Stockholm Doctoral Course Program in Economics

Topics in Applied Microeconometrics:
Using GIS

Lecture 2

Spatial Join

Masayuki Kudamatsu

IIES, Stockholm University

6 September, 2013

What is *spatial join*?

Matching *features* from two different vector datasets by location

- Surveyed villages & Weather data grid cells
- Islands & Wind data grid points
- Zip-code zones & Air pollution monitors (Currie & Neidell 2005)

Benefits from spatial join

- Expand the set of feasible research questions
  - e.g. Weather data often available at earth grid points, not at districts
- Control for FEs of very small areas (within administrative districts)
  - $\Leftarrow$ No need to aggregate to merge the data

Outline

1. Feyrer and Sacerdote (2009)
2. Kudamatsu, Persson, & Stromberg (2012)
3. Replicate Kudamatsu et al (2012) in ArcGIS
4. How to use a Python script for ArcGIS geoprocessing

1. Feyrer and Sacerdote (2009)

- An example of constructing instruments from spatial data
- And the instruments are merged by spatial join.

1.1 Research question

Impact of colonial rule on income today for islands

- Original?
  - Island dataset
  - Wind patterns as instruments
- Feasible?
  - Satellite wind data spatially joined to Island data
- Interesting?

1.2 Data

- UNEP Island Directory (islands.unep.ch)
  - Geographic coordinates of each island
- Eastward & northward wind speeds from CERSAT (www.ifremer.fr/cersat)
  - 1-degree longitude/latitude grid

$\Rightarrow$ Can be spatially joined

1.3 Main empirical specification

$$y_i = \alpha + \beta c_i + \mathbf{x}_i'\gamma + \varepsilon_i$$

- y_i Log GDP per capita / Infant mortality per 1,000 in island *i*
- c_i Length of colonial rule (in centuries) / year of colonization
- $\mathbf{x}_i$ Latitude (in absolute value), Area, Pacific dummy, Atlantic dummy

How to choose control variables

$$y_i = \alpha + \beta c_i + \mathbf{x}_i' \gamma + \varepsilon_i$$

- Why do the authors control for Latitude (in absolute value), Area, Pacific dummy, and Atlantic dummy?

1.3 Main empirical specification

$$y_i = \alpha + \beta c_i + \mathbf{x}_i' \gamma + \varepsilon_i$$

- c_i instrumented by mean & sd of east-west wind speed

“Theory” behind the 1st-stage

- Until 20c, sailing required steady wind (Age of Sail)
- Latitude sailing was the main form of nagivation
  - $\Leftarrow$ Longitude of islands unknown

$\Rightarrow$ East-west wind speed is crucial for where colonizers end up

1.4 Results

Next page taken from Table 2 of Feyrer & Sacerdote (2009)

	(1) Log GDP per Capita	(2) Log GDP per Capita	(3) Log GDP per Capita—IV
Number of centuries a colony	0.42	0.491	0.712
	(0.076)***	(0.110)***	(0.253)***
First year a colony			
Final year a colony			
Remained a colony in 2000			
Abs (latitude)		0.053	0.054
		(0.012)***	(0.011)***
Area in millions of sq km		−20.374	−21.738
		(3.894)***	(3.970)***
Island is in Pacific		0.752	1.018
		(0.464)	(0.559)*
Island is in Atlantic		0.425	0.188
		(0.395)	(0.477)
Constant	7.472	6.033	5.484
	(0.205)***	(0.552)***	(0.834)***
Observations	81	81	81
R-squared	0.273	0.527	0.498

How to report regression results

- Add control variables step by step
  - Changes in the coefficient size tell you how these control variables correlate with the main regressor
- Report OLS estimates as well as IV estimates
  - Difference between the two tells you what causes OLS estimates biased
  - And more reasons to be discussed below

Issues on IV estimation

- Exclusion restriction
- Weak instruments
  - cf. Murray (2006)
- Local average treatment effect
  - cf. Imbens (2010)

IV Issue 1: Exclusion restriction

- Wind: no longer important for sailing after steamships became the norm in 20c (Age of Steam)

$\Rightarrow$ No direct impact on GDP per capita / infant mortality today

- Are you convinced?

IV Issue 1: Exclusion restriction

- Wind: no longer important for sailing after steamships became the norm in 20c (Age of Steam)

$\Rightarrow$ No direct impact on GDP per capita / infant mortality today

- Are you convinced?
- How about pre-colonial institutions?

Checking exclusion restriction

- Regress pre-treatment outcomes (& other covariates) on instruments, if available
- It would be nice if income level of islands before colonization were available...

IV issue 2: Weak instruments

Quick recap: With weak IVs, 2SLS is problematic:

- Inconsistency (Bound et al. 1995)
- Finite sample bias gets large, even in very large sample (Bound et al. 1995)
  - Biased towards OLS estimate
- S.E. too small

Recommended practices

- Report the 1st stage & *F*-stat on excluded instruments
  - *F*-stat should be at least 10
  - If 2+ endogenous regressors, report Cragg-Donald statistics (Stock, Wright, & Yogo 2002)
  - In Stata, use *ivreg2* with *first* or *ffirst* option.
- Report the OLS estimate as well.
  - If $|\hat{\beta}_{OLS}| < |\hat{\beta}_{IV}|$, zero effect can be rejected at least.

Table A1 of Feyrer & Sacerdote (2009)

	(1) Number Centuries a Colony
East-west vector of wind	−0.236
	(0.070)***
Monthly standard dev. of east-west vector	0.508
	(0.238)**
Final year a colony	
Remained a colony after 2000	
Abs(latitude)	0.015
	(0.013)
Area in millions of sq km	8.532
	(4.671)*
Island is in Pacific	−1.494
	(0.354)***
Island is in Atlantic	0.782
	(0.362)**
Constant	0.756
	(0.833)
Observations	81
R-squared	0.681
F statistic for Instruments	5.81
p-value	.005

IV issue 3: LATE

- Estimated impact is the LATE for areas colonized due to wind patterns
- Is this a coefficient of interest when we think of the impact of colonization?

