Stockholm Doctoral Course Program in Economics

Topics in Applied Microeconometrics: Using GIS

Lecture 5

Zonal Statistics

Masayuki Kudamatsu

IIES, Stockholm University

24 September, 2013

What is Zonal Statistics?

- Calculate summary statistics of raster data values for each polygon
  - Mean
  - Standard deviation
  - Minimum
  - Maximum
  - Sum
  - Range
  - Count
  - (if raster value is integer) Median, Majority, Minority, Variety

- Various global raster datasets out there
  - Population
  - Elevation
  - Land use
- Zonal Statistics lets you aggregate these datasets to whatever polygons you prefer
  - Countries
  - Sub-national districts
  - Grid cells

Application of zonal statistics in economics

- Dell, Jones & Olken (2012)
    - Total population in each weather data grid cell (for calculating population-weighted average temperature for each country)
- Michalopoulos (2012)

Outline

1. Michalopoulos (2012)
2. Replicate Michalopoulos (2012)'s data

1. Michalopoulos (2012)

1.1 Research Question

- Does geographic diversity (variability in agricultural suitability / altitude) affect ethnic diversity?
    - Interesting?
        - Ethnic diversity associated w. growth, govt quality, etc.
        - Thought to be exogenous in economics
    - Original?
        - Cross-“virtual country” regression
    - Feasible?
        - Spatial datasets & GIS software

1.2.1 Data on spatial distribution of languages

- World Language Mapping System
  - Polygons of linguistic groups' homelands in 1990-95
  - Based on SIL International *Ethnologue: Languages of the World* (15th edition)

1.2.2 Data on land endowments

- Suitability for Agriculture by Ramankutty et al. (2002)
  - Raster of 0.5 x 0.5 degree resolution
  - Fraction of a cell that is cultivable around 1990
  - Obtained by “regressing” observed cultivated areas (satellite images) on climate (temperature, precipitation, potential sunshine hours) & soil conditions (carbon density, pH)
- Elevation (and other covariates) from Geographically Based Economic Data

1.3 Empirical challenge

- Cross-country regression suffers from omitted variable bias
  - e.g. Countries that could create a centralized modern state may have reduced # of languages & homogenized land quality across the country at the same time
- But geography is time-invarient. Cannot use panel regression to control for country FEs

1.3 Empirical challenge (cont.)

- With global raster data, one can run cross-subnational region regression with country FE controlled for
- But sub-national region boundaries are endogenous
  - e.g. Splitting districts by ethnicity may be more difficult with more diverse geography

1.3 Empirical challenge (cont.)

- Solution 1: cross-“virtual country” regression
  - Divide the earth into 2.5 x 2.5 degree grid cells
- Solution 2: Dyadic regression
  - For each raster cell, pair it with 8 adjacent cells
  - Measure differences/similarity within each pair
  - Control for raster cell FEs
- Without ArcGIS 10, neither of these is infeasible

1.4 Cross-virtual country regression

$$
\begin{aligned}
\ln(\#languages_i) = \; & \beta_0 + \beta_1 Latitude_i \\
& + \beta_2 SD(elevation)_i \\
& + \beta_3 SD(land\_quality)_i \\
& + \mathbf{X}_i' \beta_{\mathbf{4}} + \xi_i
\end{aligned}
$$

i : virtual country (2.5 by 2.5 degree grid cell on the land)

1.4 Cross-virtual country regression

$$
\begin{aligned}
\ln(\# \textit{languages}_i) &= \beta_0 + \beta_1 \textit{Latitude}_i \\
&\quad + \beta_2 SD(\textit{elevation})_i \\
&\quad + \beta_3 SD(\textit{land\_quality})_i \\
&\quad + \mathbf{X}_i' \boldsymbol{\beta}_4 + \xi_i
\end{aligned}
$$

$\mathbf{X}_i'$ can include country dummies

Role of ArcGIS 10

- Virtual country boundaries
  $\longrightarrow$ Create Fishnet (cf. Lecture 2)
- $\#languages_i$
  $\longrightarrow$ Union + Dissolve
- $SD(elevation)_i$ & $SD(land\_quality)_i$
  $\longrightarrow$ Zonal Statistics as Table
- Country dummy
  $\longrightarrow$ Feature To Point + Spatial Join

Standard errors?

