Historical Persistence & Geospatial Economics

Applied Economics Research Course

Bas Machielsen

Introduction

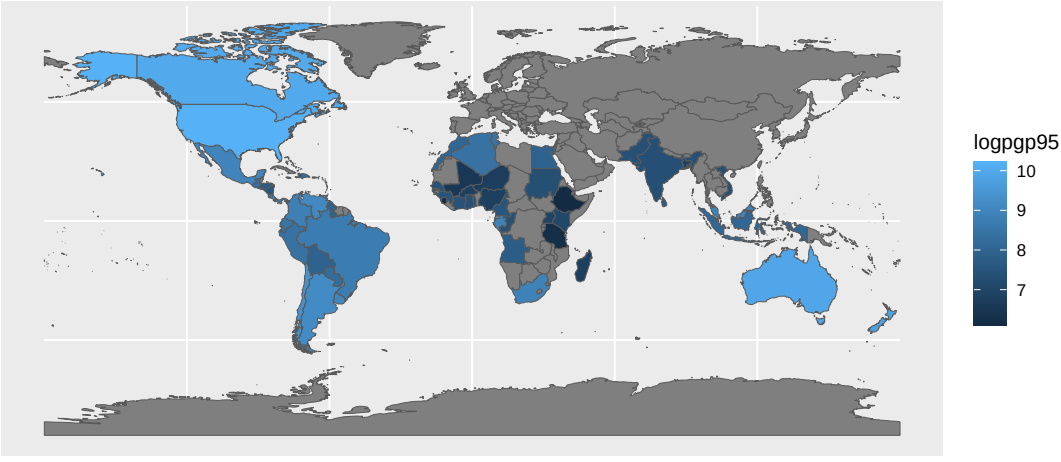
Institutions

- Economic or other outcomes are not only influenced directly by relative prices, costs and benefits.
- They are also influenced by more latent, long-term factors, often called institutions.
  - Example: Acemoglu, Robinson and Johnson (2001):

(..) estimate the effect of institutions on economic performance. Europeans adopted very different colonization policies in different colonies, with different associated institutions. In places where Europeans faced high mortality rates, they could not settle and were more likely to set up extractive institutions. (..) Exploiting differences in European mortality rates as an instrument for current institutions, we estimate large effects of institutions on income per capita.

Institutions

- Here's a world map with GDP/capita in 1995 for a selected number of countries:



Institutions

- In this paper, Acemoglu, Johnson and Robinson (2001) take institutions to be *expropriation risk*: Measures risk of government appropriation of foreign private investment on a scale from 0 (least risk) to 10 (most risk), averaged over all years from 1985-1995.
- If one wants to find the relationship between institutions and economic well-being, one runs into the problem of *endogenity*:
  - Institutions causally affects economic growth
  - But economic growth also affects institutions!
  - Formally:
    - $\text{GDP/Cap}_i = \alpha_0 + \alpha_1 \text{Institutions}_i + \epsilon_i^1$
    - $\text{Institutions}_i = \beta_0 + \beta_1 \text{GDP/Cap}_i + \epsilon_i^2$
- If this is the *true* data-generating process, an OLS regression will not yield the correct coefficient you are interested in (α_1).

Instrumental Variables

- People have devised various strategies to solve this problem. The most often-used strategy is to use an *instrumental variable*
  - An instrumental variable is a variable that exogenously causes a change in X while not directly affecting Y.
  - In the AJR (2001) case: $\text{Institutions}_i = \beta_0 + \beta_1 \text{GDP/Cap}_i + \underbrace{\beta_2 \text{SettlerMortality}_i}_{\text{Instrumental Variable}} + \epsilon_i^2$
  - Now, GDP per capita and Institutions are *endogenously* determined as a result of *exogenous* settler mortality

