GIS for Economists 3

Giorgio Chiovelli Sebastian Hohmann Tanner Regan

18/05/2023

Table of contents

Overview

The plan for today

Replication: Michalopoulos (2012)

- Introduction to the paper
- Empirical strategy and how it relates to GIS data
  - Cross-Country analysis
  - Cross-Virtual-Country analysis
  - Dyadic analysis
- Replication with geopandas

Michalopoulos AER (2012)

Research question

Michalopoulos, Stelios. (2012). "The Origins of Ethnolinguistic Diversity," American Economic Review, 102(4): 1508-1539

What determines ethnolinguistic diversity within and across countries?

Main idea: Diversity in land endowments across regions $\Rightarrow$ formation and persistence of ethnic diversity.

1. Variation in regional land quality $\Rightarrow$ region specific human capital
2. Differences in region specific human capital $\Rightarrow$ barrier to population mixing
3. Limited population mixing between regions $\Rightarrow$ emergence of differential ethnic traits

This was a significant contribution to a large literature that had focused on ethnic diversity as a RHS variable.

This section is based on presentation slides available on the author's website

The empirical analysis establishes that geographic variability, captured by variation in regional land quality and elevation, is a fundamental determinant of contemporary linguistic diversity.

How to do this in practice?

Three main approaches:

- Country level analysis
- Virtual-countries
- Dyadic-analysis

Michalopoulos AER (2012)

Research design: Cross-country analysis

At the **country**-level, what is the effect of land quality variation on the number of languages?

$$\log(\textit{Number of languages}_i) = \beta_0 + \beta_1 \textit{Variation in Land Quality}_i + \gamma \mathbf{X_i} + \eta_i, \quad (1)$$

$\rightarrow$ where i indexes countries.

- What is a concern?

Michalopoulos AER (2012)

Research design: Cross-country analysis

At the **country**-level, what is the effect of land quality variation on the number of languages?

$$\log(\textit{Number of languages}_i) = \beta_0 + \beta_1 \textit{Variation in Land Quality}_i + \gamma \mathbf{X_i} + \eta_i, \quad (1)$$

$\rightarrow$ where i indexes countries.

- What is a concern?
  - Modern centralized states (which often formed along geographic boundaries) affected the distribution of languages (education, language policies, conquest, genocide).
  - Have to account for state-specific histories.

Michalopoulos AER (2012)

Research design: Cross-virtual-country analysis

Idea: **Virtual countries**

- Divide earth into cells of equal size ("virtual countries")
- Then run, as before (note $\mathbf{X_i}$ can include country fixed effects):

$$\log(\textit{Number of languages}_i) = \beta_0 + \beta_1 \textit{Variation in Land Quality}_i + \gamma \mathbf{X_i} + \eta_i, \quad (2)$$

$\rightarrow$ where i now indexes virtual countries.

- Could we still be concerned?

Michalopoulos AER (2012)

Research design: Cross-virtual-country analysis

Idea: **Virtual countries**

- Divide earth into cells of equal size ("virtual countries")
- Then run, as before (note $\mathbf{X_i}$ can include country fixed effects):

$$\log(\textit{Number of languages}_i) = \beta_0 + \beta_1 \textit{Variation in Land Quality}_i + \gamma \mathbf{X_i} + \eta_i, \quad (2)$$

$\rightarrow$ where i now indexes virtual countries.

- Could we still be concerned?
  - Standard concern: omitted variable bias, η_i could be correlated with $\textit{Variation in Land Quality}_i$
  - Can we focus on *similar* regions that differ only in land quality?

Michalopoulos AER (2012)

Research design: Pairwise analysis of adjacent regions

Idea: **Dyadic** analysis of adjacent regions

- Divide earth into cells of equal size (1/25 the size of the previous virtual countries)
- Then run dyad pairs with cell fixed effects:

$$\begin{aligned} \textit{Percentage of common languages}_{ij} = & \alpha_i + \alpha_j + \\ & \beta_1 \textit{Absolute difference in Land Quality}_{ij} + \\ & \gamma \mathbf{X_{ij}} + \xi_{ij}, \end{aligned} \tag{3}$$

$\rightarrow$ where now i and j index adjacent cells.

- Advantage of dyadic structure
  - Minimize concerns that differences in unobservables drive differences in number of languages since focus on adjacent cells
  - See related modern methods paper by Druckenmiller and Hsiang (2019) on Spatial First-Differences.

Michalopoulos AER (2012)

Replication with geopandas

We will cover GIS methods to create data for:

- Countries
- Virtual-countries
- Dyads

For each of these analyses we need data on:

- Languages
  - Linguistic groups' homelands from WLMS, accurate between 1990 and 1995.
- Land quality
  - Agricultural suitability, elevation, climate, proximity to coast, etc.