One more recommended pratice

- Show reduced-form results, too
  - For those skeptical on exclusion restriction (Deaton 2010)
  - To see if weak IV issues are serious (Angrist & Pischke 2009, p. 213)
    - Reduced-form coefficient $\propto$ true causal impact size
  - For those who don't think LATE is of interest (Deaton 2010)

2. Kudamatsu, Persson, & Stromberg (2012)

- Use continent-wide natural experiments with a very large number of observations to estimate the impact of disease/nutrition environments on a rare event like infant mortality
- Which is infeasible with RCTs

2.1 Research question

How do annual weather fluctuations kill infants in Africa?

- Interesting?
  - Infant death: big issue in Africa
  - Implications for climate change impact
- Original?
  - No continent-wide study with comparable data
  - RCTs on combatting infant mortality: no statistical power to detect the impact
- Feasible?

2.2 Data

- DHS for infant death at surveyed clusters
- ERA-40 for weather at 1.25-degree grid points

Spatially joined by ArcGIS

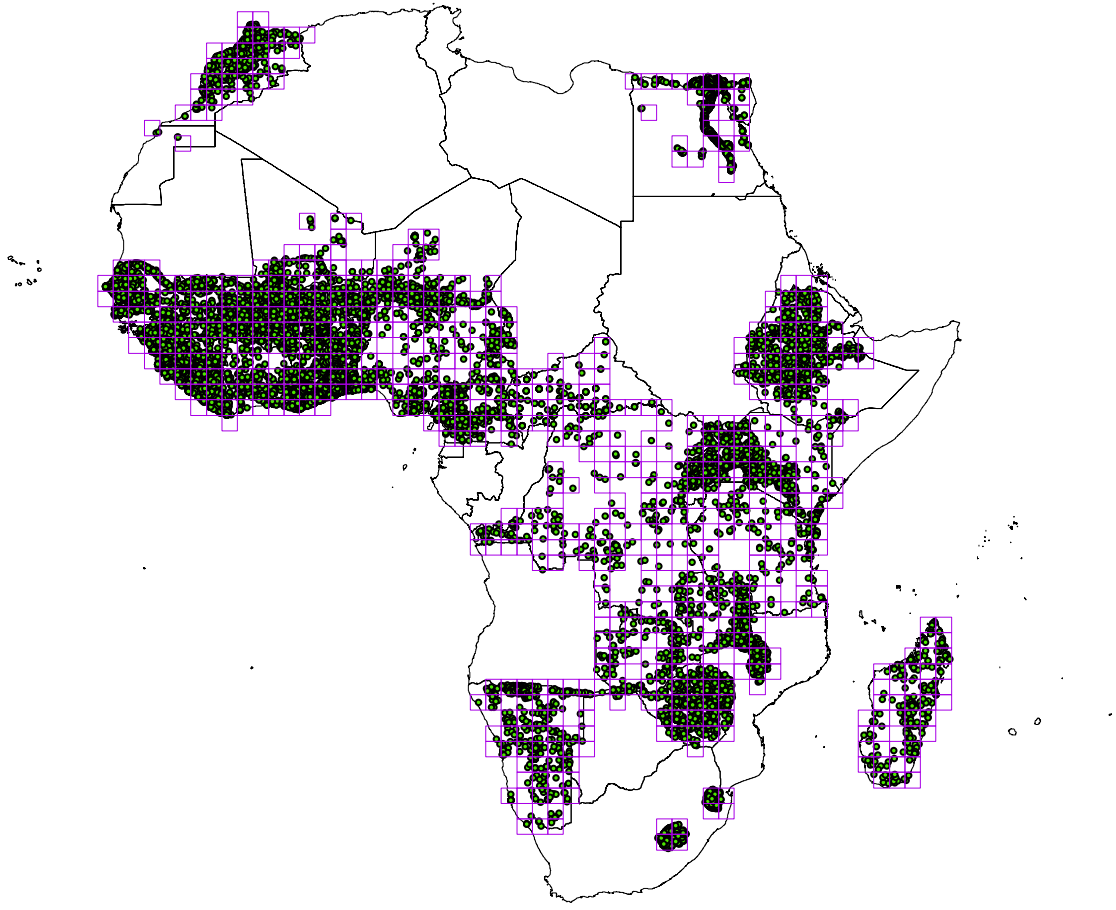

Figure 2 – DHS clusters and ERA-40 grid in sample

Benefits from spatial join

- Estimating heterogeneous impacts by non-administrative boundaries
  - Severity of malaria infection
  - Climate zones
- Controlling for country-by-year FEs

2.3 Empirical specification

$$m_{i,c,t} = \alpha_{c,s} + \alpha_{x,y} + \mathbf{z}'_{g,t}\boldsymbol{\beta} + \varepsilon_{i,c,t}$$

$m_{i,c,t}$ Indicator for death before age of 12 months for baby i born in cluster c in date t (in month)

$\mathbf{z_{g,t}}$ Weather variables at grid point g (the nearest one from cluster c)

2.3 Empirical specification

$$m_{i,c,t} = \alpha_{c,s} + \alpha_{x,y} + \mathbf{z}'_{g,t}\boldsymbol{\beta} + \varepsilon_{i,c,t}$$

$\alpha_{c,s}$ FE for cluster c in calendar month s (of date t)

$\alpha_{x,y}$ FE for country x (where cluster c is located) in year y (of date t)

Identification

Cluster-month FE

$\Rightarrow$ Deviation from the monthly mean

identifies the weather impact

- Otherwise the location-specific seasonality contaminates the estimation

e.g. adaptation to bad months for baby survival

$\Rightarrow$ underestimate the impact of weather shocks

Country-year FE

$\Rightarrow$ Country-wise trends in weather and infant mortality are non-parametrically controlled for

e.g. Rain & Democracy (Bruckner & Ciccone 2011), Democracy & Infant mortality (Kudamatsu 2012)

- Would be infeasible if weather data were available at country level

Standard errors

$$m_{i,c,t} = \alpha_{c,s} + \alpha_{x,y} + \mathbf{z}'_{g,t}\boldsymbol{\beta} + \varepsilon_{i,c,t}$$

At which level should we cluster standard errors?

Standard errors (cont.)

Clustered at grid point level

- across clusters matched with the same grid point:

  error term: correlated

  weather variables: same

- across time within the same grid point

  error term: serially correlated

  weather variables: serially correlated

Weather variables

For infant survival in Africa, more rain is

- Good in terms of crop yields & nutritional intake
- Bad in terms of malaria infection risk

Weather variables (cont.)