Instrumental Variables

- You can rearrange these two equations to find that:

$$\text{Institutions}_i = \cdots + \underbrace{\left(\frac{\beta_1\alpha_1\beta_2}{1-\alpha_1\beta_1} + \beta_2\right)}_{=\frac{\beta_2}{1-\alpha_1\beta_1}} \cdot \text{SettlerMortality}_i + \ldots$$

$$\text{GDP/Cap}_i = \cdots + \left(\frac{\alpha_1\beta_2}{1-\alpha_1\beta_1}\right) \cdot \text{SettlerMortality}_i + \ldots$$

- Now you can find (*identify*) α_1 by dividing Cov(GDP/Cap, SettlerMort) by Cov(SettlerMort, Institutions)
- The key (and untestable) assumption here is that settler mortality does not directly influence GDP per capita in 1995.
- Loosely speaking, settler mortality represents an exogenous shock to institutions, allowing us to identify the influence of a change in institutions on GDP per capita

Historical Persistence

Back to historical persistence

- In general, we are dealing with outcomes that are determined endogenously: outcomes now and outcomes in the past are determined by other, latent factors
- To identify the influence of these latent factors, economists often use large-scale, influential events that shock these institutions. They compare outcomes in places that were initially similar, but some of them have been coincidentally exposed to certain events, whereas others have not.
- Usually, this research is based on spatial regression discontinuity designs, comparing places at one side of the border with places at the other side of the border (e.g. Dell, 2010, Lowes and Montero, 2021).

Historical Persistence

- The set-up in general is:
  - Outcome Y_i has "deep roots" caused by some D_i long ago
  - This D_i usually did not came to be randomly, but at the margin near some threshold, it did
  - Example (Dell 2011): Spanish conquest of Latin America was bounded by mountainous areas
- Comparing villages just within vs. just outside of the old imperial borders
- Other seminal papers:
  - Acemoglu et al. (2011), Occupied areas by Napoleon in Germany do better
  - Dell & Olken (2020): Positive development effects of extractive colonialism in Indonesia
  - Voigtlander & Voth (2012): Deep roots of persecution in Nazi Germany

Spatial Regression Discontinuity

- Consider the example of **Roman roads**. Let's suppose that Roman roads, by influencing trade networks, stimulate development:
  - $Y_i = \alpha_0 + \alpha_1 \text{RomanRoad}_i + \alpha_2 \text{Environment}_i + \epsilon_i^1$
- But Roman Roads have not been built randomly by the Romans, but might have built in areas with more suitable environments for development:
  - $\mathbb{P}[\text{RomanRoad}_i] = \beta_0 + \beta_1 \text{Environment}_i + \epsilon_i^2$
- Then, comparing areas with roads to areas without roads doesn't work due to differences in the environment:
  - $\mathbb{E}[Y|RR = 1] = \alpha_0 + \alpha_1 + \alpha_2 \mathbb{E}[\text{Environment}|\text{RR} = 1]$
  - $\mathbb{E}[Y|RR = 0] = \alpha_0 + \alpha_2 \mathbb{E}[\text{Environment}|\text{RR} = 0]$

Identification Strategy

- If we can find two places with the same expected value of getting a road, but only *one* of them gets a road..
- According to the equations on the previous slide, this means two places with the same environment (take the expected value of Y_i conditional on Environment to see that)
- Then, we can estimate α_1 by:
  - $\mathbb{E}[Y|$RomanRoad & P[RomanRoad]] - $\mathbb{E}[Y|$NoRomanRoad & P[RomanRoad]]
- This is what we attempt to do in a spatial regression discontinuity design: compare two arguably identical places with the same probability of treatment, where only one of them actually gets the treatment

