GIS for Economists 3

Giorgio Chiovelli   Sebastian Hohmann

17/03/2020

Table of contents

Overview
- The plan for the session

Paper Replication: Michalopoulos AER (2012)
- Research question
- Research design
  - Cross-country analysis
  - Cross-virtual-country analysis
  - Pairwise analysis of adjacent regions
- Further robustness and channel
- Replication in QGIS
  - Inputs
  - Cross-Country analysis
  - Cross-Virtual-Country analysis
  - Dyadic analysis

Overview

The plan for today

Replication: Michalopoulos (2012)

- Introduction to the paper
- Cross-Country analysis
- Cross-Virtual-Country analysis
- Dyadic analysis

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

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

Ethnic diversity has been used as RHS variable

- Ethnic divisions and economic performance across countries (Easterly and Levine, (1997))
- Fractionalization and public good provision (Banerjee el al (2006)), civil conflict (Fearon, Alesina et al (2003))
- Inequality across ethnic groups (Loury (1977), Esteban and Ray (2007))
- Optimal State formation (Alesina and Spolaore, (1997), Alesina et al (2006))

Ethnic diversity as LHS variable (its economic origins) are less well understood
Main idea: Diversity in land endowments across regions ⇒ formation and persistence of ethnic diversity.

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

Michalopoulos AER (2012)

Research design: Cross-country analysis

Author begins by running the following regression at the **country**-level

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

where i indexes countries.

- Suppose $\hat{\beta}_1 > 0$, significant.
- What is the concern?

Michalopoulos AER (2012)

Research design: Cross-country analysis

Author begins by running the following regression at the **country**-level

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

where i indexes countries.

- Suppose $\hat{\beta}_1 > 0$, significant.
- What is the 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 <u>country</u> fixed effects):

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

where now i indexes *virtual* countries.

- Suppose $\hat{\beta}_1 > 0$, significant.
- Could we still be concerned?

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):

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

where now i indexes *virtual* countries.

- Suppose $\hat{\beta}_1 > 0$, significant.
- Could we still be concerned?
- Standard concern: omitted variable bias, η_i could be correlated with $\textit{Variation in Land Quality}_i$
- Can we somehow focus only on otherwise *similar* regions that differ only in land quality?

Idea: Dyadic analysis of adjacent regions

- Divide earth into cells of equal size (1/25 the size of the previous virtual countries)
- Then run

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

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
- Can include cell fixed effects

Michalopoulos AER (2012)

Further robustness and channels

Indirect evidence: recent migrations

- When focusing on countries, virtual countries, pairs where less than 40% of population can trace ancestry back to 1500AD, results disappear

Channel: location-specific human capital

- many possible channels: ethnic identity formation along geographic lines to defend against invaders, homogenous territories may be easier to defend, geographic differences increase migration costs leading to isolation and ethnic drift, location specific human capital
- If location-specific human capital is the channel, then nonadjacent partitions of a language group should exhibit similar modern modes of subsistence.

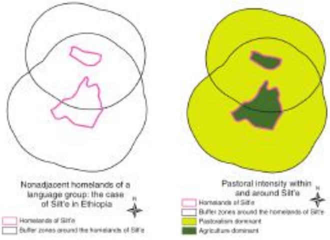

Figure 9A Figure 9B

- Compute "buffers" to account for common regional factors

Channel: location-specific human capital, continued

Run the specification

$$\text{specialization}_{i,g} = \alpha + \beta_1 \text{buffer specialization}_{i,g} + \beta_2 \text{specialization}_{j,g} + \beta_3 \text{land quality}_{i,g} + \epsilon_{i,g}, \quad (4)$$

where i, j are partitions of group g.

$\hat{\beta}_2 > 0$, significant provides evidence in support of location specific human capital.

