Rio Grande do Sul state, as indicated by the yellow area in Figure (8). Conversely, population density in the remaining areas appears to be considerably lower, with large swaths of land characterized as sparsely populated. Of particular note is Corrientes in Argentina, which appears to be largely uninhabited according to the population density map. This observation highlights the diverse spatial distribution of population within the region, with urban centers experiencing high population density while rural and remote areas exhibit lower levels of human settlement.

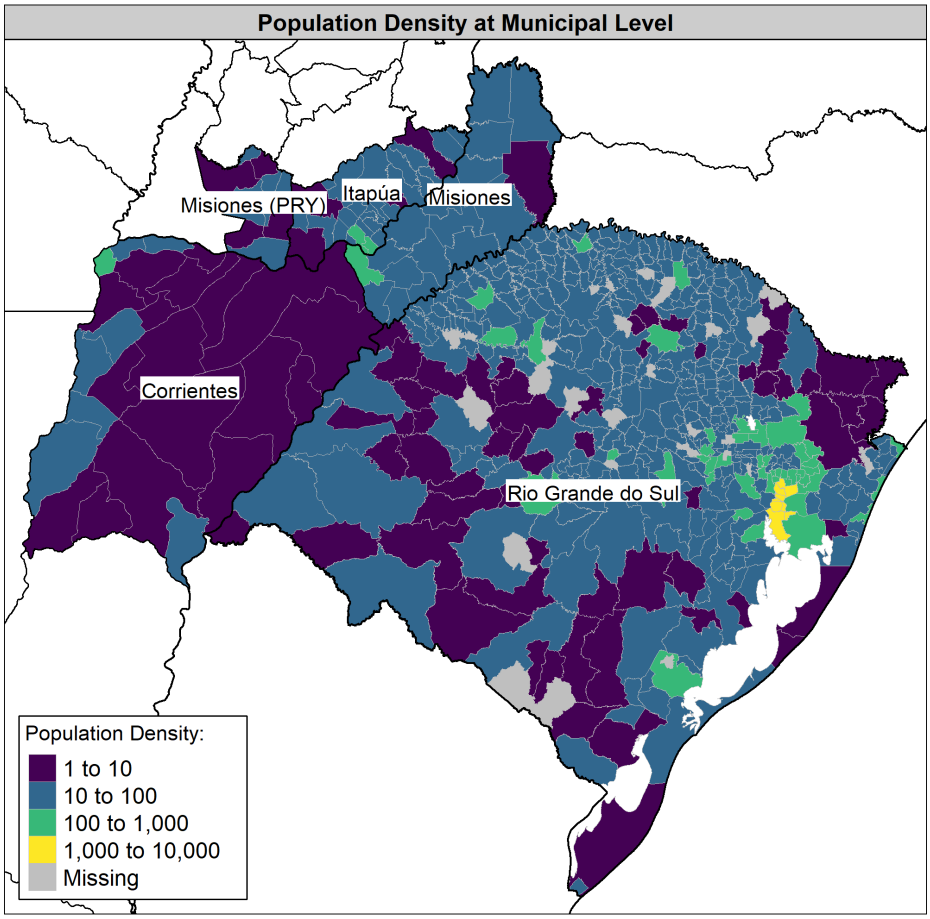

Figure 8: Population density level in the former Guarani Jesuit Missions regions

2.2 Replication Tables

In order to replicate the results reported on Table II of Valencia Caicedo (2018b) we performed an OLS regression on `illiteracy` based on the distance from the Guarani Jesuits Missions and a series of Geographical and State Fixed Effects controls. Notice that although Valencia Caicedo (2018b) says that the latitude and longitude specification are only part of the GEO controls, they are also employed in the baseline estimation as it can be seen from the replication data (Valencia Caicedo, 2018a).

The replication table is reported in the next page in Table (4).

Notice that for clarity and conciseness we collapsed the regressors for distance to the nearest coast, distance to the nearest river, altitude, ruggedness, temperature, area, rainfall, latitude, and longitude, besides the Meso region fixed effects (only for Brazil) under the category `Geo controls`.

Concerning the replication with Conley Standard Errors, we achieved this by employing the function `conleyreg()`. Given that the distance cutoff measure in the paper was set at 0.1 degrees, we had problems in the replication of the exercise given that this formula only allows for distance in km. So we multiplied the degrees of separation of latitude by 110575 to get the corresponding linear distances in meters [1].

Although not perfectly equal, the Conley Standard Errors reported in Table (5) are very similar to the ones obtained by Valencia Caicedo (2018b).

[1] Distance obtained from the National Geospatial-Intelligence Agency Calculator available *here*

	Dependent variable: ***Illiteracy***							
	Baseline Model		Brazil		Argentina		Paraguay	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
distmiss	0.0105***	0.0112**	0.0164**	0.0297***	0.0157*	0.0669***	0.0043	0.0138
	(0.004)	(0.005)	(0.007)	(0.009)	(0.008)	(0.023)	(0.016)	(0.026)
GEO Controls	No	Yes	No	Yes	No	Yes	No	Yes
State Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	548	548	467	467	42	42	39	39
R^2	0.0419	0.0730	0.0940	0.1348	0.1651	0.6689	0.0039	0.2513
Adjusted R^2	0.0294	0.0486	0.0762	0.1041	0.0749	0.5475	−0.0514	0.0190
Residual Std. Error	3.9514 (df = 540)	3.9121 (df = 533)	4.0396 (df = 457)	3.9782 (df = 450)	2.9238 (df = 37)	2.0450 (df = 30)	2.1750 (df = 36)	2.1009 (df = 29)
F Statistic	3.3704*** (df = 7; 540)	2.9977*** (df = 14; 533)	5.2696*** (df = 9; 457)	4.3824*** (df = 16; 450)	1.8297 (df = 4; 37)	5.5089*** (df = 11; 30)	0.0713 (df = 2; 36)	1.0818 (df = 9; 29)
Note:								*p<0.1; **p<0.05; ***p<0.01

Table 4: Replication Baseline Model

	Dependent variable: ***Illiteracy***							
	Baseline Model		Brazil		Argentina		Paraguay	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
distmiss	0.0105***	0.0112**	0.0164*	0.0297***	0.0157**	0.0669***	0.0043	0.0138
	(0.004)	(0.005)	(0.009)	(0.010)	(0.007)	(0.019)	(0.011)	(0.023)
GEO Controls	No	Yes	No	Yes	No	Yes	No	Yes
State Fixed Effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Constant	−35.3484***	−53.7412	−35.2736	−143.8689**	69.2628*	−41.0579	8.6768***	−80.7225**
	(11.9086)	(32.9684)	(38.0588)	(66.4845)	(34.5324)	(46.1689)	(0.6657)	(37.7357)
Note:								*p<0.1; **p<0.05; ***p<0.01

Table 5: Models with Conley SE

2.3 Modelling Distance

When analyzing spatial data, transforming distance using various mathematical functions can significantly impact the interpretation of spatial relationships in regression models. For the assignment we adopted the following transformations:

- logarithm,
- negative exponential,
- square root.