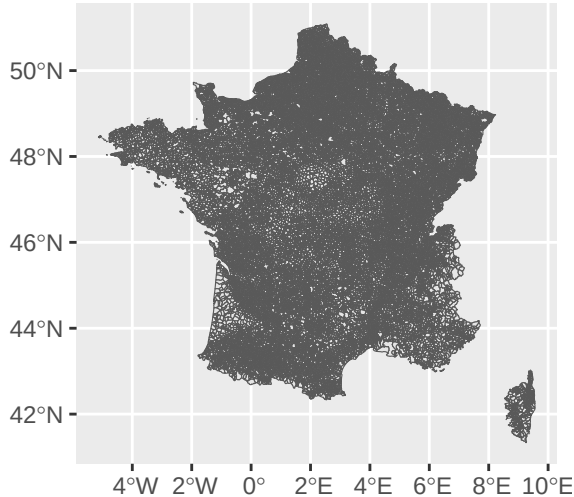
The Data

Data

- I have two a couple of data sets on offer to you:
  - A dataset of French and German municipalities overlaid with Roman roads (`france_germany.shp`)
  - This dataset contains the distance of each municipality to a Roman road, as well as an indicator, whether a road streches through a municipality or not
  - A dataset of Dutch municipalities overlaid with the border of the former Roman empire (`netherlands_rome.geojson`)
  - Again including a variable for distance to the border, and an indicator of whether a present-day municipality was inside or outside the former Roman empire
- You are also free to use or compile a custom dataset using different settings and countries.

Demonstration

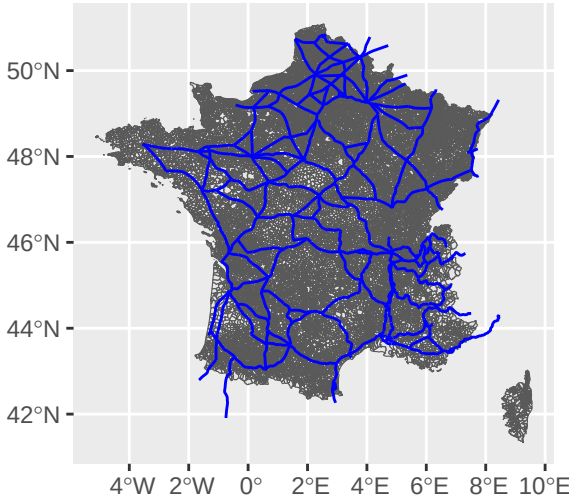
Demonstration



What does this data.frame look like?

```
Simple feature collection with 5 features and 8 fields
Geometry type: POLYGON
Dimension:     XY
Bounding box:  xmin: 4.248763 ymin: 49.54231 xmax: 5.31495 ymax: 50.12165
Geodetic CRS:  WGS 84
  GISCO_ID CNTR_CODE LAU_ID             LAU_NAME POP_2019 POP_DENS_2019
1 FR_08026        FR  08026  Aubigny-les-Pothées      318      21.48498
2 FR_08027        FR  08027 Auboncourt-Vauzelles      102      19.06222
3 FR_08028        FR  08028              Aubrives      874      82.53602
4 FR_08029        FR  08029              Auflance       85      13.92512
5 FR_08030        FR  08030                  Auge       61      13.42680
   AREA_KM2 YEAR                  _ogr_geometry_
1 14.801038 2018 POLYGON ((4.404236 49.75043...
2  5.350899 2018 POLYGON ((4.465646 49.56644...
3 10.589316 2018 POLYGON ((4.770839 50.09282...
4  6.104075 2018 POLYGON ((5.289269 49.62845...
5  4.543151 2018 POLYGON ((4.260257 49.86763...
```