Aggregate rainfall (& temperature for malaria) in different ways to separately estimate these two opposing effects of rainfall

- Crop yield: Total rainfall during the growing season
- Malaria risk: At least 60mm monthly rainfall in the past 3 months etc. (Tanser et al. 2003)

2.4 Results on malaria

- Epidemic areas

  # of malaria months per year on average $\in (0, 4]$

  are vulnerable to annual weather fluctuations

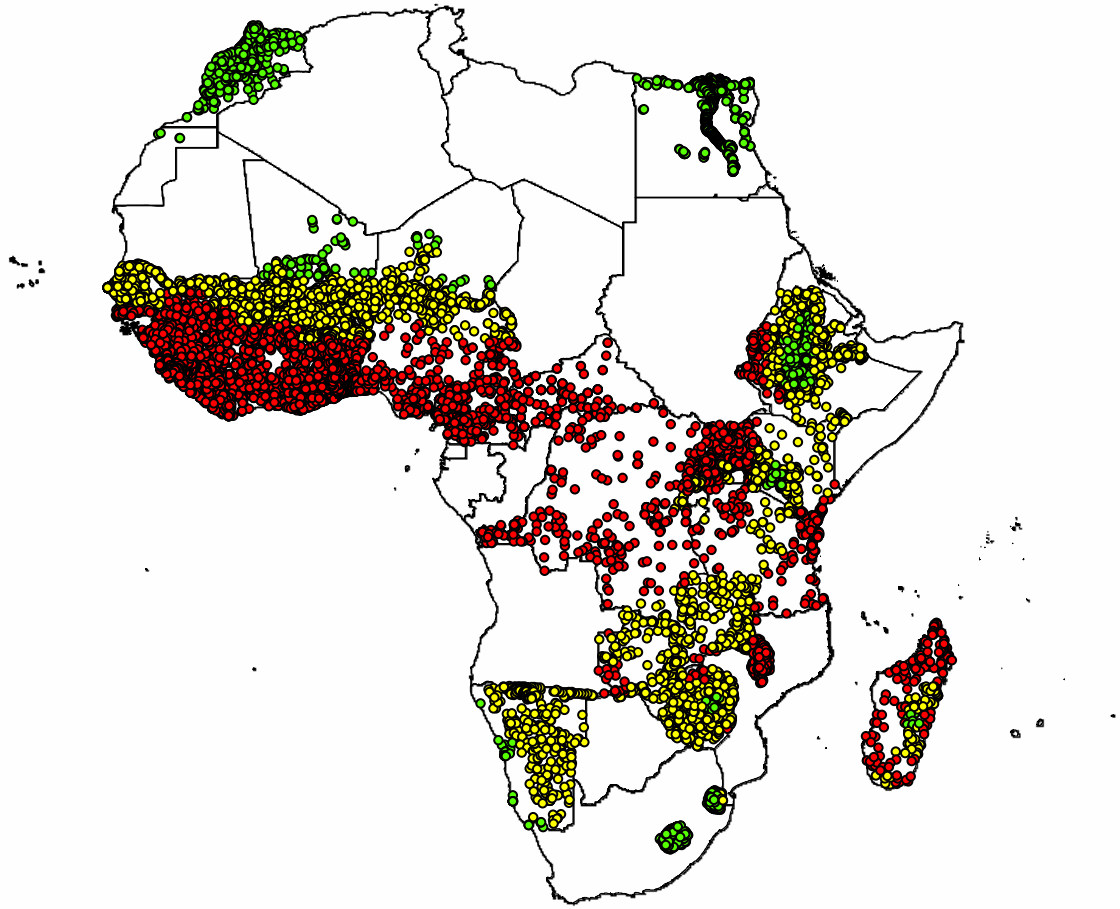

Figure 3 – Malaria exposure zones in Africa

Taken from Kudamatsu et al. (2012)

Dep. Var.: Infant death dummy $\times$ 1000		
Sample	Endemic	Epidemic
# of malaria months $t-11$ to t	-0.16	0.94*
	(0.53)	(0.54)
Country-year FE	YES	YES
Cluster-month FE	YES	YES
# of ERA-40 cells	365	275
Observations	389116	377361

S.E. clustered at ERA-40 cells in parentheses

* significant at 10%, ** 5%, *** 1%

Dep. Var.: Infant death dummy × 1000		
Sample	Less epidemic	More epidemic
# of malaria months		
$t-11$ to t is:		
0	0.30	-3.40
	(2.76)	(2.97)
1-2		-7.15***
		(2.23)
3-4	1.31	
	(3.80)	
5-6	15.62	-4.92
	(11.89)	(3.67)
7+	38.44**	15.67**
	(15.62)	(7.70)
Country-year FE	YES	YES
Cluster-month FE	YES	YES
# of ERA-40 cells	150	125
Observations	187858	189503

2.5 Results on nutrition

- Arid areas (high temperature, low rainfall): vulnerable to annual weather fluctuations