Figure (9) illustrates the variable `distmiss`, which denotes the distance from the closest Guarani Distance Mission, which Valencia Caicedo (2018b) intends as measured from the centroid of the municipality. The shading in the map suggests that the centers of these missions were primarily situated in the North of the Rio Grande do Sur state, near the border with Argentina, thus located in the heart of the relevant regions under investigation, specifically in the most remote and internal areas.

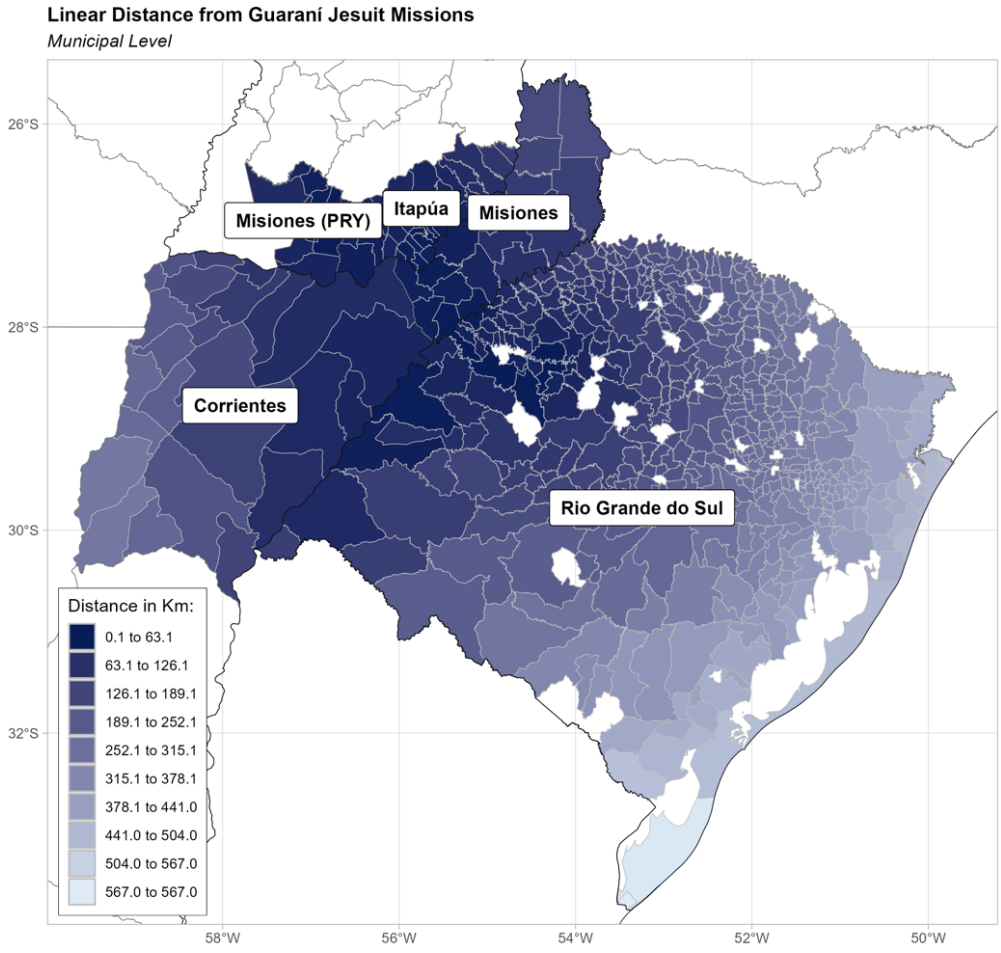

Figure 9: Distance from the former Guarani Jesuit Missions

Logarithm transformation is effective for compressing the scale of distance variables, especially when dealing with large distances. It is particularly useful when the relationship between distance and the outcome variable diminishes at a decreasing rate. Taking the logarithm of distance allows

us to interpret coefficients as the percentage change in distance leading to a one-unit change in the transformed variable.