Merge this with the roads

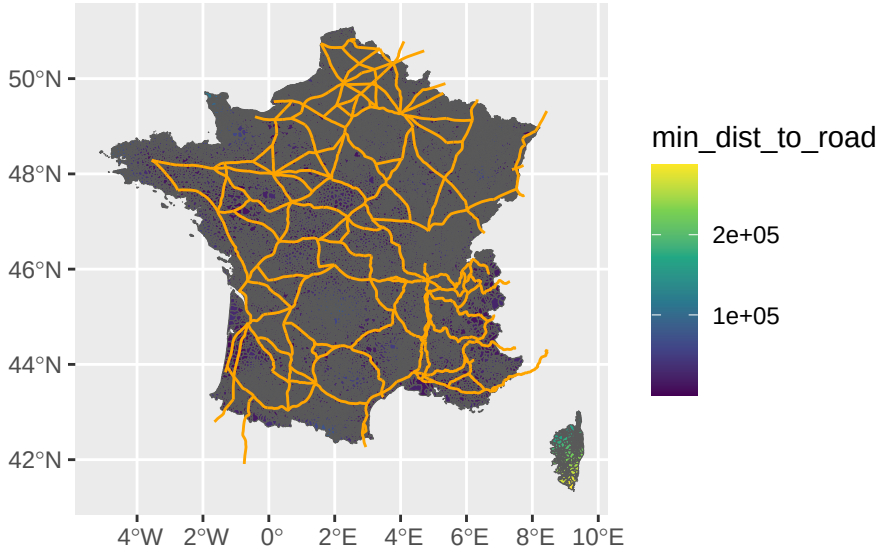


Compute the distance

```
minimum_distances <- fr |>
  st_centroid() |>
  st_distance(roads_in_fr) |>
  apply(1, min)
```

```
 [1] 12389.1098  5728.6071 34012.5501  7624.8761 24728.5649 11728.6840
 [7]  2126.6036  7900.3122   877.4699  9450.5789
```

Plot the result



A small analysis

	Pop. Density
(Intercept)	195.676***
	(4.821)
min_dist_to_road	-0.002***
	(0.000)
R2	0.004
Num.Obs.	35227

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Your task

- Find municipality-level or geographical outcomes that suit an *interesting research question*
  - Present-day or more historical/macro-level outcomes
- Make sure you control for environments!
  - Overlapping regions (provinces, supersets of municipalities)
  - But also possible: terrain, other initial conditions (coastal municipalities)
  - Make it plausible that you compare apples with apples
  - Some roads have also been built with a destination in mind: how to control for that?
- Explore mechanisms through which the effect can work

Geospatial Economics

Introduction

- In economics, many research questions can be answered using geospatial data. For example, questions related to economic development (Beyer et al., 2021; Besley et al., 2022) make frequent use of variation between different geographic units and compare them in aspects such as nightlight density and electricity consumption.
- Similarly, questions related to environmental economics can also be answered using similar data on the basis of geospatial variation (Castells-Quintana et al., 2021; Felbermayer et al., 2022).
- This theme will focus on similar research questions with the aim of introducing students to geospatial data wrangling and econometric methods.

Example

- This theme is a little bit broader than the Historical Persistence theme
- Many research questions are possible.
- Usually, your outcome variable is based on a *shapefile* containing an interesting outcome, e.g. nightlights, species distribution, pollution
- As an example, I have prepared a dataset and analysis of the following research question:
  - Does the air quality influence the performance of French schools?

Loading the Data

- Load the school performance dataset (source: data.gouv.fr)

```
library(here)
school_performance <- read_delim(here('data', 'schools', 'fr-en-indicateurs
```

- Filter so that the year is 2022:

```
school_performance <- school_performance |>
  filter(Session == 2022)
```

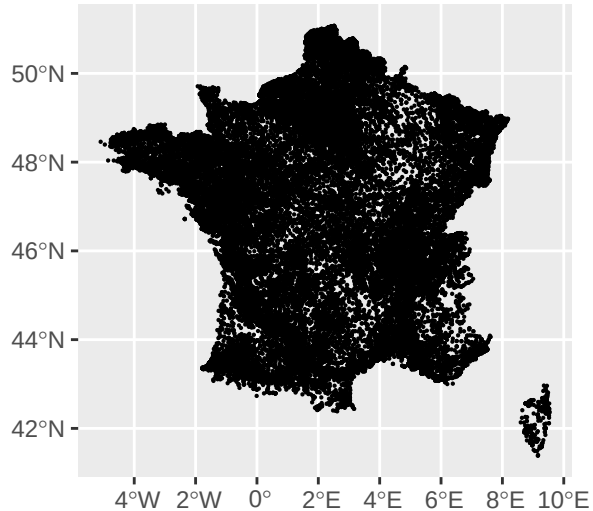
Geocoding the Schools

- We have to geocode the schools using a dataset containing the coordinates of these schools:

```
coordinates <- read_delim(here('data', 'schools', 'fr-en-adresse-et-geoloca
  select(numero_uai, latitude, longitude) |>
  filter(!is.na(latitude), !is.na(longitude)) |>
  st_as_sf(coords=c('longitude', 'latitude'), crs=4326)
```