Michalopoulos AER (2012)

Replication with geopandas: full list of inputs

- Languages: Michalopoulos uses WLMS `http://www.worldgeodatasets.com/language/`. We have an old version of this called *langa.shp*
- Agricultural suitability: `https://nelson.wisc.edu/sage/data-and-models/atlas/data.php?incdataset=Suitability%20for%20Agriculture`
- Population density for different years `http://themasites.pbl.nl/tridion/en/themasites/hyde/download/index-2.html`
- Country boundaries `http://www.naturalearthdata.com/downloads/10m-cultural-vectors/10m-admin-0-countries/`
- Coastline `http://www.naturalearthdata.com/downloads/10m-physical-vectors/10m-coastline/`
- Lakes `http://www.naturalearthdata.com/downloads/10m-physical-vectors/10m-lakes/`
- Elevation `https://cgiarcsi.community/data/srtm-90m-digital-elevation-database-v4-1/`
- Temperature and rainfall `https://www.worldclim.com/current`

All inputs (except elevation, temperature, rainfall) on google drive

Michalopoulos AER (2012)

Replication with geopandas: Settings

Each stage of the python script is listed on the following slides. Check the table of contents in the Jupyter notebook to find the corresponding code location.

1. Main settings
    - Import python packages
    - Set paths for file locations on your computer
    - Set file locations for inputs, temporary, and output files
    - Define other settings (e.g. coordinate systems)

Michalopoulos AER (2012)

Replication with geopandas: Cleaning

2.A. Clean language data

- add unique ID, rename columns, and drop unnecessary fields

2.B. Clean agricultural suitability data

- add coordinate system

2.C. Clean population data

- convert from ASCII raster type to GeoTiff

2.D. Clean elevation data

- convert from AIG raster type to GeoTiff

2.E. Clean climate data

- take averages across months

3.A. Aggregate data by country

- We have a bunch of raster data (agricultural suitability, elevation, temperature, rainfall, population density in different years)
- We want to compute *zonal statistics* (such as mean and standard deviation of agricultural suitability and elevation) in a country.
- Pre-assign all the variables
- Write a loop where each iteration computes Zonal Statistics of a different raster
- Output the results to *.csv*

Michalopoulos AER (2012)

Replication with geopandas: Cross-Country analysis

3.B. Count languages for countries

- *Intersect* WLMS and countries
- *groupby* to count number of languages that intersect each country

3.C. Calculate distances between countries and coast

- Re-set the country coordinate system to an equal area projection
- Find country centroids
- Calculate distance to coast from each centroid.

3.D. Calculate country areas

- Re-set the country coordinate system to an equal area projection
- Calculate area in square kilometers

- Output all the results to *.csv*

Michalopoulos AER (2012)

Replication with geopandas: Cross-Virtual-Country analysis

4.A. Build Grid for virtual countries

- Create a grid of 2.5×2.5 degree cells covering the world
- Add unique ID for each virtual country
- *Intersect* the cells with the actual countries to remove oceans
- *Dissolve* the intersections to get single units for each virtual country
- Clean up holes created by unaligned country borders
- Save grid as *.shp*

4.B. Count languages for virtual countries

- *Intersect* virtual countries with WLMS
- *groupby* to count number of languages that intersect each country
- Output the language counts to *.csv*
- Get virtual countries without languages with *Difference*
  - This creates a few 'broken' geometries that we need to fix.
- Save virtual countries with and without languages as *.shp*

Michalopoulos AER (2012)

Replication with geopandas: Cross-Virtual-Country analysis

4.C. Calculate land and water areas

- Calculate virtual country area
- Intersect virtual countries with lakes, calculate water area for each virtual country

4.D. Calculate distances between virtual countries and coast

- Calculate coordinates of each virtual country centroid
- Calculate distance between centroids and coast
- Get coordinates of point on coast nearest to virtual country

4.E. Aggregate data by virtual countries

- As for the countries, we loop over the different rasters and each iteration uses *Zonal Statistics*

- Output all the results to *.csv*

Michalopoulos AER (2012)

Replication with geopandas: Dyadic analysis

5.A. Create dyad cells

- As for virtual countries above, just change the resolution to 0.5×0.5 decimal degrees.

5.B. Languages spoken in dyad cells

- Before we only cared about the number. Now we want the *percentage common* to the dyad. $\Rightarrow$ need the actual languages. $\Rightarrow$ *Spatial Join* the cells to WLMS
- Output languages per virtual country to *.csv*

5.C. Calculate land and water area for dyad cells

- Same as above for virtual countries

5.D. Calculate distances between dyad cells and coast

- Same as above for virtual countries

5.E. Aggregate data by dyad cells

- Similar to countries and virtual countries, but now use *point query* to extract values from points
- We do this partly because the suitability raster has resolution 0.5×0.5 degrees, and partly because we want to show a new tool, but we could also do this with Zonal Statistics.
- Output the results to *.csv*

5.F. Create polygon neighbours

- Loop through every dyad cell
- Identify each cell's neighbours (neighours are all cells that are NOT *disjoint*)
- Remove self as neighbour
- Output the neighbour pairs to *.csv*