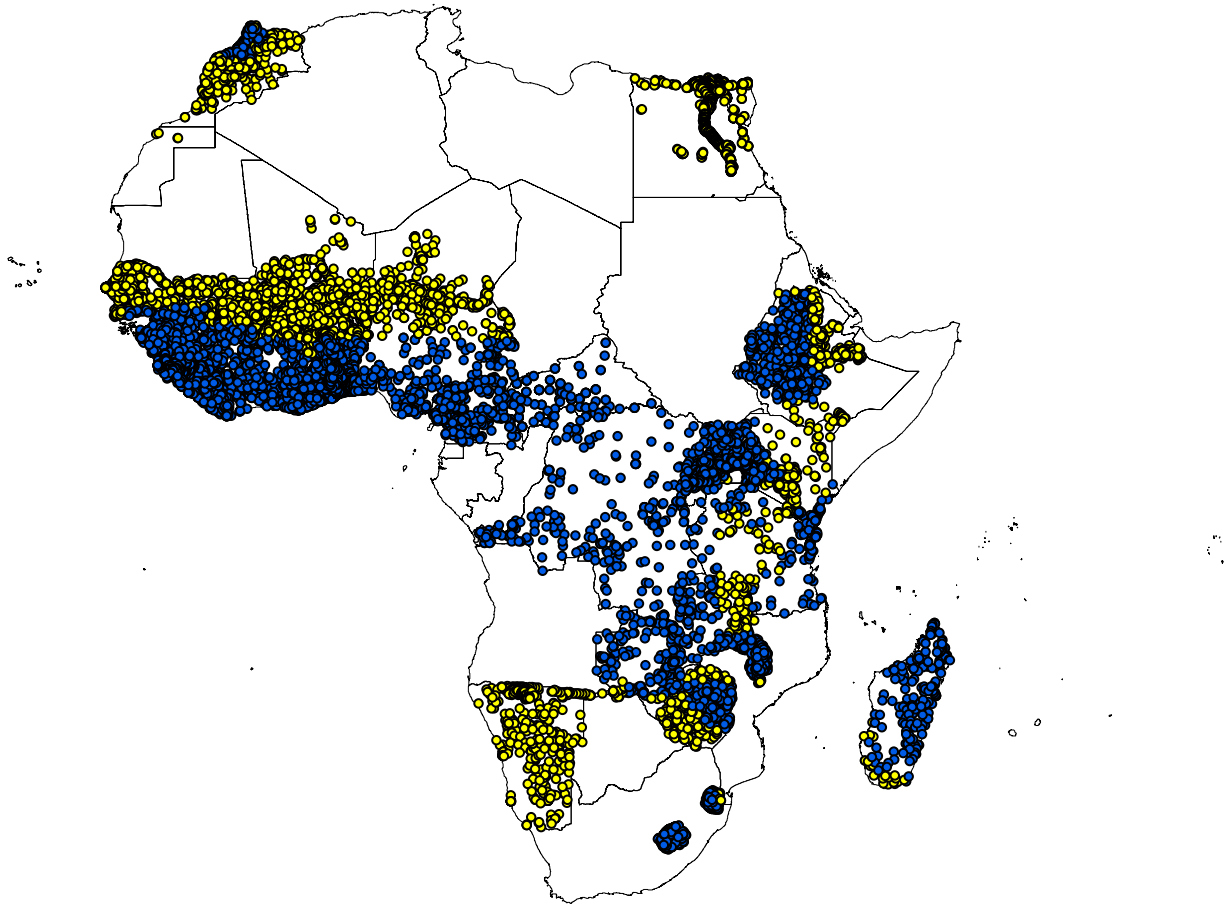

Figure 8 – Arid and rainy climate zones in Africa

Taken from Kudamatsu et al. (2012)

Dep. Var.: Infant death dummy × 1000		
Sample	Rainy	Arid
Growing season	-0.025	-0.039
rainfall (cm)	(0.025)	(0.110)
Drought dummy	-2.64	23.1***
	(12.2)	(8.49)
Flood dummy	-0.99	-1.75
	(3.94)	(3.79)
Country-year FE	YES	YES
Cluster-month FE	YES	YES
# of ERA-40 cells	439	304
Observations	481018	481453

S.E. clustered at ERA-40 cells in parentheses

* significant at 10%, ** 5%, *** 1%

- Launch ArcMap now.
- Download the folder “Lecture2” from:

Network > INK00001 > teacher > Economics

3. Replicate Kudamatsu et al (2012) in ArcGIS

We will learn two things:

- Create a shapefile of graticule
- Use the Spatial Join tool

Why a graticule shapefile?

- In some datasets (weather data in particular), the unit of observations is a grid point (latitude and longitude at equal intervals).
  - WorldClim (Dell, Jones, & Olken 2009)
  - GCPC (Miguel et al. 2004)
  - TOMS air pollution index (Jayachandran 2009)
- A graticule shapefile allows you to spatially merge such data with other data

Exercise 1: Create a graticule shapefile

By using Model Builder, create ERA-40 grid point data

- Geoprocessing tools to use:
  - Create Fishnet
  - Define Projection

    (cf. Lecture 1 Exercise 7)
  - Feature To Point
  - Add XY Coordinates
  - Spatial Join

ERA-40 data

- Spatial resolution: 1.25 °x 1.25 °
- Africa: roughly spans between
  - 40°S to 40°N
  - 20°W to 60°E
- We want to create square polygons whose centroid is from (-20, -40) to (60, 40) at the interval of 1.25°
  - See below for why polygons, not points

Create Fishnet

- Fishnet Origin Coordinate: (-20.625, -40.625)
- Y-Axis Coordinate: (-20.625, -10.625)
- Cell Size Width: 1.25
- Cell Size Height: 1.25
- Number of Rows: 65
- Number of Columns: 65
- Uncheck "Create Label Points"
- Geometry Type: POLYGON

Create Fishnet (alternatively...)

- Top: 40.625, Bottom: -40.625
- Left: -20.625, Right: 60.625
- Cell Size Width: 1.25
- Cell Size Height: 1.25
- Number of Rows: 0
- Number of Columns: 0
- Uncheck "Create Label Points"
- Geometry Type: POLYGON

Define Projection

- The Create Fishnet tool does not assign any coordination system to the output file
- As the DHS cluster locations use WGS 1984, choose the same.
- How? See Lecture 1 Exercise 7

- Our purpose is to assign the ERA-40 grid point data coordinate to each DHS cluster
- By default, the Create Fishnet tool does not include the centroid coordinate as attributes for each polygon
- We need to add centroid coordinates to each polygon by using the next three tools

Feature To Point

- This tool creates a point feature class of centroids of the input polyline/polygon features
- Don’t try to use the INSIDE option for the point location. It usually doesn’t work properly (at least for ArcGIS version 9.3).
- But this tool itself doesn’t add coordinates to the output features. So...

Add XY Coordinates

- Works with point features as inputs
- This tool overwrites the input data. If you prefer keeping the input data, use Copy Features first (as in Lecture 1 Exercise 7).

- Now we want to attach these centroid coordinates to the original polygon features.
  - We prefer polygons to points, as merging so many point features (over 700 in ERA-40) with another set of so many point features (over 17000 in DHS) usually doesn't work properly in ArcGIS.
- Spatial Join tool does this.