Filtering

- Filter only Metropolitan France (for the sake of visualization)

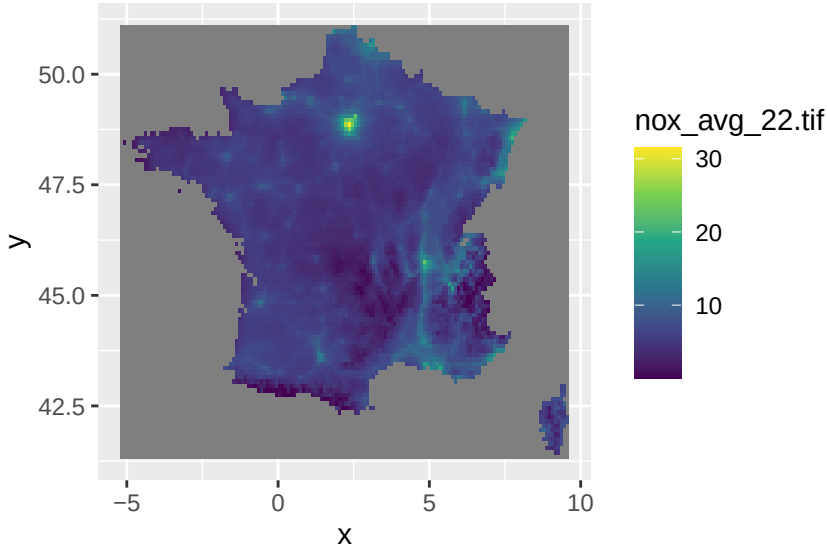


Importing Pollution Data

- This is annual NOx concentration data (Source: European Environment Agency):
  - NOx stands for Nitrogen Oxides—a group of harmful gases (mainly NO and NO) produced when fossil fuels (like gasoline, diesel, coal) burn at high temperatures. Common sources are cars & trucks, power plants and factories
- Read the file, downsample it (to keep it computationally manageable)
- Put it in the same CRS, keep only raster squares that are in France:

```
library(stars); library(dplyr)
sf_use_s2(F)
nox <- read_stars(here('data', 'schools', 'nox_avg_22.tif'))
nox_lowres <- nox |> st_transform(st_crs(france))
nox_lowres <- st_warp(src=nox_lowres, cellsize = c(0.1, 0.1), crs=4326)
nox_lowres <- nox_lowres |> st_crop(st_union(france))
```

Plot Result



Link Coordinates to Schools

```
geocoded_schools <- inner_join(coordinates, school_performance, by = c('numero_uai'='UAI'))
geocoded_schools |> head(5)
```

```
Simple feature collection with 5 features and 29 fields
Geometry type: POINT
Dimension:     XY
Bounding box:  xmin: -2.097861 ymin: 47.83106 xmax: -1.19397 ymax: 48.34177
Geodetic CRS:  WGS 84
# A tibble: 5 x 30
  numero_uai             geometry num_ligne Session `Nom de l'établissement`
  <chr>               <POINT [°]>     <dbl>   <dbl> <chr>
1 0350759K    (-1.68343 48.11588)      1289    2022 COLLEGE ECHANGE
2 0350762N    (-1.19397 48.34177)      1251    2022 COLLEGE THERESE PIERRE
3 0350868D     (-1.3157 47.83106)      1269    2022 COLLEGE PRIVE ST JOSEPH
4 0350869E   (-1.993239 47.89385)      1327    2022 COLLEGE PRIVE STE MARIE
5 0350878P   (-2.097861 47.99924)      1284    2022 COLLEGE PRIVE DE L'HERMINE
# i 25 more variables: Commune <chr>, `Code région académique` <chr>,
#   `Région académique` <chr>, `Code académie` <chr>, Académie <chr>,
#   `Code département` <chr>, Département <chr>, Secteur <chr>,
#   `Nb candidats G` <dbl>, `Taux de réussite G` <dbl>,
#   `VA du taux de réussite G` <dbl>, `Nb candidats P` <dbl>,
#   `Taux de réussite P` <dbl>, `Note à l'écrit G` <dbl>,
#   `VA de la note G` <dbl>, `Note à l'écrit P` <dbl>, ...
```