- Standard errors clustered at country level
- In principle, Conley's (1999) method can also be used (cf. Lecture 3)
- But for the whole globe, the location of each unit cannot be recorded in meters

Results: Table 4 columns (4)-(6))

Dep. Var.: Log # of languages

Sample	Full	Tropics	Non-tropics
SD(elevation)	0.082***	0.118**	0.093**
	[0.030]	[0.057]	[0.043]
SD(land_quality)	0.116***	0.103**	0.173***
	[0.033]	[0.048]	[0.055]
Country FE	YES	YES	YES
Other controls	YES	YES	YES
Observations	1663	536	1127

One identification concern

- Land quality variable incorporates climate + soil conditions
- Soil conditions may be endogenous

  e.g. Ethnically homogeneous areas $\rightarrow$ more cultivation due to good governance $\rightarrow$ soil erosion $\uparrow$

- Look at climate-induced suitability for agriculture only (Table 5B col 4)

1.5 Dyadic regression

Unit of analysis: a pair of neighboring 0.5 x 0.5 degree cells (w/i a country)

$$\begin{aligned}\%\mathit{common\_languages}_{ij} &= \gamma_0 + \gamma_1 \Delta(\mathit{land\_quality})_{ij} \\ &\quad + \gamma_2 \Delta(\mathit{elevation})_{ij} + \mathbf{X}_{ij}\gamma_{\mathbf{3}} + \xi_{ij}\end{aligned}$$

- $\mathbf{X}_{ij}$ can include cell FEs for *i* and *j*

Results (Table 6 columns (2)-(5))

Dep. Var.: % common language

Sample	Full	Africa	Europe	Asia
Δ Land_Quality	-0.038***	-0.054***	-0.048**	-0.056***
	[0.012]	[0.018]	[0.021]	[0.016]
Δ Elevation	-0.051***	-0.050***	-0.046**	-0.053***
	[0.006]	[0.016]	[0.022]	[0.009]
Cell FE	YES	YES	YES	YES
Other controls	YES	YES	YES	YES
Observations	156570	35305	11975	74830

1.6 Migration since 1500

- Today’s population: mostly being offsprings of recent migrants for regions like the Americas
- Ethnic diversity should be less influenced by geographic diversity for such regions
  - Too short for geography to form distinct languages
  - Recent years: geography less important for economic activities

1.6 Migration since 1500 (cont.)

- Split the sample by:
  - whether the country (where the cell centroid is located) has more (or less) than 40% of the population being indigenous (ie. not offsprings of migrants since 1500)
    - $\Leftarrow$ World Migration Matrix by Louise Putterman

Results: Table 7 columns (1)-(4)

VARIABLES	Ln Number of Languages		% of Common Languages	
Variation in Elevation	0.470***	-0.352**		
	(0.132)	(0.168)		
Variation in Land Quality	1.220***	0.256		
	(0.406)	(0.361)		
Difference in Land Quality			-0.048***	-0.001
			(0.011)	(0.008)
Difference in Elevation			-0.052***	-0.038
			(0.007)	(0.023)
Country FE	Y	Y	N	N
Region FE	N	N	Y	Y
Continental FE	N	N	N	N
Observations	1352	311	124522	32048

1.7 Taking stock

- Ethnic diversity: cannot be treated as an exogenous variable anymore
- With raster data, one can do a lot better than standard cross-country regression

- Download T:/economics/Lecture5 now
- Launch ArcMap 10 now

In the downloaded Lecture 5 folder,

- Unzip GREG.zip
- Unzip suit.zip in a subfolder (say, called suit)
    - A raster data in the ESRI grid format consists of various files stored in different subfolders
    - Better to keep everything in one subfolder (so it's not confusing when you by mistake browse it in Windows Explorer)