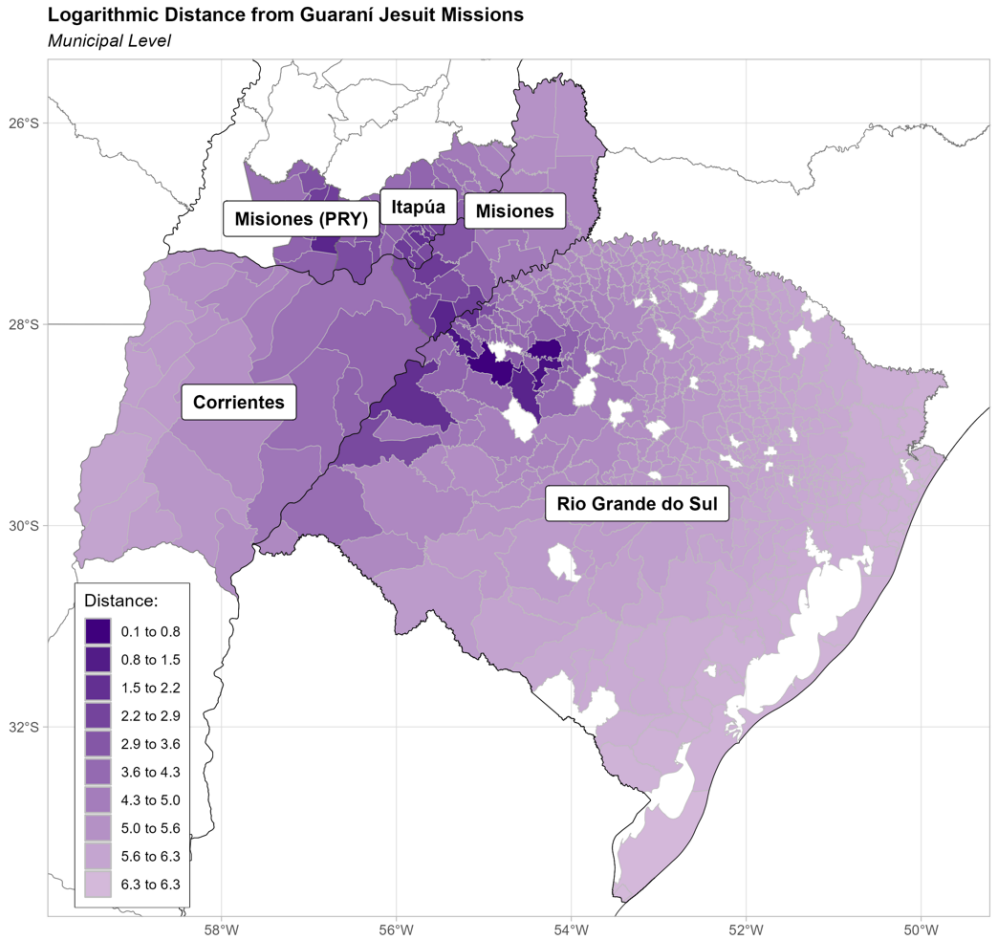

Figure 10: Logarithmic Distance

The negative exponential transformation (reported in Figure (11)) is a valuable tool in spatial analysis, particularly for emphasizing the impact of smaller distances while diminishing the influence of larger distances. This transformation is advantageous when there is a rapid decline in the effect of distance up to a certain threshold.

In the case of the Guarani Jesuit Missions, this transformation is particularly relevant. The positive effects of education and training provided by the missions likely diminish as one resided farther away. This effect is amplified in regions that are secluded and difficult to access, where the challenges of transportation exacerbate the distance barrier.

The coefficients associated with the negative exponential-transformed distance represent the rate of decay in the effect of distance on the outcome variable. In our analysis, a coefficient of -0.01 was employed, signifying that the outcome variable decreases by 1% for each additional unit increase in distance.

Finally concerning square root transformation, depicted in Figure (12), it moderates the influence of extreme distances and may highlight the effect of distances closer to the mean distance. It is advantageous when the relationship between distance and the outcome variable is non-linear but flattens out at higher distances.

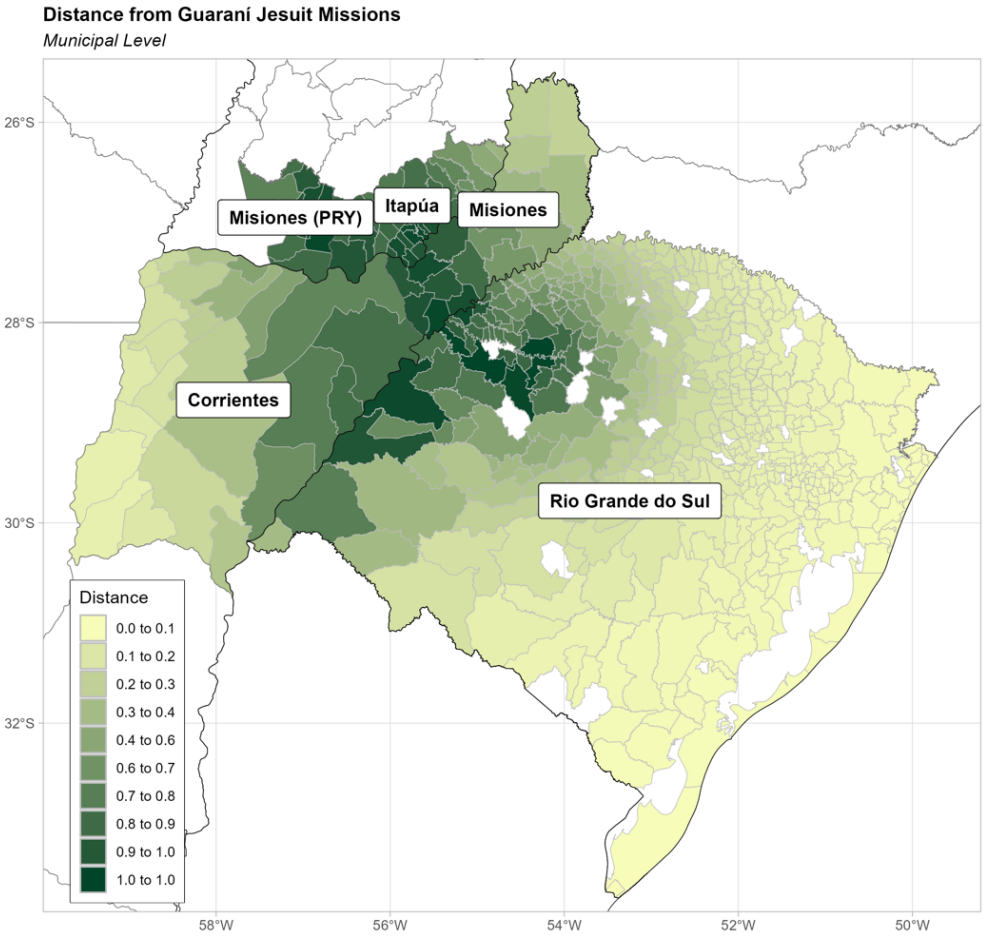

Figure 11: Decay Rate

In conclusion, the choice of transformation for distance variables in regression models significantly impacts the interpretation of spatial relationships. Logarithm, negative exponential, and square root transformations offer ways of understanding relationships between distance and the outcome variable. Researchers must carefully select the appropriate transformation based on the characteristics of the data and the research question at hand to ensure accurate and meaningful interpretations of spatial relationships in regression analysis. This becomes evident in Table (6), where we regressed the baseline model with different transformed distances and can see that not all `distmiss` versions are now statistically significant.