Spatial Join in general

- Merge two datasets spatially
- Soon later we'll use this tool to merge point features (DHS clusters) with polygon features (whose centroid is ERA-40 grid point) that contain them
- You can merge any combination of vector datasets (cf. L3, L5, L6)

Spatial Join for ERA-40 cells

- Target Features: cell polygons
- Join Features: cell centroid points
- Join Operation: JOIN ONE TO ONE
- Match Option: INTERSECT or WITHIN
- Field Map of Join Features: delete everything
- Keep All Target Features: checked (doesn't matter in this case)

A quirk for Spatial Join

- Delete everything from the Field Map of Join Features field
  - Otherwise ArcGIS tends to randomly drop *fields* (ie. columns in the attributes table) from the original data
- Don’t try to drop fields in ArcGIS. Once read in Stata, drop unnecessary variables.

Exercise 2: Merge DHS clusters with ERA-40 grid points

Add to the model created in Model Builder for Exercise 1:

- First, convert DHS cluster XY data into a shapefile of point features (Lecture 1 Exercise 3)
- Then, use the Spatial Join tool.

Due to confidentiality requirements for data use, the DHS cluster data used in this exercise is modified by dropping cluster and survey identifiers.

Make XY Event Layer

- XY Table: dhs.txt
  (downloaded from T:/Economics)
- X Field: longnum
- Y Field: latnum
- Spatial Reference: WGS 1984

Then use Copy Feature to make it permanent data

Spatial Join

- Target Features: DHS cluster points
- Join Features: ERA-40 grid cell polygons
- Join Operation: JOIN ONE TO ONE
- Match Option: INTERSECT or WITHIN
- Keep All Target Features: checked

  (we may want to know which clusters cannot be matched with ERA-40 cells)

- Now we have a shapefile whose attribute table can be used in Stata to merge DHS data with ERA-40 data
- Annoyingly, ArcGIS cannot convert the file format for attribute tables (dBASE) into other formats that can be read in Stata.

How to read attribute tables in Stata

1. Open by Excel and save as an Excel spreadsheet
    - Since version 12, Stata can read Excel data directly
2. Use Stat Transfer to convert into Stata data format (.dta)
3. Set up ODBC. For Windows XP, see

www.ats.ucla.edu/stat/stata/faq/odbc.htm

- Stata can read .dbf files by the "odbc load" command
- This command does not work if the name of a dBASE file is in 6+ letters

$\Rightarrow$ Always name the output files with 5 letters or less

```
cd directoryname

odbc load, table("file.dbf") dsn("dBASE Files") lowercase clear
```

- First, change the directory to the one in which the dBASE file is saved
- Use quotation marks for the file name
- The lowercase option makes the variable names in lower case
    - ArcGIS makes all the variable names in upper case.

4 Python programing for ArcGIS

(Equivalent of writing Stata do files)

- Essential for replication
- Convenient to repeat the same geo-processing
  - Within a script by using a loop
  - When you obtain the updated version of the datasets

4.1 How to write Python scripts

- Use Model Builder to write a draft script
  - As we will see, it's impossible to write a script from scratch without making a mistake
- Then edit the draft script
  - Throughout the course, we will learn various scripting tips

Exercise 3: Export a Python script from the Model Builder

- (if the model was saved before and is not opened yet) Right-click the model and click “Edit” in Catalogue Window
- Click in the menu bar “Model > Export > To Python Script”
- Don’t forget to add “.py” at the end of the file name

Exercise 4: Read Python scripts in IDLE

First, open scripts in the IDLE (default text editor for Python scripts)

- Click Windows Start button.
- Then click “ArcGIS > Python 2.6 > IDLE (Python GUI)”
- Click “File > Open” to read scripts

- To edit the exported Python script, it's a good practice to copy and paste commands to your own template script.
- Here's my own template script ("template4L2.py" included in the downloaded data folder)

A template Python script

```
############################## Preamble ###################################
### Read the ArcGIS object ###
print "Launching ArcGIS 10"
import arcpy

### Set environment ###
print "Setting the environment"
# Allow the overwriting of the output files
arcpy.env.overwriteOutput = True # This command is CASE-SENSITIVE
# Set the working directory.
arcpy.env.workspace = "D:/temp" # NEVER USE single backslash (\).

### Local variables ###
////PASTE HERE//// # local variables from the exported script

############################## Geoprocessing ##########################
try:
    ////PASTE HERE//// # geoprocessing commands

except:
    print arcpy.GetMessages()

### Release the memory ###
print "Closing ArcGIS 10"
del arcpy
```

- Using this template script, we now learn the basics of Python scripting for ArcGIS

4.2 Essentials for Python scripting for ArcGIS

1. Object
2. String variable
3. Try-Except statement
4. Print
5. Comments

1. Object

- An *object* is a set of commands

```
### Read the ArcGIS object ###
import arcpy
```

e.g. "arcpy" is an object containing all the commands for ArcGIS data processing

- The "import" command reads this object (i.e. launch ArcGIS)

1. Object (cont.)

- An object may contain smaller objects

  e.g. “arcpy” contains an object called “env”

- Two types of commands:

  properties and methods

- A *property* contains a value (e.g. for environment settings)

  ex.1 the working directory name

  ex.2 whether the overwriting of output files is allowed or not

- To get a value of the property, type

object.property

- To assign a value to the property, type

object.property = value

```
### Read the ArcGIS object ###
import arcpy

### Set environment ###
# Allow the overwriting of the output files
arcpy.env.overwriteOutput = True # This command is CASE-SENSITIVE
# Set the working directory.
arcpy.env.workspace = "D:/temp" # NEVER USE single backslash (\).
```

An ArcGIS quirk

- To turn on the overwriting of output files, the correct command is

  arcpy.env.overwriteOutput = True
- Note that this is CASE-SENSITIVE (unbelievable...)

  cf. Python is case-sensitive in general

Pathname in Python

- Use \\
  e.g. C:\\TEMP\\lecture2
- / can be used, but it sometimes causes an error
- Never use \
  - In Python, \ is used for line continuation
  - In Windows, \ is used for pathname...

1. Object (cont.)

- Two types of commands: properties and methods
- A *method* is a function (ie. processing inputs to produce outputs)

  ex.1 geoprocessing tools

  ex.2 error message

- To use a method, type

  object.method(arg1, arg2, ...)