2. Replicate the data used in Michalopoulos (2012)

- Exercise 1: Cross-virtual country data
- Exercise 2: Adjacent cell pair data
    - For polygons of languages spoken, here we use the Geo-referencing of Ethnic Groups dataset (GREG.shp)

  $\Leftarrow$ World Language Mapping System: not for free

2.1 Cross-virtual country data

a. Create the virtual country polygons
   - A review of lecture 2: create grid cell polygons

b. Obtain # of languages spoken
   - Union, Select & Dissolve

c. Obtain land quality measures
   - Zonal Statistics as Table

2.1a: Virtual country polygons

1. Create 2.5-degree grid cell polygons by:
    - Create Fishnet
    - Define Projection

- Cell size: $2.5^\circ \times 2.5^\circ$
- Longitute: 180° West to 180° East
- Latitude: 65° South to 85° North

(see the bottom of page 1522)

Create Fishnet

- Fishnet Origin Coordinate: (-180, -65)
- Y-Axis Coordinate: (-180, -55)
- Cell Size Width: 2.5
- Cell Size Height: 2.5
- Number of Rows: 60 $(= (85 - (-65))/2.5)$
- Number of Columns: 144 $(= (180 - (-180))/2.5)$
- Uncheck "Create Label Points"
- Geometry Type: POLYGON

Alternatively...

- Top: 85
- Bottom: -65
- Left: -180
- Right: -180
- Cell Size Width: 2.5
- Cell Size Height: 2.5
- Number of Rows: 0
- Number of Columns: 0
- Uncheck "Create Label Points"
- Geometry Type: POLYGON

Define Projection

- WGS 1984

$\Leftarrow$ Language polygon data uses WGS1984

2.1a: Virtual country polygons (cont.)

2. Add the unique positive integer ID
   - Add Field
   - Calculate Field

$\Leftarrow$ Zonal Statistics works only if each zone is assigned w/ unique <u>positive</u> integer

cf. In Lec 4 Ex 1, we used these two tools to calculate the area of each polygon

Add Field

- Field Name: cell_id (for example)
- Field Type: SHORT
  - SHORT for integer ranging -32,767 to 32,767
  - LONG for integer ranging -2,147,483,647 to 2,147,483,647
  - # of virtual countries: $60 \times 144 = 8640$

cf. In Lec 4 Ex 1, we chose FLOAT as the area takes fractional values

Calculate Field

- Expression:

$$!FID! + 1$$

Why?

  - FID field gives the unique integer ID including 0
  - Field name needs to be enclosed with ! for Python to calculate values

cf. In Lec 4 Ex 1, we used “float(!SHAPE.AREA!)” for area calculation

- Now run the model and check if virtual country polygons are properly constructed.
- We’ve created the unit of analysis. Next up: dependent variable (# of languages spoken)

2.1b # of languages spoken

- First, we need to separate each virtual country into the parts matched with language polygons and the parts not.
    - Land quality measures are obtained for those matched parts only (see section III.B)
- We can do this by:
    - Union (Analysis)
    - Select

Union (Analysis)

- Essentially the same as the Intersect tool
- Except that it keeps those parts of the input features that are not intersected
- Output features include FIDs from all input features
- For those not intersected, -1 is assigned for FID from the other input features

Why Union, not Intersect?

- To deal with sample selection bias
  - Language map makers might use land quality data to draw the boundary
- Compare the areas with language polygons and those without (Table 3)
- Restrict the sample to virtual countries with the whole area covered by language polygons (Table 5B col. 1)

Union (cont.)

- Input Features:
  - virtual country polygons
  - language group polygons (GREG.shp)
- Uncheck “Gaps Allowed”
  - If checked, a blank area enclosed by input polygons will be an output polygon

- Now run the model. Check the attributes table of the output shapefile.
- Check “FID_GREG”. When it has a value of -1?

Select

- This tool creates a shapefile containing a subset of features from the original shapefile
- The subset is defined by attribute(s)

Select (cont.)