	Dependent variable: ***Illiteracy***			
	Baseline Model	Alternative Distance		
	(1)	(2)	(3)	(4)
distmiss	0.0112**			
	(0.0046)			
distmiss_log		0.1610		
		(0.3326)		
distmiss_dec			−2.1720	
			(1.5114)	
distmiss_root				0.1814*
				(0.1023)
GEO Controls	Yes	Yes	Yes	Yes
State Fixed Effects	Yes	Yes	Yes	Yes
Observations	548	548	548	548
R^2	0.0730	0.0630	0.0662	0.0681
Adjusted R^2	0.0486	0.0384	0.0417	0.0436
Residual Std. Error (df = 533)	3.9121	3.9331	3.9264	3.9224
F Statistic (df = 14; 533)	2.9977***	2.5598***	2.6993***	2.7818***

Note: *p<0.1; **p<0.05; ***p<0.01

Table 6: Modeling different distances

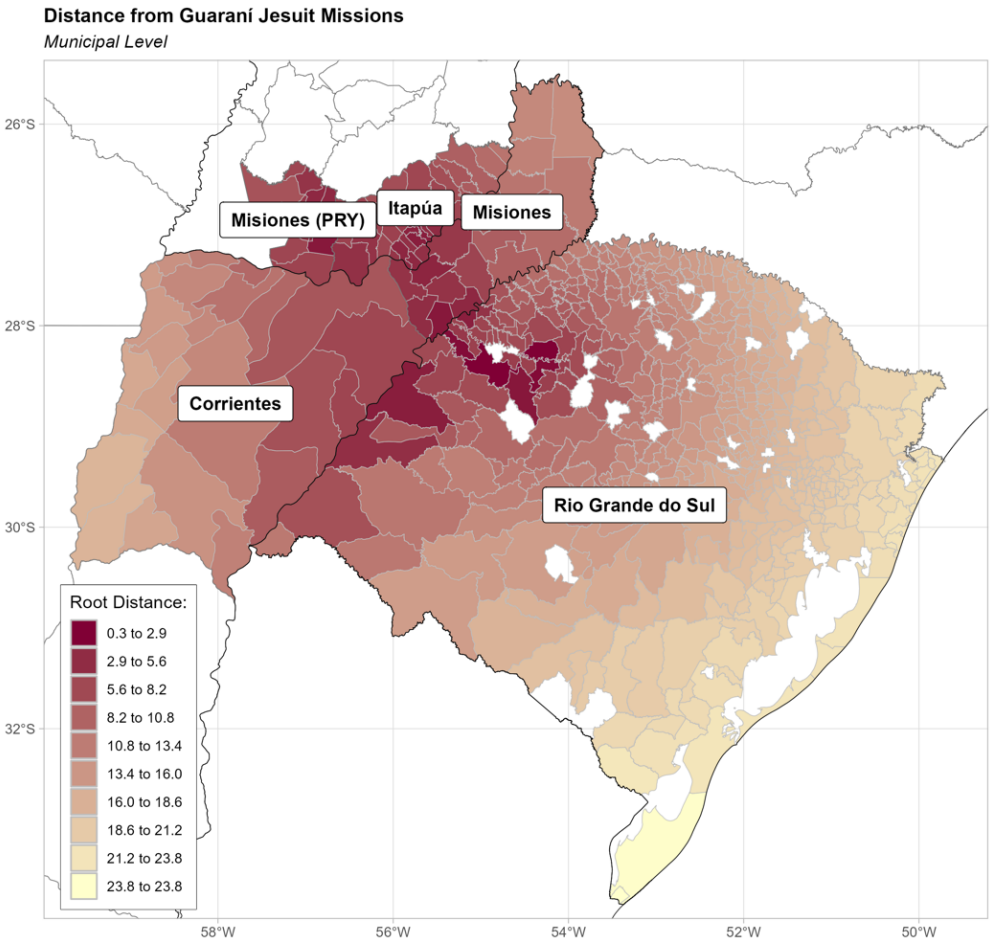

Figure 12: Root Distance

2.4 Accounting for Geography and Spatial Variables

Spatial economics emphasizes the significance of spatial variables in understanding economic dynamics, particularly when analyzing data over extended periods.

Geography plays an important role as a confounding variable in regressions. Consider a scenario where we investigate the correlation between economic growth and education levels across diverse regions over an extended period. The geographical characteristics of these regions, such as their proximity to urban centers, accessibility to transportation networks, or abundance of natural resources, can significantly impact both economic growth and education levels. Failing to incorporate geographical factors such as the natural, local endowment of some regions into the analysis could lead to distorted estimations of the true relationship between economic growth and education, given that over time, the variables of interest build upon that bundle.

Spatial dependence, characterized by the tendency of observations to exhibit similar traits or behaviors of their neighbours, poses a challenge in regressions with a long time gap. Over time, persistent spatial patterns in economic variables can emerge, contributing to spatial autocorrelation. Disregarding spatial dependence may result in biased estimations. For example, if income disparities show spatial autocorrelation between North and South, overlooking this phenomenon could lead to the over- or underestimation of the influence of other variables on income.

Spatial heterogeneity highlights the diversity among geographic regions concerning economic structure, policy environments or cultural influences. Neglecting spatial heterogeneity can lead to misinterpretations for instance, when scrutinizing the determinants of entrepreneurship rates across time. If one overlooks spatial disparities in regulatory frameworks or market accessibility it might flaws the analysis of drivers of entrepreneurship.
Therefore, accounting for geography and spatial variables is fundamental when conducting regressions with a substantial time gap. Failure to acknowledge their influence can introduce confounding effects and biases, compromising the validity of the analysis.

Disclaimer: the corresponding R-script for this exercise can be found in the Appendix of this document.

3 Exercise C

The perils of *(ignoring)* peer effects

"The perils of *(ignoring)* peer effects" addresses, on the one hand, the issues associated with the misinterpretation of results as peer effects due to econometric and methodological challenges when causality is actually unclear as stated by Angrist (2014). On the other hand, ignoring peer effects altogether poses problems as well, potentially reducing the validity and reliability of research findings.

3.1 Drawing valid inference, the trade-off between internal and external validity, and the goals of scientific research

Being able to draw valid causal inference is essential in (social) scientific research to understand complex dynamics of e.g. economic relationships or human behavior. Ignoring crucial effects can thereby undermine the validity of research results. Ignoring peer effects can distort causal relationships, create biased coefficient estimates, and make it impossible to draw valid inference, in turn undermining the understanding of the underlying dynamics. An important problem when analyzing peer effects is endogeneity, as peer effects are often endogenous (since the results of individuals are not independent of the results of their peers), thereby complicating the estimation of causal relationships.