Example methods in exported script

```
# Process: Create Fishnet
arcpy.CreateFishnet_management(era40_shp, "-20.625 -40.625", "-20.625 -10.625", "1.25", "1.25", "65", "65",
"", "NO_LABELS", "DEFAULT", "POLYGON")

# Process: Make XY Event Layer
arcpy.MakeXYEventLayer_management(dhs_csv, "longnum", "latnum", dhs_Layer, "GEOGCS['GCS_WGS_1984',DATUM
['D_WGS_1984',SPHEROID['WGS_1984',6378137.0,298.257223563]],PRIMEM['Greenwich',0.0],UNIT['Degree',
0.0174532925199433]];-400 -400 1000000000;-100000 10000;-100000
10000;8.98315284119521E-09;0.001;0.001;IsHighPrecision", "")

# Process: Copy Features
arcpy.CopyFeatures_management(dhs_Layer, dhs_shp, "", "0", "0", "0")

# Process: Define Projection
arcpy.DefineProjection_management(era40_shp, "GEOGCS['GCS_WGS_1984',DATUM['D_WGS_1984',SPHEROID['WGS_1984',
6378137.0,298.257223563]],PRIMEM['Greenwich',0.0],UNIT['Degree',0.0174532925199433]]")

# Process: Feature To Point
arcpy.FeatureToPoint_management(era40_shp__2_, era40points_shp, "CENTROID")

# Process: Add XY Coordinates
arcpy.AddXY_management(era40points_shp)

# Process: Spatial Join
arcpy.SpatialJoin_analysis(era40_shp__2_, era40points_shp__2_, era40final_shp, "JOIN_ONE_TO_ONE",
"KEEP_ALL", "", "CONTAINS", "", "")

# Process: Spatial Join (2)
arcpy.SpatialJoin_analysis(dhs_shp, era40final_shp, dhsera40_shp, "JOIN_ONE_TO_ONE", "KEEP_ALL", "",
"WITHIN", "", "")
```

- As you notice, ArcGIS method names mix uppercase and lowercase letters
- Arguments for these methods are often very long with lots of [ ] ‘ ’ ; ,

$\Rightarrow$ Better to first use Model Builder to export the script

To find help for each geoprocessing method

- For more general help information, search in the ArcGIS Desktop Help on the web

  resources.arcgis.com/en/help/main/10.1/index.html

  - There's a built-in help, too, but the web version is more often updated

2. Try-Except statement

- If you simply run commands and get an error, you won't get a message on what went wrong
- So we need to use the try-except statement & the GetMessages method

- Try-Except statement:

  If any of the *indented* commands after “try:” causes an error,

  $\Rightarrow$ *indented* commands after “except:” will be executed

$\Rightarrow$ By inserting the GetMessages method after "except:”, you will get an error message when there is an error

Exercise 5: Use the Try-Except statement

- Copy all the geoprocessing commands in the exported script
- Paste them between “try:” and “except:” in the template script
- Indent all the pasted commands
  - Select commands to be indented
  - Click “Format > Indent Region”
  - Avoid using the tab key to do this.

```
############################## Geoprocessing ##########################
try:
    # Process: Create Fishnet
    arcpy.CreateFishnet_management(era40_shp, "-20.625 -40.625", "-20.625 -10.625", "1.25",
"1.25", "65", "65", "", "NO_LABELS", "DEFAULT", "POLYGON")

    # Process: Define Projection
    arcpy.DefineProjection_management(era40_shp, "GEOGCS['GCS_WGS_1984',DATUM
['D_WGS_1984',SPHEROID['WGS_1984',6378137.0,298.257223563]],PRIMEM['Greenwich',0.0],UNIT['Degree',
0.0174532925199433]]")

    # Process: Feature To Point
    arcpy.FeatureToPoint_management(era40_shp, era40points_shp, "CENTROID")

    # Process: Add XY Coordinates
    arcpy.AddXY_management(era40points_shp)

    # Process: Spatial Join
    arcpy.SpatialJoin_analysis(era40_shp, era40points_shp, era40final_shp, "JOIN_ONE_TO_ONE",
"KEEP_ALL", "", "CONTAINS", "", "")

    # Process: Make XY Event Layer
    arcpy.MakeXYEventLayer_management(dhs_csv, "longnum", "latnum", dhs_Layer, "GEOGCS
['GCS_WGS_1984',DATUM['D_WGS_1984',SPHEROID['WGS_1984',6378137.0,298.257223563]],PRIMEM
['Greenwich',0.0],UNIT['Degree',0.0174532925199433]];-400 -400 1000000000;-100000 10000;-100000
10000;8.98315284119521E-09;0.001;0.001;IsHighPrecision", "")

    # Process: Copy Features
    arcpy.CopyFeatures_management(dhs_Layer, dhs_shp, "", "0", "0", "0")

    # Process: Spatial Join (2)
    arcpy.SpatialJoin_analysis(dhs_shp, era40final_shp, dhsera40_shp, "JOIN_ONE_TO_ONE",
"KEEP_ALL", "", "WITHIN", "", "")

except:
    print arcpy.GetMessages()
```

3. String variable

- It's a good practice to give a local macro name to each data file
  - You may later want to use a different input file for the same data processing
  - It's tedious to search for which commands in a script needs to be edited to change the input file name
- For this purpose, we need to learn how to create a string variable in Python

3. String variable

- To create a variable called *file1* which contains a string *data.shp*, type

$$\text{file1} = \text{“data.shp”}$$

  - Stata's equivalent command: local

- Create a variable for each data file name at the beginning of the script

$\Rightarrow$ Easy to learn what files are used for inputs or created as outputs and intermediate files

Exercise 6: Edit file names

- Copy local variables in the exported script
- Paste them before “try:” in the template script
- Sort these names by inputs, intermediates, and outputs
- For intermediates and outputs, delete the directory path
  - We already set the working directory as the output directory

Exercise 6 (cont.)