To keep the virtual countries matched with language polygons:

- Expression: "FID_GREG" <> -1

For those unmatched:

- Expression: "FID_GREG" = -1

$\Leftarrow$ Notice the field name is enclosed by double quotation marks

A Tip for the Select tool

- First create the input file for the Select tool by running the model.
- Then double-click the field name instead of typing it on your own. (This automatically adds quotation marks in the proper way.)
  - If the input file is not created yet, double-clicking does not add quotation marks...

- Now we have virtual country by language polygon intersections
- But we need statistics at the level of virtual country
- We use the Dissolve tool to aggregate
    - If we didn’t need to throw away unmatched parts, we could use Spatial Join from the beginning (and use Stata to count # of languages).

Dissolve

- This tool aggregates polygons by attribute(s)
- You can obtain the aggregate statistics, too.

Dissolve (cont.)

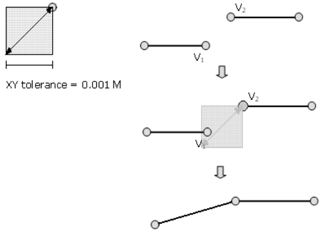
- Dissolve Fields: cell_id
- Statistics Field(s):
  Choose FID_GREG. Then select COUNT for Statistics Type
  - This gives you # of languages in each virtual country
- Check “Create multipart features”

Do the same for unmatched parts of virtual countries, because they may also have multiparts

- Now run the model and see the attribute table to check if FID_GREG properly shows # of languages spoken.
- Check cell_id 1340 (in Chile). Does its FID_GREG equal to 4 as the GREG.shp suggests?
- If not, check the output from the Union tool. They may contain tiny intersection polygons on the border of language groups.

- Polygons in the GREG.shp seems slightly overlapping over each other.
- To ignore this in the geoprocessing, set XY Tolerance a little bit higher in the Union tool.

XY Tolerance



- The size of a square treated as a point in geoprocessing
- By default, it's 0.001 meters.

XY Tolerance (cont.)

- Changing it to a little bit larger value (say, 1 meter) usually solves the problem.
- If geoprocessing tools that create a new shape file by merging data spatially (Spatial Join, Intersect, Union, etc.) do not work properly, see if increasing XY Tolerance a bit solves the issue.

- We’ve created the dependent variable. Next up: the main explanatory variable (s.d. of land quality)

2.1c Land quality

- Suitability for Agriculture data: 0.5 x 0.5 degree raster
- This data is not projected
- So first use Copy Raster to create a copy (so the original is kept)
- Then use Define Projection to assign WGS 1984

(cf. Lecture 1, Exercise 7)

Remember?

A file name for ESRI grid format raster

- cannot be longer than 13 characters.
- has no extension needed

2.1c Land quality (cont.)

- Then use Zonal Statistics As Table to obtain
  - Variation (s.d.) in Land Quality
  - Mean Land Quality
  - # of raster cells

    $\Rightarrow$ Drop virtual countries w/ less than 10 cells in Stata (page 1524)
  - Dispersion in Land Quality (Table 5B col. 3)

Zonal Statistics as Table

- Aggregates input raster data values within each input polygon
- Then create a dBASE table in which
  - Each row: each polygon
  - Columns: various statistics

cf. There is another tool called Zonal Statistics, which creates a raster data file with raster values being one specified zonal statistics (so raster values are the same under the same polygon)

Zonal Statistics as Table (cont.)

- Make sure the polygon features have the unique positive integer ID field
- Choose the polygon shapefile for “Input raster or feature zone data”
- Choose the raster value data for “Input value raster”
- Specify the output table as ***.dbf. Then use Stat/Transfer to convert it to your desired file format

Now, if you cannot launch a window for Zonal Statistics as Table...

- In the menu bar, click Customize > Extensions...
- Check “Spatial Analyst”

If this doesn’t solve the issue, you need to purchase the license for Spatial Analyst extension

- Which includes Zonal Statistics, Extract Values to Points (see below) and other useful geoprocessing tools

Zonal Statistics as Table (cont.)