Angrist (2014) also provides insights into another analysis of peer effects, in which peer characteristics should be manipulated independently of individual characteristics in order to isolate the causal effects of social interactions. However, this poses practical limitation problems, leading us to the trade-off between internal and external validity. In a controlled experimental setting, researchers can manipulate and isolate certain variables, like interactions between peers, and control for other factors. However, such an experimental setting, while potentially establishing a clear causal relationship and thereby high internal validity, often lacks the ability to fully capture the complexity of real-world events and thus lacks generalizability and therefore external validity. A real-life scenario, such as a classroom, represents a more generalizable setting and therefore higher external validity, as it better reflects actual social dynamics. However, internal validity may suffer due to confounding and unobservable factors, and establishing a causal relationship will be more difficult. The trade-off becomes apparent. What adds to the problem, not considering peer effects at all can lead to problems in drawing valid causal inference and thus internal validity. Furthermore, ignoring peer effects can pose an issue for external validity, especially in experimental settings, since not considering social dynamics that occur in a real-world scenario makes it difficult/impossible to generalize the results to the wider population.

The goal of applied scientific research is to investigate and understand real-world phenomena (economic developments, social behavior, etc.) and propose practical solutions to problems in the form of e.g. recommendations for policy decisions. Methodological scientific research is more concerned with developing statistical models, methods, and approaches for investigating complex phenomena and testing statistical hypotheses. Ignoring peer effects can have a negative impact on both, as important mechanisms may be overlooked and results may be biased. To summarize and in conjunction with the previous points: The goals of scientific research are everything we have discussed

so far: drawing valid causal inference as well as achieving internal and external validity, all of which can be undermined by ignoring peer effects.

3.2 How network dependence (spatial, social, etc.) may impact validity and relevance of a certain instrument.

In general, *instrumental variable* (IV) estimators fail to identify the parameters of an endogenous regressor if the instruments are not exogenous or relevant. Let us talk more specifically about the *quarter of birth instrument* (QOBI) by Angrist and Krueger (1991). Studies based on US data (see Buckles and Hungerman (2013)) have provided evidence that the QOBI may be endogenous, leading to inconsistent estimates when used as an IV, since women with a high socioeconomic status more often plan births outside the winter months. Conversely, the planning of births for families with a lower socioeconomic status could be based on other criteria or precise planning might not even be possible due to e.g. the lack of contraceptives. Fan et al. (2014), in contrast, who studied Taiwanese data, could not find these differences between women in high and low socioeconomic status, which nicely illustrates that IVs can vary in validity depending on the specific network.

For the QOBI to be valid, the quarter in which a person is born must be independent of other factors that potentially influence educational outcomes (which is already contradicted by the example mentioned about women in the US). The social network a person is born into can vary systematically over the course of the year, as vacation periods, tax incentives, etc. can have different effects on different socio-economic groups. The relevance of the QOBI could also vary by location, as there might be e.g. differences in seasonal employment affecting decisions to have children (spatial network dependence). In areas characterized by agriculture, births could be influenced by the agricultural calendar, which might in turn correlate with e.g. educational opportunities in a region (the world is full of endogeneity). If an IV is used to solve endogeneity problems, this does not work if the instrument itself is endogenous due to (spatial or social) network dependence, reducing both the validity and the relevance of the instrument.

Word count: 796

4 Exercise D

What is the format of the image?

Figure 13: Picture of dog wearing an Elizabethan collar

This very paw-some image has, as already stated in the footnote of the task assignment, a fur-midable raster format. This means that the image is composed of a grid of pixels, each pixel containing information about e.g. color. The image therefore consists of many tiny squares (pixels) positioned in a specific order, with each pixel contributing to the formation of the overall image.

Here is another one:

Figure 14: Another picture of a cute dog [2]

[2]This picture also has a raster format.

References

Angrist, J. (2014). The perils of peer effects. *Labour Economics*, *30*(100), 98–108. https://EconPapers.repec.org/RePEc:eee:labeco:v:30:y:2014:i:c:p:98-108

Angrist, J., & Krueger, A. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, *106*(4), 979–1014. Retrieved April 13, 2024, from http://www.jstor.org/stable/2937954

Buckles, K., & Hungerman, D. (2013). Season of birth and later outcomes: Old questions, new answers. *The Review of Economics and Statistics*, *95*(3), 711–724. Retrieved April 13, 2024, from http://www.jstor.org/stable/43554790

Fan, E., Liu, J.-T., & Chen, Y.-C. (2014, August). *Is the 'Quarter of Birth' Endogenous? Evidence From One Million Siblings in Taiwan* (w20444). National Bureau of Economic Research. Cambridge, MA. https://doi.org/10.3386/w20444

Valencia Caicedo, F. (2018a). Data.zip. In *Replication data for: The mission - human capital transmission, economic persistence, and culture in south america*. Harvard Dataverse. https://doi.org/10.7910/DVN/ML1155/H6UPL5

Valencia Caicedo, F. (2018b). The Mission: Human Capital Transmission, Economic Persistence, and Culture in South America*. *The Quarterly Journal of Economics*, *134*(1), 507–556. https://doi.org/10.1093/qje/qjy024

Appendix

Appendix A: Exercise A

For a more thorough analysis of the code for Exercises A and B please consult the public GitHub repository available *here*.