- With file-overwriting tools (e.g. Define Projection, Add XY Coordinates), the Model Builder assigns multiple variables to the same file name
- Better delete these duplicates and correct geoprocessing command arguments accordingly

```
# Set the working directory.
arcpy.env.workspace = "D:/temp/Lecture2/outputs"

### Local variables ###
# Inputs
dhs_csv = "D:\\temp\\Lecture2\\dhs.txt" # DHS cluster XY data
# Intermediates
era40_shp = "era40.shp"
era40points_shp = "era40points.shp"
dhs_Layer = "dhs_Layer"
# Outputs
era40final_shp = "era40final.shp" # ERA-40 grid polygons with centroid coordinates
dhs_shp = "dhs.shp" #DHS cluster point features
dhsera40_shp = "dhs40.shp" # DHS cluster point features with ERA-40 grid point coordinate.
```

4. Print

- To show the message on the IDLE screen while running a script, type

print “message”

  - Useful to know which part of the script is being run now
  - Can be used to display the value that a method returns
  - Stata’s equivalent command: display

Exercise 7: Add print commands

- For each geoprocessing, use print command to display what is being processed

```
### Geoprocessing starts here ###
try:
    # Process: Create Fishnet
    print "Creating ERA-40 grid polygons: step 1"
    arcpy.CreateFishnet_management(era40_shp, "-20.625 -40.625", "-20.625 -10.625", "1.25", "1.25", "65", "65", "", "NO_LABELS", "DEFAULT",
"POLYGON")

    # Process: Define Projection
    print "Creating ERA-40 grid polygons: step 2"
    arcpy.DefineProjection_management(era40_shp, "GEOGCS['GCS_WGS_1984',DATUM['D_WGS_1984',SPHEROID['WGS_1984',6378137.0,298.257223563]],PRIMEM
['Greenwich',0.0],UNIT['Degree',0.0174532925199433]]")

    # Process: Feature To Point
    print "Creating ERA-40 grid polygons: step 3"
    arcpy.FeatureToPoint_management(era40_shp, era40points_shp, "CENTROID")

    # Process: Add XY Coordinates
    print "Creating ERA-40 grid polygons: step 4"
    arcpy.AddXY_management(era40points_shp)

    # Process: Spatial Join
    print "Creating ERA-40 grid polygons: step 5"
    arcpy.SpatialJoin_analysis(era40_shp, era40points_shp, era40final_shp, "JOIN_ONE_TO_ONE", "KEEP_ALL", "", "CONTAINS", "", "")

    # Process: Make XY Event Layer
    print "Creating DHS cluster point features: step 1"
    arcpy.MakeXYEventLayer_management(dhs_csv, "longnum", "latnum", dhs_Layer, "GEOGCS['GCS_WGS_1984',DATUM['D_WGS_1984',SPHEROID['WGS_1984',
6378137.0,298.257223563]],PRIMEM['Greenwich',0.0],UNIT['Degree',0.0174532925199433]];-400 -400 1000000000;-100000 10000;-100000
10000;8.98315284119521E-09;0.001;0.001;IsHighPrecision", "")

    # Process: Copy Features
    print "Creating DHS cluster point features: step 2"
    arcpy.CopyFeatures_management(dhs_Layer, dhs_shp, "", "0", "0", "0")

    # Process: Spatial Join (2)
    print "Spatially joining DHS clusters with ERA-40 grid points"
    arcpy.SpatialJoin_analysis(dhs_shp, era40final_shp, dhsera40_shp, "JOIN_ONE_TO_ONE", "KEEP_ALL", "", "WITHIN", "", "")

except:
    print arcpy.GetMessages()
```

5. Comments

- To insert comments, type # at the beginning of a comment
  - Useful for others to understand (and for you to remember) what each command does
  - Also useful if you want to skip some commands for the time being
  - Stata's equivalent: *

Exercise 8: Skip some commands temporarily

- Select commands to be skipped in the next run of the script
- Click “Format > Comment Out Region”
- ## will appear at the beginning of the command lines

Exercise 9: Run the script

- Save the script (File > Save / Save As)
- Click “Run > Run Module”
- Cross your fingers :-)
- Read output shapefiles in ArcMap (Lecture 1 Exercises 1-2) to see if everything worked out

- Probably your script crashes with an error message containing “cannot get exclusive schema lock”.
    - We get this error message thanks to the Try-Except statement in the script.
- This happens when the output shapefiles are shown in ArcMap (because we created these from the Model Builder).
- Before you run the script, close all the shapefiles in ArcMap.

If output files are not what you intended to create...

- Do your best to revise the script.
- Before you re-run the revised script, close all the output shapefiles you read in ArcMap. Otherwise the script crashes.
- A revised script very often crashes due to the previous output data files haunting somewhere in the computer.

- If you cannot find what's wrong with the script, quit ArcMap and IDLE (and even Windows). Relaunch and then run the script.
- If this still doesn't solve the problem, look up at http://forums.arcgis.com. Most likely someone else has encountered the same problem and been answered by those knowledgeable.

- The model Python script for this lecture is available as “lec2script.py” in the downloaded data folder for Lecture 2

5. What we've learned on ArcGIS

- Create a graticule shapefile
- Merge two vector datasets spatially

Do you remember which geoprocessing tools you used for each of these tasks?

Appendix 1: Weak instruments in more detail

With weak IVs, 2SLS is problematic:

- Inconsistency (Bound et al. 1995)
- Finite sample bias gets large, even in very large sample (Bound et al. 1995)
  - Biased towards OLS estimate
- S.E. too small

- These problems (bias / small S.E.) get worse with more instruments.
- But when just-identified (# of IVs = # of endogenous regressors)
  - When just-identified , 2SLS is "median-unbiased" (Angrist & Pischke 2009, p. 209)
  - S.E. is still fairly large so you won't end up rejecting the zero effect when it is zero.

How to deal with weak IV

Two issues

- How to tell if instruments are weak
  - Answer differs by # of endogenous regressors
- How to cope with consequences of weak instruments
  - Answer differs by just-identified or over-identified and by # of endogenous regressors

Diagnosis if only one endogenous regressor

- Look at F-stat on excluded instruments from the 1st-stage regression
  - For # of IV $\geq$ 3, Stock & Yogo (2005) provide threshold values below which bias in 2SLS is >10% of bias in OLS (which is around 10)
  - Staiger & Stock (1997) write "the F-stat rarely exceeds 10 in applications"

$\Rightarrow$ 10 becomes a rule of thumb (it seems)

- With heteroskedasticity, $F > 10$ may not be enough (Montiel Olea & Pflueger 2013)
- Stock & Yogo (2005) also provide the critical value for the actual size of 5% test being less than 10%

Diagnosis if 2+ endogenous regressors

- Use the Cragg-Donald statistics (see Stock, Wright, & Yogo 2002)
- Or follow Angrist & Pischke (2009: 218), if interested in bias of each individual coeff estimate

(both implemented by ivreg2 in Stata)

Solutions if just-identified

- No better point estimator available
- With only 1 endogenous regressor, Moreira (2003)'s CLR confidence intervals are valid
  - Andrews, Moreira, and Stock (2006) prove its optimality regardless of the strength of instrument
  - Stata ado: downloadable from Marcelo Moreira's website

Solutions if just-identified (cont.)