Link Geocoded Schools to Pollution

```
pollution_values <- st_extract(nox_lowres, geocoded_schools, FUN=mean, na.rm=T)
geocoded_schools <- geocoded_schools |>
  mutate(pollution_values=pollution_values$nox_avg_22.tif)
geocoded_schools |> head(5)
```

```
Simple feature collection with 5 features and 30 fields
Geometry type: POINT
Dimension:     XY
Bounding box:  xmin: -2.097861 ymin: 47.83106 xmax: -1.19397 ymax: 48.34177
Geodetic CRS:  WGS 84
# A tibble: 5 x 31
  numero_uai             geometry num_ligne Session `Nom de l'établissement`
  <chr>               <POINT [°]>     <dbl>   <dbl> <chr>
1 0350759K    (-1.68343 48.11588)      1289    2022 COLLEGE ECHANGE
2 0350762N    (-1.19397 48.34177)      1251    2022 COLLEGE THERESE PIERRE
3 0350868D     (-1.3157 47.83106)      1269    2022 COLLEGE PRIVE ST JOSEPH
4 0350869E   (-1.993239 47.89385)      1327    2022 COLLEGE PRIVE STE MARIE
5 0350878P   (-2.097861 47.99924)      1284    2022 COLLEGE PRIVE DE L'HERMINE
# i 26 more variables: Commune <chr>, `Code région académique` <chr>,
#   `Région académique` <chr>, `Code académie` <chr>, Académie <chr>,
#   `Code département` <chr>, Département <chr>, Secteur <chr>,
#   `Nb candidats G` <dbl>, `Taux de réussite G` <dbl>,
#   `VA du taux de réussite G` <dbl>, `Nb candidats P` <dbl>,
#   `Taux de réussite P` <dbl>, `Note à l'écrit G` <dbl>,
#   `VA de la note G` <dbl>, `Note à l'écrit P` <dbl>, ...
```

Small Analysis

- Clean up variable names, and run a regression

	Avg. Grade
(Intercept)	-0.001
	(0.020)
pollution_values	-0.012***
	(0.002)
nb_mentions_tb_g	0.004***
	(0.000)
R2	0.023
Num.Obs.	5883

+ p < 0.1, * p < 0.05, ** p < 0.01, *** p < 0.001

Logistics

Course Schedule

- Upload research plan in Osiris 13 May 2025 at the latest
- Complete paper and upload in Osiris 26 June 2025 at the latest
- Lecture 1: Plenary Lecture, Introduction
- Lecture 2: Discussion of RQs (Send them on Teams!)
- Lecture 3 + 4: Plenary Lectures, Geospatial Data
  - Depending on time: room to deliberate
- Lectures 5-7: Feedback sessions, room to work and discuss
- Lecture 8: Presentations, feedback from supervisor and co-reader

Expected output

- At the end of week 3 you will upload your research plan in Osiris, including problem statement, (preliminary) literature review and research approach. This plan and your performance in the research group so far will be the input for your supervisors' go/no go decision (crucial with respect to your compliance with the effort requirements).
- In the meeting in week 4, the focus is on the research plan, which is a crucial element in the overall research process. It should contain the motivation for the central problem statement and a plan how to do the actual research.
- Clearly, all members of the supervision group should receive each other's' draft research plan in time, that is, a few days before the meeting itself; to make the meeting as useful as possible, everyone should have carefully read the different research plans, should have given each of these a thought and should have prepared some questions or suggestions.

