Introduction to GIS Methods in Economics

Giorgio Chiovelli Sebastian Hohmann

Bonn, 28/05/2019

Table of contents

Overview

The plan for today

Replication: Michalopoulos (2012)

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

Michalopoulos AER (2012)

Research question

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 $\Rightarrow$ 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

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 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

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

where now i indexes *virtual* countries.

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

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

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

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 *Variation in Land Quality*$_i$
- Can we somehow focus only on otherwise *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

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

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.

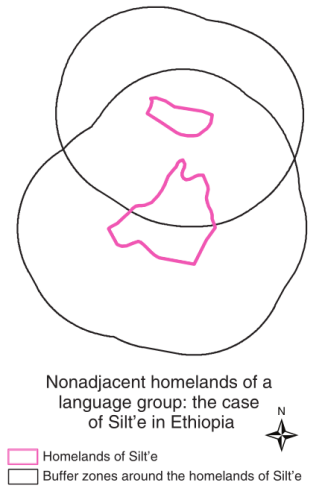

FIGURE 9A

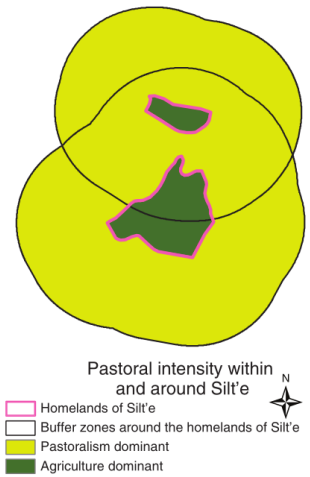

FIGURE 9B

- 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 ArcGIS: Cross-Country analysis

Inputs

- Languages: Michalopoulos uses WLMS (not free). We will use GREG: `https://icr.ethz.ch/data/greg/`
- 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 `http://www.worldclim.org/`
- 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/`

Michalopoulos AER (2012)

Replication in ArcGIS: Cross-Country analysis

Preparing the GREG shapefile

- GREG has some languages occupying more than one polygon, and polygons shared between languages.
- Assigning one (multi-part) polygon per language group can be done, but it is a bit lengthy. Check *_1_cleanGREG.py* (based on Kudamatsu's code) if you are interested.
- We will use the cleaned file.

Preparing the agricultural suitability raster

- Use *Copy Raster* to make a backup of the raster data.
- Use *Define Projection* to define the spatial reference to be WGS 1984.

→ *_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 as Table* and uses *Table to Table (Conversion)* to output the result to *.txt*

→ *_3_zonalstats_countrylevel.py*

Remaining variables for country-level

- Number of languages in each country: *Intersect* GREG and countries, *Dissolve* the intersection by country-ID, using COUNT as an option to count the number of languages inside each country.
- Distance to coast: Find country centroids with *Feature to point*, use *Near* with GEODESIC as option to find distance to coast from country centroid.
- Country areas: *Project* countries to an equal area projection, *Add Field* (DOUBLE, and NULLABLE) for country area, use *Calculate FIELD* with `!SHAPE.AREA@SQUAREKILOMETERS!` and PYTHON_9.3 as options.
- *Table to Table (Conversion)* to output the result of all three.

→ *_4_other_countrylevel.py*

Michalopoulos AER (2012)

Replication in ArcGIS: Cross-Virtual-Country analysis

Creating virtual countries

- *Create Fishnet* for a global raster of 2.5×2.5 degree cells, *Define Projection* to WGS 1984, *Add Field* to create a cell-identifier, *Calculate Field* to populate the cell-identifier, *Intersect* the cells with the actual countries (Why are we doing this?), *Dissolve* to obtain the virtual countries used in analysis.

Obtain the number of languages

- *Intersect* virtual countries with GREG, *Dissolve* the intersection by cell-ID, using COUNT as an option to count the number of languages inside each virtual country, *Intersect* the result again with the countries (since want to include country fixed effects in the regressions)

Obtain the areas without languages

- *Union* virtual countries and GREG, and *Select* the areas without any GREG identifier.

Additional variables

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

→ *_5_vcountry_features.py*

Zonal statistics

- As for the countries, we loop over the different rasters
- Each iteration uses *Zonal Statistics as Table* and uses *Table to Table (Conversion)*

→ *_6_zonalstats_vcountry.py*

Michalopoulos AER (2012)

Replication in ArcGIS: 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. $\Rightarrow$ *Spatial Join* of the cells to GREG, JOIN_ONE_TO_MANY, and use INTERSECT as the join option.

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 *Extract Multi 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.

Michalopoulos AER (2012)

Replication in ArcGIS: Dyadic analysis

Dyadic structure

- Use *Polygon Neighbours* to crate a table with all the neighbouring identifiers. (`http://pro.arcgis.com/en/pro-app/tool-reference/analysis/polygon-neighbors.htm`)
- The first two arguments are the input features and the output table
- The next field sets the ID by with unique polygons of the input features are identified.
- area_overlap determines how to deal with overlapping polygons. Since we don't have any, stick with the default NO_AREA_OVERLAP
- Set NO_BOTH_SIDES to avoid duplicating dyads
- Ignore cluster tolerance, leave the default for out_linear_units (the total length of the coincident edge of the neighbouring polygons – useless to us), and leave the default for out_area_units (units in which the overlap would be reported if we had selected to do so).

Outputting

- *Table to Table (Conversion)* to output the results of all geoprocessing to *.txt*.

→ *_7_dyads_features.py*