Michalopoulos AER (2012)

Replication in QGIS: Cross-Country analysis

Inputs all except (large) elevation, temperature, rainfall on google drive

- 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`
- Elevation: `http://topex.ucsd.edu/WWW_html/srtm30_plus.html`
- Temperature and rainfall `https://www.worldclim.com/`
- 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/`

Preparing the WLMS shapefile

- *Fix geometries* to process shapefile (will always have to do)
- *Add autoincremental ID field* for country IDs
- *Field calculator* to change variable names
- *Delete column* to drop some fields

→ *_1_cleanWLMS.py*

Preparing the agricultural suitability raster

- Use *GDAL Warp reproject* to project the raster go WGS 84.
- Use *GDAL Extract projection* to create a permanent projection for the suitability raster.

→ *_2_cleansuit.py*

Looping over the raster files

- 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*
- At the end, use custom function (saw in lecture 2) to output the results to *.csv*

→ *_3_zonalstats_countrylevel.py*

Remaining variables for country-level

- Number of languages in each country: *Intersect* WLMS and countries, *Statistics by categories* the intersection by country-ID, using *ADMIN* as the *CATEGORIES_FIELD_NAME* and *None* as the *VALUES_FIELD_NAME* (produces just the count).
- Distance to coast: Find country centroids with *Centroids*, use *GRASS v.distance* to find distance to coast from country centroid (there are a bunch of steps involved here, we will cover them carefully but note that the implementation could probably be simpler – let us know if you find a way!).
- Country areas: *Reproject Layer* countries to an equal area projection, *Add Field* (DOUBLE, and NULLABLE) for country area, use *Field Calculator* with *area($geometry)/1000000* as *FORMULA*.

→ *_4_other_countrylevel4a-c.py*

Creating virtual countries

- *create grid* for a global raster of 2.5×2.5 degree cells, *Add autoincremental ID field* to create a cell-identifier, *Intersect* the cells with the actual countries (Why are we doing this?), *GRASS v.clean* to clean up the intersection, *Dissolve* to obtain the virtual countries used in analysis.

Obtain the number of languages

- *Intersect* virtual countries with WLMS, *Statistics by categories* the intersection, *Intersect* the result again with the countries (since want to include country fixed effects in the regressions)

Obtain the areas without languages

- *GRASS v.overlay* virtual countries and WLMS to obtain areas without any WLMS identifier.

Additional variables

- Virtual country area, centroid coordinates, area under water, distance to the coast: as before, consult the python script if you are interested.

→ *_5a-e_vcfeats.py*

Zonal statistics

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

→ *_6_zonalstats_vcountry.py*

Michalopoulos AER (2012)

Replication in QGIS: Dyadic analysis

Creating cells

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

Obtain languages spoken in each cell

- Before we only cared about the number. Now we want the *percentage common* to the dyad. $\Rightarrow$ need the actual languages. $\Rightarrow$ *Join attributes by location* of the cells to WLMS

Control variables

- Area, water area, centroids, and coordinates. Straightforward, consult the python script if you are interested.

Obtain cell raster values, intersect with actual countries

- Could do this with Zonal Statistics as table
- Partly because the suitability raster has resolution 0.5×0.5 degrees, and partly because we want to show a new tool, we will use *Add raster values to points* which extracts the value of the underlying raster and assigns it to the point feature lying inside the cell. (What is the issue with this for other rasters?)
- *Intersect* the cell centroids with actual countries to allow for inclusion of country fixed effects and country-level variables in the regression.

$\rightarrow$ *_7a-e_xyz.py*

Michalopoulos AER (2012)

Replication in QGIS: Dyadic analysis

Dyadic structure

- Use *Polygon Neighbours* to crate a table with all the neighbouring identifiers.
- Custom script, modified version based on the script by Ujaval Gandhi

  `https://www.qgistutorials.com/en/docs/find_neighbor_polygons.html`
- Loop through all features
- First, create a list of indices of features that intersect the bounding box
- Then check for all features intersecting the bounding box if they are (a) not the feature itself and (b) not disjoint. If so, they are neighbors and we add their names and feature-IDs to the list of neighboring features

Outputting

- use the *DictWriter* method from the *csv* class (need to import) to export to csv

→ *_7f_dyadfeats_polygon_neighbors.py*