- With multiple endogenous regressors, it's an open question how to build valid confidence intervals (at least as of 2006, when Murray 2006 was written)

Solutions if over-identified

- Use LIML with s.e. correction by Bekker (1994) (recommended by Imbens 2007)
  - Flores-Lagunes (2007) shows LIML does as well as other proposed solutions
  - Available in Stata
- With only 1 endogenous regressor, Moreira's (2003) CLR confidence interval is nearly-optimal (proved by Andrews, Moreira, and Stock (2006) (cf. L4)

However...

- If $|\hat{\beta}_{OLS}| < |\hat{\beta}_{IV}|$, the possibility of $\beta = 0$ can be ruled out if 2SLS bias is smaller than OLS bias
- Relative bias of 2SLS to OLS is

$$\frac{l}{n\tilde{R}^2}$$

l	# of instruments
n	# of obs
$\tilde{R}^2$	R-squared in 1st stage
cf.	page 124 of Murray (2006)

Appendix 2: Clustering standard errors in detail

- Individual outcomes & group-level regressors
  - $\Rightarrow$ Cluster at group level (Moulton 1990)
- In panel data, both outcomes and regressors are serially correlated within the same unit
  - $\Rightarrow$ Cluster at unit level (Bertrand et al 2004)

- Always report # of clusters used for s.e. calculation
  - $\Leftarrow$ If small, s.e. is underestimated (Bertrand et al. 2004)
- In such cases with observational data, use wild bootstrap-t procedure by Cameron et al. (2008)
  - Create a distribution of t-values by making the residuals negative for randomly chosen half of clusters in each step of bootstrapping
  - Sample Stata do file: available at Douglas Miller's website

- For RCTs with a small # of clusters, randomization inference is a way to go (cf. L3)

References for Lecture 2

Andrews, Donald W. K., Marcelo J. Moreira, and James H. Stock. 2006. "Optimal Two-Sided Invariant Similar Tests for Instrumental Variables Regression." *Econometrica* 74 (3): 715-752.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Bekker, Paul A. 1994. "Alternative Approximations to the Distributions of Instrumental Variable Estimators." *Econometrica* 62(3): 657-681.

Bound, John, David A. Jaeger, and Regina M. Baker. 1995. "Problems with Instrumental Variables Estimation when the Correlation between the Instruments and the Endogenous Explanatory Variable is Weak." *Journal of American Statistical Association* 90: 443-450.

Brückner, Markus, and Antonio Ciccone. 2011. "Rain and the Democratic Window of Opportunity." *Econometrica* 79(3): 923–947.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2008. "Bootstrap-based Improvements for Inference with Clustered Errors." *Review of Economics & Statistics* 90(3): 414–427.

Currie, Janet, and Matthew Neidell. 2005. "Air Pollution and Infant Health: What Can We Learn From California's Recent Experience?" *Quarterly Journal of Economics* 120 (3): 1003-1030.

Deaton, Angus. 2010. "Instruments, Randomization, and Learning about Development." *Journal of Economic Literature* 48(2): 424-455.

Dell, Melissa, Benjamin F. Jones, and Benjamin A. Olken. 2009. "Temperature and Income: Reconciling New Cross-Sectional and Panel Estimates." *NBER Working Paper* 14680.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2007. "Using Randomization in Development Economics Research: A Toolkit." In *Handbook of Development Economics*, eds. T. Paul Schultz and John A. Strauss. Elsevier, p. 3895–3962.

Feyrer, James, and Bruce Sacerdote. 2009. "Colonialism and Modern Income: Islands as Natural Experiments." *Review of Economics and Statistics* 91(2): 245-262.

Flores-Lagunes, Alfonso. 2007. "Finite sample evidence of IV estimators under weak instruments." *Journal of Applied Econometrics* 22(3): 677-694.

Imbens, Guido W. 2007. "Weak Instruments and Many Instruments." http://

www.nber.org/WNE/lect_13_weakmany_iv.pdf

Imbens, Guido W. 2010. “Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009).” *Journal of Economic Literature* 48(2): 399-423.

Jayachandran, Seema. 2009. “Air Quality and Early-Life Mortality.” *Journal of Human Resources* 44(4): 916-954.

Kudamatsu, Masayuki. 2012. “Has Democratization Reduced Infant Mortlaity in Sub-Saharan Africa? Evidence from Micro Data.” *Journal of the European Economic Association* forthcoming.

Kudamatsu, Masayuki, Torsten Persson, and David Stromberg. 2012. “Weather and Infant Mortality in Africa.” CEPR Discussion Paper, no. 9222.

Miguel, Edward, Shanker Satyanath, and Ernest Sergenti. 2004. “Economic Shocks and Civil Conflict: An Instrumental Variables Approach.” *Journal of Political Economy* 112 (4): 725-753.

Montiel Olea, Jose Luis, and Carolin Pflueger. 2013. “A Robust Test for Weak Instruments.” Journal of Business and Economic Statistics, 31(3): 358-369.

Moreira, Marcelo J. 2003. “A Conditional Likelihood Ratio Test for Structural Models.” *Econometrica* 71(4): 1027-1048.

Moulton, Brent R. 1990. “An Illustration of a Pitfall in Estimating the Effects of Aggregate Variables on Micro Units.” *Review of Economics and Statistics* 72(2): 334–338.

Murray, Michael P. 2006. “Avoiding Invalid Instruments and Coping with Weak Instruments.” *Journal of Economic Perspectives* 20(4): 111-132.

Stock, James H, Jonathan H Wright, and Motohiro Yogo. 2002. “A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments.” *Journal of Business and Economic Statistics* 20(4): 518-529.

Stock, James H., and Motohiro Yogo. 2005. “Testing for Weak Instruments in IV Regression.” In *Identification and Inference for Econometric Models: A Festschrift in Honor of Thomas Rothenberg*, eds. Donald W. K. Andrews and James H Stock. Cambridge University Press, p. 80-108.

Tanser, Frank C, Brian Sharp, and David le Sueur. 2003. “Potential Effect of Climate Change on Malaria Transmission in Africa.” *Lancet* 362: 1792-1798.