Stockholm Doctoral Course Program in Economics

Topics in Applied Microeconometrics:
Using GIS

Lecture 7

# Spatial Regression Discontinuity Design

Masayuki Kudamatsu

IIES, Stockholm University

9 October, 2013

# What is spatial regression discontinuity design?

- Forcing variable: location (2-dimensional vector)
- Cutoff: boundary of treated areas
- With ArcGIS, easy to implement

# Examples

- Dell (2010)
- Michalopoulos & Papaioanno (2012)

# Outline

1. Dell (2010)
    - Review on RD design
2. Replicate spatial data for Dell (2010)

1. Dell (2010)

## 1.1 Research Questions

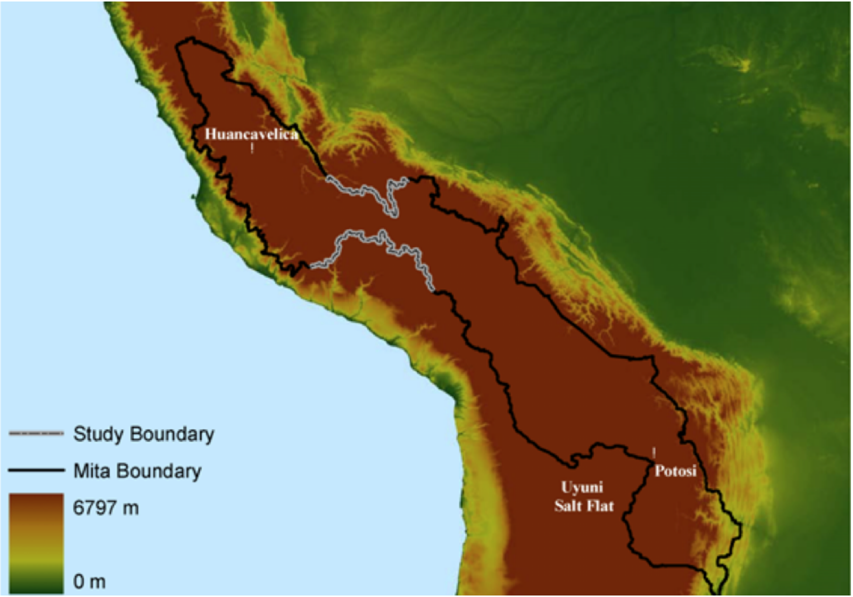
- Does the forced mining labor system during the Spanish colonial rule affect living standards in Peru today?
- If so, why?

# 1.1 Research Questions (cont.)

- Interesting?
  - Yes if the why question is answered
- Original?
  - Focus on the mechanism of the long-run effect of history
  - Spatial RD design
  - Findings against the well-known hypothesis by Engerman & Sokoloff
- Feasible?

## 1.2 Mita

- 1573-1812: Spanish colonial rule forced 1/7 of adult male population in parts of Peru & Bolivia to work in the Potosi silver & Huancavelica mercury mines
- Communities subject to mita: clearly defined geographically


Huancavelica
Potosi
Uyuni
Salt Flat
Study Boundary
Mita Boundary
6797 m
0 m

# 1.2 Mita (cont.)

- Determinants of mita assignments:
    - Short distance to mines
    - High elevation

# Digression: what causes treatment?

- In observational study, always investigate the reasons for treatment assignment
- If possible, argue that those factors are orthogonal to outcomes
- Otherwise, control for those factors in regression
  - If exogenous, use them in the baseline
  - Otherwise, use them for robustness checks

## 1.3 Data

For dependent variables:

- HH consumption: ENAHO 2001
- Heights of all 6-9 years old pupils: MoE census
    - School children data: useful in developing countries if enrollment ratio close to 100%
- Road network in Peru

  ⇐ We will calculate road lengths in ArcGIS later

## 1.3 Data (cont.)

For regressors:

- Which districts were assigned to mita
- Location of district capitals
- District boundaries
- Elevation: SRTM30 (cf. Lecture 6)