Expected output

- In each presentation (meeting 4, meeting 6, meeting 8) session, every student gives a short and focused presentation of his/her research project – corresponding to the stage the project is in – and elaborates both on the contents and the used research method and methodology, on choices that have been made and on solved or unsolved problems they have encountered in their work.
- Each presentation is followed by a general discussion where all participants are supposed to contribute by asking questions for further explanation as well as critical questions about the actual research and by providing constructive suggestions and comments. *This feedback needs to be written down in a document and handed in at the start of each session*
- **Final Paper Due**: 26 June 2025

Contact

- You can always contact me at a.h.machielsen@uu.nl
- Also: a **Microsoft Teams group**
- Data available on course website.

Data Sources

- World Values Survey: https://www.worldvaluessurvey.org - Geographically coded information about norms & values
- Demographics & Health Survey: https://dhsprogram.com/data/ - Also geocoded information about D&H
- PRIO GRID Gridded Data
- Eurostat: https://ec.europa.eu/eurostat/web/main/data/database - Database about European countries on aggregate (country) and more disaggregated levels
- Waar Staat Je Gemeente?: Dutch Municipality Data
- Statistiques Locales (FR): French Municipality Data
- Global Climate Database: http://www.worldclim.org/
- My own website data overview (includes municipality data): here
- Clio-Infra: https://github.com/basm92/Clio
- Standard scholarly measures of politics: https://github.com/xmarquez/democracyData
- Correlates of war, armed conflicts: https://rdrr.io/cran/peacesciencer/
- Political protests: https://acleddata.com/
- Labor conflicts: https://datasets.iisg.amsterdam/dataverse/labourconflicts

References & Literature

Acemoglu, D., Cantoni, D., Johnson, S., & Robinson, J. A. (2011). The consequences of radical reform: The French Revolution. American economic review, 101(7), 3286-3307.

Beach, B., & Hanlon, W. W. (2022). Culture and the historical fertility transition. The Review of Economic Studies.

Dell, M. (2010). The persistent effects of Peru's mining mita. Econometrica, 78(6), 1863-1903.

Dell, M., & Olken, B. A. (2020). The development effects of the extractive colonial economy: The dutch cultivation system in java. The Review of Economic Studies, 87(1), 164-203.

Lowes, S., & Montero, E. (2021). Concessions, violence, and indirect rule: evidence from the Congo Free State. The Quarterly Journal of Economics, 136(4), 2047-2091.

Lowes, S. (2022). Kinship Structure and the Family: Evidence from the Matrilineal Belt (No. w30509). National Bureau of Economic Research.

References & Literature

Jones, B. F., & Olken, B. A. (2005). Do leaders matter? National leadership and growth since World War II. The Quarterly Journal of Economics, 120(3), 835-864.

Voth, H. J. (2021). Persistence–myth and mystery. In The handbook of historical economics (pp. 243-267). Academic Press.

Voigtländer, N., & Voth, H. J. (2012). Persecution perpetuated: the medieval origins of anti-Semitic violence in Nazi Germany. The Quarterly Journal of Economics, 127(3), 1339-1392.

Besley, T., Burgess, R., Khan, A., & Xu, G. (2022). Bureaucracy and development. Annual Review of Economics, 14(1), 397-424.

Beyer, R. C., Franco-Bedoya, S., & Galdo, V. (2021). Examining the economic impact of COVID-19 in India through daily electricity consumption and nighttime light intensity. World Development, 140, 105287.

References & Literature

Donaldson, D., & Storeygard, A. (2016). The view from above: Applications of satellite data in economics. Journal of Economic Perspectives, 30(4), 171-198.

Castells-Quintana, D., Dienesch, E., & Krause, M. (2021). Air pollution in an urban world: A global view on density, cities and emissions. Ecological Economics, 189, 107153.

Felbermayr, G., Gröschl, J., Sanders, M., Schippers, V., & Steinwachs, T. (2022). The economic impact of weather anomalies. World Development, 151, 105745.

Beach, B., & Hanlon, W. W. (2018). Coal smoke and mortality in an early industrial economy. The Economic Journal, 128(615), 2652-2675.

Hanlon, W. W. (2020). Coal smoke, city growth, and the costs of the industrial revolution. The Economic Journal, 130(626), 462-488.