- Zone field: cell_id (or the name you chose for Add Field above)
- Output table: landq.dbf (5 letters or less if you use ODBC to let Stata read it directly)
- Check “Ignore NoData in calculations
- Statistics type: ALL
  - $\Leftarrow$ We need MEAN, STD, COUNT, and RANGE

- Now you can merge the attribute table from the Dissolve tool and landq.dbf by cell_id and run cross "virtual-country" regressions.
- To obtain control variables and sample selection indicators, do the following assignments:

Assignment 5a

Use Zonal Statistics As Table for:

- Identify which virtual country has less than 3000 inhabitants (to be dropped from the sample; see page 1523) or at least 50000 (column 3, Table 5A)

Obtain the data source on your own:

- See the data source section of the paper for the web address

Assignment 5b

Obtain other covariates

- Absolute Latitude & Migratory Distance from East Africa
  - Feature to Point to create centroid point features
  - Add XY Coordinates

cf. Lecture 4 Exercise 2

Assignment 5b (cont.)

- Variation in Elevation, Mean Elevation, Dispersion in Elevation (column 3, Table 7b), Mean Precipitation, Mean Temperature
  - Data source is G-Econ, which provides an Excel sheet with each row representing 1 x 1 degree cell
  - Follow what we did for Lecture 2 Exercises 1-2, to match virtual countries with G-Econ cells
  - Then use Stata to calculate mean etc.

Assignment 5b (cont.)

- Area of virtual countries
    - Project + Add Field + Calculate Field

  cf. Lec 4 Ex 1
- Ln(Population Density in 1995)
    - Use Area and Population (obtained in Assignment 5a)
- Sea Distance
    - Near with centroid point features as inputs
    - Then use Stata's globdist ado

  cf. Lec 4 Ex 2

Assignment 5b (cont.)

- Water Area (use Natural Earth as inputs)
  - Intersect + Project + Add Field + Calculate Field

  cf. Lec 4 Ex 1
- Number of Countries
  - Use Natural Earth country boundary data
  - Spatial Join (virtual countries as target)
  - Then use Stata's collapse (count) command)

  cf. Lec 4 Ex 1

2.2 Adjacent cell pair data

In ArcGIS, create two tables

a. Table 1
   - Each row is the 0.5 by 0.5 degree cell
   - Columns: cell ID, language1, language2, ..., land quality, elevation, ...

b. Table 2
   - Each row is a pair of adjacent cells
   - Columns: cell i ID, cell j ID

- Then in Stata...
  - Merge table 2 w/ table 1 by cell i ID
  - Merge table 2 w/ table 1 by cell j ID
  - Create the pair level variables

2.2a Create 0.5-degree cell polygons

Create Fishnet

- Same as before except:
  - Cell Size Width: 0.5
  - Cell Size Height: 0.5

Then Define Projection to assign WGS1984

And assign cell ID by Add Field and then by Calculate Field with !FID!+1

- We do not need unique integer ID, as we don’t use Zonal Statistics this time
- But to merge properly in Stata later, it’s a good idea to define each cell’s ID here

2.2b Land quality

- We use:
  - Feature To Point
  - Extract Values To Points

(We can instead use Zonal Statistics as Table, but there’s no aggregation needed.)

Feature To Point

- Creates centroid point features for each input polygon/polyline (cf. Lec 2)

Extract Values To Points

- “Merge” point features with raster data by location
- A new field called RASTERVALU will be added in which the raster data value at the location of each point is stored
- If the underlying raster has no data, the value -9999 is assigned.

- Now run the model.
- It takes quite a while as there will be so many 0.5-degree cells across the world.

Assignment 5c: Cell areas

- Need to calculate each cell area to restrict the sample to pairs whose total area is at least 1000 sq km (see page 1529)
- Project + Add Field + Calculate Field

Assignment 5d: Languages spoken

- In this case, we can use Spatial Join
- Then in Stata, reshape the file structure to the wide format so that each row is the 0.5-degree cell

- We’ve extracted data to construct dependent and explanatory variables at the 0.5 degree cells
- Now create the unit of observations: adjacent cell pairs.