# 1.4 Empirical strategy

$$\begin{aligned} c_{idb} &= \gamma\, mita_d + \mathbf{X}'_{id}\beta \\ &\quad + f(location_d) + \Phi_b + \varepsilon_{idb} \end{aligned}$$

| | |
|---|---|
| $c_{idb}$ | Outcome for observation *i* in district *d* in segment *b* of mita boundary |
| $mita_d$ | Mita district indicator |
| $\mathbf{X}_{id}$ | Controls (e.g. district mean elevation/slope) |

# 1.4 Empirical strategy (cont.)

$$\begin{aligned} c_{idb} &= \gamma\, mita_d + \mathbf{X}'_{id}\beta \\ &\quad + f(location_d) + \Phi_b + \varepsilon_{idb} \end{aligned}$$

- $f(\cdot)$: cubic polynomials
- $location_d$:
  - A longitude & latitude
  - B distance to Potosi
  - C distance to mita boundary
- $\Phi_b$: Segment fixed effect

# 1.4 Empirical strategy (cont.)

- Adopts polynomial approach
  - Coordinates at the individual level: unavailable

  ⇒ Local linear regression approach: infeasible

  cf. Michalopoulos & Papaioanno (2012) have 0.125 x 0.125 degree cell observations (of night-time light). But still use polynomial approach (in distance to national borders by ethnic groups)
- Restrict the sample to w/i 50km from mita boundary

# 1.4 Empirical strategy (cont.)

- Forcing variable is 2-dimensional (longitude & latitude of district capital)
    - Data-demanding, though
    - Perhaps overfitting

$\Rightarrow$ Project the coordinate into 1-dimensional space to reduce the data requirements

- Distance to Potosi
    - This is not standard RD: no clear cut-off value
    - Economically most important, though
- Distance to mita boundary

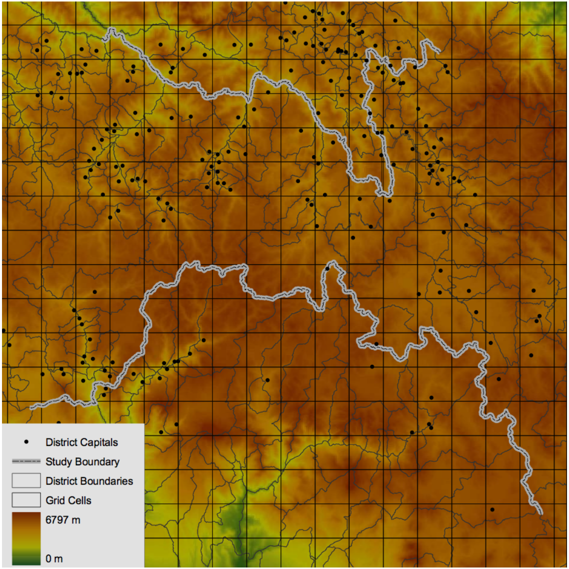
# 1.4 Empirical strategy (cont.)

- Project the coordinate into 1-dimensional space to reduce the data requirements

⇒ Boundary segment fixed effects: essential

  - Otherwise you will compare mita and non-mita districts with same distance to Potosi (or mita boundary) that are located far away from each other

# Figure A1



District Capitals
Study Boundary
District Boundaries
Grid Cells
6797 m
0 m

# 1.4 Empirical strategy (cont.)

- Choose cubic polynomial (somehow ad hoc, but perhaps due to feasibility)
    - Check other orders (1, 2, 4)
    - Allow polynomials to differ between mita and non-mita districts for 1st-/2nd-order polynomial in distance to Potosi / mita boundary
    - Results reported in Table III and Table A3

# 1.5.1 Results for today's living standards

- Mita districts: worse off
  - Household consumption 22-33% lower
  - Stunted children: 6-9%pt more (mean 40%)

## 1.5.2 Results for mechanisms

Mita districts historically:

- Lower % of rural population living in haciendas (large land-holding with attached labor force) in 1689, ca. 1845, & 1940 (Table VI)
  - ⇐ Spanish authority restricted the formation of haciendas
  - But long after mita was abolished in 1812, this difference persisted
- Less educated in 1876, 1940, & 2001 (Table VII)

# 1.5.2 Results for mechanisms (cont.)

Mita districts today:

- Lower density of regional / paved roads today (Table VIII)
- Less participation in market (ie. more subsistence farming) (Table IX)

These findings suggest:

- Landed elite in non-mita districts provided public goods (education & roads), which led to more economic prosperity than mita districts

$\Rightarrow$ This is totally opposite to influential Engerman-Sokoloff hypothesis on why Latin America lags behind North America

# 1.6 Validity check 1: No discontinuity for covariates?

- Elevation, Slope, Ethnicity (Table I)
  - We will see how ArcGIS helps create Table I
- 1572 tribute rate, % of 1572 tribute allocated to elite groups (Tables I & V)
- Population shares of adult men, boys, and women (Table V)

# 1.6 Validity check 2: Compliance?

- Out-migration from mita districts may have been substantial...
- Table IV columns 7 & 12: omit 4.8% of the non-mita sample from the upper tail of the outcome distribution
    - In-migration rate: higher by 4.8% for non-mita districts (according to 1993 census)

# Digression: Dealing with non-compliance

- Obtain in-migration rates between treated and control groups
- If treated (control) group has higher rate, trim the upper or lower tail of the outcome distribution in treated (control) group by the excessive in-migration rate
- Then re-estimate treatment effect

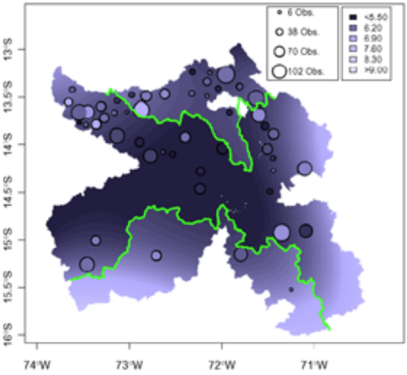
- Same can be done for out-migration.
- For Dell (2010), out-migration rates are comparable between treated and control groups
- As long as treatment induces out-migration monotonically, the mean outcome of emigrants should then be comparable

- See Manski (1990) for the basic idea
- Same approach can also be taken for differential attrition between treated and control groups (Lee 2009, who formalizes the approach)

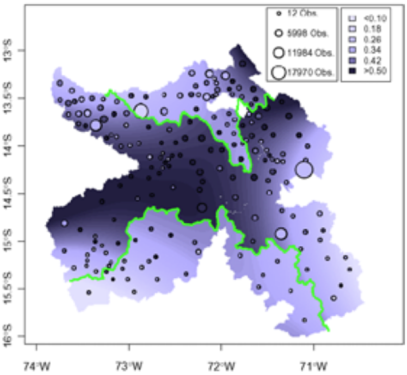
# 1.6 Validity check 3: jump at non-cutoff values

- Infeasible for longitude & latitude or distance to Potosi (where’s the median?)
- Could be done with distance to mita boundary

# 1.7 Graphical representation of the main results



(a) Consumption (2001)



(b) Stunting (2005)

## 1.8 External validity

Observations around the cutoff may be exceptions...

- OLS with sample BETWEEN 25-100 km from mita boundary
- Similar estimates to the RD
- See page 1887

- Download T:/Economics/Lecture7
- This folder contains:
  - mita.gdb (replication spatial data provided by Dell 2010)
  - ***.py (replication Python scripts provided by Dell 2010)
  - data4exercises folder (data for this lecture)
- Unzip three zip files in data4exercise folder
- Launch ArcMap

# 2. Replicate spatial data for Dell (2010)

1. Geodatabase
2. Main data
3. Table I
4. Figure II
5. Understand the replication Python script

# 2.1 Geodatabase

- We will learn how to create some of the spatial data files provided by Dell (2010) as replication data
- These files are saved in a *file geodatabase* called “mita.gdb”
- What is a file geodatabase?