Productivity Growth Visualization:

```
#read geo data and check
shp <- st_read(dsn = "./data", layer ="EU27") #reads in the shapefile
head(shp)
st_crs(shp)

#countries of interest
shp_6 <- shp %>%
  filter(grepl("^AT|^DE|^IT|^FR|^ES|^PT", Id))

#load productivity data and check
load(file="./data/data1.rda")
glimpse(data1)

selected_vars <- c("IDb", "pr80b","pr103b", "lninv1b", "lndens.empb")
data2 <- data1[, selected_vars]

#merge data and check
shp_6merged <- left_join(shp_6, data2, by=c("Id"="IDb"))
str(shp_6merged)
```

```
#growth rate of productivity from 1980 to 2013
shp_6merged$gr <- (shp_6merged$pr103b-shp_6merged$pr80b)/shp_6merged$pr80b
summary(shp_6merged$gr)
shp_6mergedc <- na.exclude(shp_6merged)
shp_6mergedc <- shp_6mergedc %>%
  mutate(gr.quant= cut(gr,
                       breaks = c(quantile(shp_6mergedc$gr, probs = seq(0,1, by=0.2))),
                       include.lowest = TRUE))
summary(shp_6mergedc$gr.quant)

mapA1 <- ggplot(data = shp_6mergedc) +
  geom_sf(aes(fill = gr.quant), colour = "black") +
  scale_fill_viridis_d(option = "viridis", direction = 1) +
  theme_void() +
  theme(
    plot.margin = margin(2, 2, 2, 2, "cm"),  # Adjust the plot margins
    legend.position = "right",
    legend.key.size = unit(0.5, "cm"),  # Adjust the size of the legend key
    legend.text = element_text(size = 8)  # Adjust the size of legend text
  ) +
  labs(title = "Productivity Growth",
       subtitle = "NUTS2",
       fill = "Quantiles")
print(mapA1)
```

Spatial Weight Matrices:

```
#centroids
coords <- st_coordinates(st_centroid(shp_6mergedc))

##Queen contiguity
queen_nb <- poly2nb(shp_6mergedc, row.names=shp_6mergedc$Id, queen=TRUE)
W.list.queen <- nb2listw(queen_nb, style = "W", zero.policy=TRUE) #row-standardized with style=B
W.queen <- listw2mat(W.list.queen)
##Distance band
distw <- dnearneigh(coords, 0, 100, row.names=shp_6mergedc$Id)
summary(distw)
k1 <- knearneigh(coords, k=1)
k1 <- knn2nb(k1)
link.max <- max(unlist(nbdists(k1, coords=coords)))
distw <- dnearneigh(coords, 0, link.max, row.names=shp_6mergedc$Id)
W.list <- nb2listw(distw, style="W", zero.policy=FALSE)
W.dist <- listw2mat(W.list)
#Distance decay
dnbdists <- nbdists(distw, coords)
gl <- lapply(dnbdists, function(x) {
  1/x
  })
W.list.gl <- nb2listw(distw, glist=gl, zero.policy=FALSE)
W.distdec <- listw2mat(W.list.gl)

#Comparing weights matrices: W.queen, W.list, W.list.gl
degree_queen <- rowSums(W.queen)
degree_dist <- rowSums(W.dist)
degree_distdec <- rowSums(W.distdec)
degree0_ids <- names(degree_queen)[degree_queen == 0]
names_for_degree0 <- shp_6mergedc$Name[shp_6mergedc$Id %in% degree0_ids]
print(names_for_degree0)

#Plot weight matrics
image(W.queen, main = "Adjacency Matrix - Queen Contiguity", axes = FALSE)
image(W.dist, main = "Adjacency Matrix - Distance",axes = FALSE)
image(W.distdec, main = "Adjacency Matrix - Distance Decay", axes = FALSE)

#Visualizing the network
##function to plot a network
plot_network <- function(W, title) {
  G <- graph.adjacency(as.matrix(W), mode = "undirected", weighted = TRUE)
  plot(G, main = title, vertex.label = NA, vertex.size = 5, vertex.color = "lightblue")
}
# Plot Queen Contiguity network
plot_network(W.queen, "Queen Contiguity Network")
# Plot Distance Band network
plot_network(W.dist, "Distance Band Network")
# Plot Distance Decay network
plot_network(W.distdec, "Distance Decay Network")
```

Spatial Autocorrelation:

```
##Evidence for spatial dependence being present: Global Moran's I
###Queen contiguity
moran.test(shp_6mergedc$gr, listw = W.list.queen, alternative = "greater")
###Distance band
moran.test(shp_6mergedc$gr, listw = W.list, alternative = "greater")
###Distance decay
moran.test(shp_6mergedc$gr, listw = W.list.gl, alternative = "greater")

#OLS estimation & spatial autocorrelation
ols_fit <- lm(gr ~ pr80b + lninv1b + lndens.empb, data = shp_6mergedc)
summary(ols_fit)

#check the errors of the OLS model for spatial dependence
lm.morantest(ols_fit, W.list.queen)
lm.morantest(ols_fit, W.list)
lm.morantest(ols_fit, W.list.gl)

#test is significant and thus we reject the H0 of uncorrelated errors
#plot the errors and the fitted values:
shp_6mergedc$u_ols <- ols_fit$residuals
shp_6mergedc$y_hat <- ols_fit$fitted.values

shp_6mergedc <- shp_6mergedc %>%
  mutate(u_ols.quant = cut(u_ols,
                           breaks = c(quantile(shp_6mergedc$u_ols, probs = seq(0,1, by=0.2))),
                           include.lowest = TRUE),
         y_hat.quant = cut(y_hat,
                           breaks = c(quantile(shp_6mergedc$y_hat, probs = seq(0,1, by=0.2))),
                           include.lowest = TRUE))

mapA2 <- ggplot(data = shp_6mergedc) +
  geom_sf(aes(fill = y_hat.quant), colour = "black") +
  scale_fill_viridis_d(option = "viridis") +
  theme_void() +
  theme(
    plot.margin = margin(2, 2, 2, 2, "cm"),  # Adjust the plot margins
    legend.position = "right",  # Move legend to the bottom
    legend.key.size = unit(0.5, "cm"),  # Adjust the size of the legend key
    legend.text = element_text(size = 8)  # Adjust the size of legend text
  ) +
  labs(title = "Fitted values", fill = "quantiles")
print(mapA2)

mapA3 <- ggplot(data = shp_6mergedc) +
  geom_sf(aes(fill = u_ols.quant), colour = "black") +
  scale_fill_viridis_d(option = "viridis") +
  theme_void() +
  theme(
    plot.margin = margin(2, 2, 2, 2, "cm"),  # Adjust the plot margins
    legend.position = "right",  # Move legend to the bottom
    legend.key.size = unit(0.5, "cm"),  # Adjust the size of the legend key
    legend.text = element_text(size = 8)  # Adjust the size of legend text
  ) +
  labs(title = "Residuals", fill = "quantiles")
print(mapA3)
```