2.2c List of adjacent cell pairs

- We use Buffer + Spatial Join
  (cf. Lecture 3)
    - Create a buffer polygon of 0.8 degrees radius (to include diagonally contiguous cell centroids) for each cell centroid
    - Spatial-join the buffer polygon with cell centroids
    - In Stata, drop rows if cell i ID $\geq$ cell j ID
- But before doing so...

Select

- First, keep only those cell centroids that have land quality measures (ie. drop cells on the water)
- Otherwise, Buffer and Spatial Join take very long
- The Select tool can be used for this purpose

Select (cont.)

- Input Features: the output from Extract Values To Points
- Expression:

  "RASTERVALU" <> -9999

  - Always enclose the field name with double quotation marks

Buffer

- Input Features: the cell centroid points
- Distance: Linear unit 0.8 Decimal degrees

- Next we use Spatial Join.
- If we want to execute Spatial Join in the Model Builder, run the model now to create input features for Spatial Join
- Otherwise, move on

(cf. Lecture 3)

Spatial Join

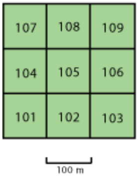
- Target Features: the 0.8-degree buffers
- Join Features: the cell centroid points
- Join Operation: JOIN ONE TO MANY
- Check "Keep All Target Features”
- Field Map...
- Match Option: INTERSECT or CONTAINS

Field Map (review)

- Keep everything if we execute Spatial Join in the Model Builder
- But before exporting a Python script, always delete everything. Otherwise Python won’t work properly.

- Instead of Buffer & Spatial Join, we can use the Polygon Neighbors tool to identify adjacent cell pairs (new in ArcGIS 10)
- Using the polygon features as inputs, it creates a dBASE file that lists the pair of adjacent polygons

INPUT POLYGONS



OUTPUT TABLE

OBJECTID *	src_myCode	nbr_myCode	LENGTH	NODE _COUNT
1	101	102	100	0
2	101	104	100	0
3	101	105	0	1
4	102	101	100	0
5	102	103	100	0
6	102	104	0	1
7	102	105	100	0
8	102	106	0	1
9	103	102	100	0
10	103	105	0	1
11	103	106	100	0
12	104	101	100	0

- Now run the model, to see what outputs we will obtain. (today we don’t have time for editing Python scripts to run)
- While we’re waiting for the Model to be run, learn a couple of new things about Python for ArcGIS

2.3 Python scripting

- Export the Pythons script.
- Browse the exported script.
- Notice the following command:

  arcpy.CheckOutExtension("spatial")

- Without this command, Zonal Statistics as Table and Extract Values To Points won’t work.
- Don’t forget to copy this command to your template (or include this command in your template)

2.3 Python scripting (cont.)

- Also notice that, for the Select tool,

  "\"FID_GREG \"<> -1"

  "\"RASTERVALU \"<> - 9999"

- " in Python: enclosing a string
- " in ArcGIS: enclosing a field name
- To prevent Python from interpreting " as string encloser, you need to write \" \".

3. What we've learned for ArcGIS

- Obtain summary statistics (mean, s.d., etc.) of a raster data variable for each polygon
- Assign a raster data variable to each point feature
- Create a new shape file of a subset of input features whose attributes satisfy some criteria

4. Where to find spatial data

(should have done in Lecture 1...)

- FAO Geonetwork (www.fao.org/geonetwork)
- Devecondata (devecondata.blogspot.com /search/label/GIS)
- Keep an eye on the publication of papers using spatial data

References for Lecture 5

Conley, T. G. 1999. “GMM estimation with cross sectional dependence.” *Journal of Econometrics* 92(1): 1-45.

Dell, Melissa, Benjamin F. Jones, and Benjamin A. Olken. 2012. “Temperature Shocks and Economic Growth: Evidence from the Last Half Century.” *American Economic Journal: Macroeconomics*, 4(3): 66-95.

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

Ramankutty, Navin et al. 2002. “The global distribution of cultivable lands: current patterns and sensitivity to possible climate change.” *Global Ecology and Biogeography* 11(5): 377-392.