Appendix B: Exercise B

```
# LOADING DATA ----------------------------------------------------------------
literacy <- read_dta("./01_Data-input/literacy_Arg-Bra-Par.dta", encoding = "latin1")

# National
shp_arg_0 <- st_read("./01_Data-input/ARG/gadm41_ARG_0.shp")
shp_ury_0 <- st_read("./01_Data-input/URY/gadm41_URY_0.shp")
shp_bra_0 <- st_read("./01_Data-input/BRA/gadm41_BRA_0.shp")
shp_pry_0 <- st_read("./01_Data-input/PRY/gadm41_PRY_0.shp")

# Regional
shp_bra_1 <- st_read("./01_Data-input/BRA/gadm41_BRA_1.shp")
shp_pry_1 <- st_read("./01_Data-input/PRY/gadm41_PRY_1.shp")
shp_arg_1 <- st_read("./01_Data-input/ARG/gadm41_ARG_1.shp")

#Provincial
shp_arg_2 <- st_read("./01_Data-input/ARG/gadm41_ARG_2.shp")
shp_bra_2 <- st_read("./01_Data-input/BRA/Brazil_province.shp")
shp_pry_2 <- st_read("./01_Data-input/PRY/gadm41_PRY_2.shp")

# PREPPING --------------------------------------------------------------------
Country <- rbind(shp_arg_0, shp_bra_0, shp_pry_0, shp_ury_0) %>%
  transform_NA(., "geometry") # user-defined function to remove "NA" in the dataset,
  it ignores the specified column $geometry

Region <- rbind(shp_arg_1, shp_bra_1, shp_pry_1) %>%
  transform_NA(., "geometry") %>%
  mutate(NAME_1 = if_else(NAME_1 == "Misiones" & GID_0 == "PRY", "Misiones (PRY)", NAME_1))

Municipality <- rbind(shp_arg_2, shp_bra_2, shp_pry_2) %>%
  transform_NA(., "geometry") %>%
  mutate(NAME_1 = if_else(NAME_1 == "Misiones" & GID_0 == "PRY", "Misiones (PRY)", NAME_1))
```

Before merging the `literacy` dataset with the Municipal shapefile, several adjustments were made to ensure compatibility and coherence between the datasets:

- Standardization of names across datasets and shapefiles: This involves ensuring that the names used to identify geographical entities (such as municipalities or regions) are consistent across both dataframes. Inconsistencies in naming conventions can lead to errors or mismatches during the merging process.
- Renaming of columns: This adjustment entails renaming columns within the datasets to align with each other or to make them more descriptive and intuitive. Renaming columns can improve clarity and facilitate understanding during data analysis.
- Introducing levels of `mesorregi`

```
ds <- literacy %>%
  mutate(state = ifelse(state == "Misiones", "Misiones (PRY)", state),
         state = ifelse(state == "Misiones1", "Misiones", state),
         muni = case_when(
           muni == "Cambyretá" ~ "Cambyreta",
           muni == "Leandro Oviedo" ~ "José Leandro Oviedo",
           muni == "Mayor Otaño" ~ "Mayor Julio D. Otaño",
           muni == "San Juan Bautista" ~ "San Juan Bautista de las Misione",
           muni == "25 de Mayo" ~ "Veinticinco de Mayo",
           muni == "Restinga Seca" ~ "Restinga Sêca",
           muni == "Sant' Ana do Livramento" ~ "Sant'Ana do Livramento",
           muni == "Vespasiano Correa" ~ "Vespasiano Corrêa",
           muni == "Westfalia" ~ "Westfália",
           TRUE ~ muni),
         country = ifelse(country == "BRA","Brazil", country),
         state = ifelse(state == "RS","Rio Grande do Sul", state),
         mesorregi = as.factor(mesorregi)) %>%
  rename(COUNTRY = country,
         NAME_1 = state,
         NAME_2 = muni,
         mis_pry = mis,
         mis = mis1)

Jesuit <- c("Misiones", "Misiones (PRY)", "Corrientes", "Rio Grande do Sul", "Itapúa")
#Used to specify relevant regions
df <- merge(Municipality[Municipality$NAME_1 %in% Jesuit, ], ds,
    by = c("COUNTRY", "NAME_1", "NAME_2"), y.all = TRUE)
```

Now that we have obtained our final dataset we can visualize the variables of interest. Provided is an example for `ggplot` and `tmap`. In fact, for the former, the same structure was used for the different visualizations, changing just the relevant variable plotted.

```
#Example ggplot:
plot_02 <- ggplot() +
  geom_sf(data = df, aes(geometry = geometry, fill = illiteracy_bins), color = "gray") +
  geom_sf(data = Region, aes(geometry = geometry), color = "#737373", fill = NA) +
  geom_sf(data = Country, aes(geometry = geometry), color = "black", fill = NA) +
  scale_fill_manual(values = custom_palette, name = "Illiteracy Level:") +
  theme_light() +
  scale_size_identity() +
  labs(x = NULL, y = NULL,
       title = "Illiteracy at Municipal Level",
       subtitle = "Former Location of the Guaraní Jesuit Missions") +
  theme(plot.title = element_text(size = 12, face = "bold"),
        plot.subtitle = element_text(size = 11, face = "italic"),
        legend.position = "inside",
        legend.position.inside = c(0.15, 0.20),
        legend.box.just = "left",
        legend.box.background = element_rect(color = "black", linetype = "solid")) +
  geom_label(data = Region[Region$NAME_1 %in% Jesuit, ], aes(x = X, y = Y, label = NAME_1),
             size = 4, color = "black",fill = "white", fontface = "bold",
             label.padding = unit(0.5, "lines")) +
  guides(fill = "legend") +
  coord_sf(xlim = c(-59.5, -49.7), ylim = c(-33.5, -25.75))
```

```
#Example tmap:
plot_04 <- tm_shape(df) +
  tm_fill("popd", title = "Population Density:", palette = "viridis", style = "log10_pretty", n = 5) +
  tm_borders(col = "gray", lwd = 0.5) +
  tm_shape(Region) +
  tm_borders(col = "black", lwd = 1) +
  tm_shape(Region[Region$NAME_1 %in% Jesuit, ]) +
  tm_text("NAME_1", size = 0.9, bg.color = "white") +
  tm_shape(Country) +
  tm_borders(col = "black", lwd = 1.5) +
  tm_layout(panel.show = TRUE,
            panel.labels = "Population Density at Municipal Level",
            panel.label.fontface = "bold",
            legend.bg.color = "white",
            legend.position = c("left", "bottom"),
            legend.frame = TRUE,
            legend.text.size = 1)
```

For the replication tables we run the following regressions:

```
model_1 <- lm(illiteracy ~ distmiss + longi + lati + corr + ita + mis_pry + mis, data = df)
model_2 <- lm(illiteracy ~ distmiss + lati + longi + alti + preci + area + tempe + rugg + river
    + coast + corr + ita + mis_pry+ mis, data = df)

df_arg <- subset(df, COUNTRY == "Argentina")
model_5 <- lm(illiteracy ~ distmiss + lati + longi + mis, data = df_arg)
model_6 <- lm(illiteracy ~ distmiss + lati + longi + tempe + alti + preci + area + rugg
    + river + coast + corr, data = df_arg)

df_pry <- subset(df, COUNTRY == "Paraguay")
model_7 <- lm(illiteracy ~ distmiss + ita, data = df_pry)
model_8 <- lm(illiteracy ~ distmiss + tempe + area + alti + preci + rugg
    + river + coast + ita, data = df_pry)

df_Brazil <- subset(df, COUNTRY == "Brazil")
model_3 <- lm(illiteracy ~ distmiss + lati + longi + mesorregi, data = df_Brazil)
model_4 <- lm(illiteracy ~ distmiss + lati + longi + area + tempe + alti + preci + rugg
    + coast + river + mesorregi, data = df_Brazil)
```

For the replication with Conley Standard Errors:

```
conley_1 <- conleyreg(illiteracy ~ distmiss + lati + longi + mis_pry + mis + corr + ita, data = df,
                      dist_cutoff = 11.0575, lat = "lati", lon = "longi")
conley_2 <- conleyreg(illiteracy ~ distmiss + lati + longi + area + tempe + alti + preci + rugg
    + river + coast + corr + ita + mis_pry + mis, data = df,
                      dist_cutoff = 11.0575, lat = "lati", lon = "longi")

conley_7 <- conleyreg(illiteracy ~ distmiss + ita, data =  df_pry,
                      dist_cutoff = 11.0575, lat = "lati", lon = "longi")
conley_8 <- conleyreg(illiteracy ~ distmiss + area + tempe + alti + preci + rugg
    + river + coast + ita, data =  df_pry,
                      dist_cutoff = 11.0575, lat = "lati", lon = "longi")

conley_3 <- conleyreg(illiteracy ~ distmiss + lati + longi + mesorregi, data = df_Brazil,
                      dist_cutoff = 11.0575, lat = "lati", lon = "longi")
conley_4 <- conleyreg(illiteracy ~ distmiss + lati + longi + preci + rugg + area + tempe + alti
    + river + coast + mesorregi, data = df_Brazil,
                      dist_cutoff = 11.0575, lat = "lati", lon = "longi")

conley_5 <- conleyreg(illiteracy ~ distmiss + lati + longi + corr, data = df_arg,
                      dist_cutoff = 11.0575, lat = "lati", lon = "longi")
conley_6 <- conleyreg(illiteracy ~ distmiss + lati + longi + area + tempe + alti + preci + rugg
    + river + coast + corr, data = df_arg,
                      dist_cutoff = 11.0575, lat = "lati", lon = "longi")
```

We conceptualized and run the regressions with different distances in the following way:

```
df_dist <- df %>%
  filter(complete.cases(distmiss)) %>%
  mutate(distmiss_log = log(distmiss+1),
         distmiss_dec = exp(-0.01 * distmiss),
         distmiss_root = sqrt(distmiss))

model_a <- lm(illiteracy ~ distmiss + lati + longi + area + tempe + alti + rugg + preci
    + river + coast + corr + ita + mis_pry + mis, data = df_dist)
model_b <- lm(illiteracy ~ distmiss_log + lati + longi + area + tempe + alti + preci + rugg
    + river + coast + corr + ita + mis_pry + mis, data = df_dist)
model_c <- lm(illiteracy ~ distmiss_dec + lati + longi + area + tempe + preci + rugg + alti
    + river + coast + corr + ita + mis_pry + mis, data = df_dist)
model_d <- lm(illiteracy ~ distmiss_root + lati + longi + area + tempe + alti + preci + rugg
    + river + coast + corr + ita + mis_pry + mis, data = df_dist)
```

For the visualization of the different distance forms, we employed the following structure, varying the variable `distmiss_` of interest:

```
#Example for Log_Distance:
breaks_log <- seq(min(df_dist$distmiss_log), max(df_dist$distmiss_log), length.out = 10)
labels_log <- sprintf("%.1f to %.1f", breaks_log, c(breaks_log[-1], max(df_dist$distmiss_log)))

Dist_2 <- ggplot() +
  geom_sf(data = df_dist, aes(geometry = geometry, fill = distmiss_log), color = "gray") +
  geom_sf(data = Region, aes(geometry = geometry), color = "#737373", fill = NA) +
  geom_sf(data = Country, aes(geometry = geometry), color = "black", fill = NA) +
  scale_fill_gradient(name = "Distance:",
                      low = "#3f007d", high = "#d4b9da",
                      breaks = breaks_log,
                      labels = labels_log) +
  theme_light() +
  scale_size_identity() +
  labs(x = NULL, y = NULL,
       title = "Logarithmic Distance from Guaraní Jesuit Missions",
       subtitle = "Municipal Level") +
  theme(plot.title = element_text(size = 12, face = "bold"),
        plot.subtitle = element_text(size = 11, face = "italic"),
        legend.position = "inside",
        legend.position.inside = c(0.09, 0.20),
        legend.box.just = "left",
        legend.box.background = element_rect(color = "black", linetype = "solid")) +
  geom_label(data = Region[Region$NAME_1 %in% Jesuit, ], aes(x = X, y = Y, label = NAME_1),
             size = 4, color = "black", fill = "white", fontface = "bold",
             label.padding = unit(0.5, "lines")) +
  guides(fill = "legend") +
  coord_sf(xlim = c(-59.5, -49.7), ylim = c(-33.5, -25.75))
```