- A file geodatabase is a folder in which vector and raster data files can be saved
- No need to add ".shp" for vector data file names.
- To distribute your spatial data, this seems convenient
- To export a dBASE table, you need to use Table Select, however.
- There must be other benefits, but I'm not sure.

# How to create a file geodatabase

- In the Catalogue window, right-click the folder in which you will create a file geodatabase
- Click “New > File Geodatabase”, to create an empty geodatabase.

# 2.2 Replicate main data

The district-level data on outcomes, the mita indicator, and the coordinates of district capitals are directly obtained from data sources. But we need to construct:

a. Distance from each district capital to mita boundary (Exercise 1)
    - polynomials & sample selection

b. Coordinate of the nearest point on the mita boundary for each district capital (Exercise 1)
    - Boundary segment FE

# 2.2 Replicate main data (cont.)

c. Mean area-weighted elevation/slope by district (Exercise 2)
    - Control variables

d. Length of roads by type (Exercise 3)
    - Intermediate outcome variable

## 2.2.1 Projection

- So we need distance, area, and length of roads
- What projection should we use?

Dell (2010) uses (see Appendix page A-3)

- UTM projection for zone 18S for area and length of roads
    - Study area spans about 500km east-west (ie. less than 5°)
    - Sensible choice
- Equidistant Cylindrical (73° W as central meridian & 13° S as standard parallel) for distance

# Equidistant Cylindrical

- Earth is projected to a cylinder tangent on the chosen standard parallel
  - If equator is chosen, it’s called Plate Carree
- No distortion in distance (only) along the standard parallel
- East-west distance expands more, the farther away from standard parallel

⇒ Don’t be fooled by the name of projection “equidistant”

The more appropriate way to calculate distance will be either

- Obtain coordinates of the points and use Stata globdist
  - Distance to Potosi (19.58° S): shorter by 6.9-16.3km
- Use UTM projection

Remember no projection preserves distance in every direction from every point

# Exercise 1

Create mita boundary polylines (“MitaBoundary” in mita.gdb) from the raw data (see Appendix p. A-3)

- District polygons (StudyDistricts.shp)
- List of districts with the mita indicator (districts.txt)

in the data4exercises folder, to obtain

a. distance to the boundary
b. nearest point on the boundary

# Check the coordinate system

- StudyDistricts.shp is already projected in UTM Zone 18S
- Right-click the data and click Properties..
- Click the Source tab.

# UTM projection parameters

- Central meridian: different by zone

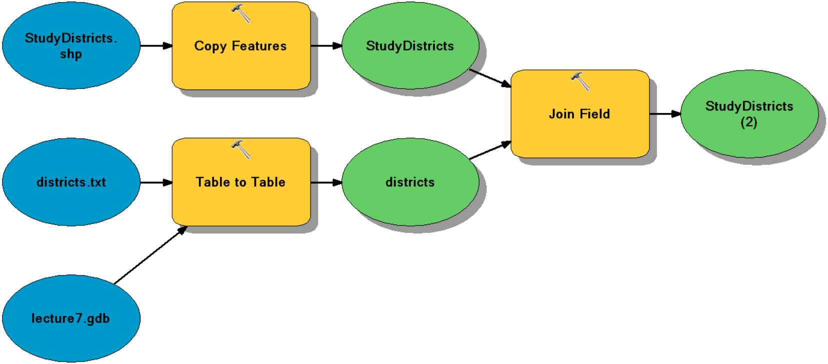
  (75 degree West for Zone 18S)
- False northing: 10,000,000 meters if southern hemisphere
  - Y coordinate on equator
- False easting: 500,000 meters
  - X coordinate on central meridian
- Scale factor: 0.9996
  - Length on central meridian is smaller by this factor than the real length

  ⇒ Minimize overall distance/area distortion within 6° longitude span

# Attach mita indicator

We first use Join Field (cf. Lec. 6)

- Copy Features to make a copy of the district polygons
  - ⇐ Join Field overwrites the input
- Table to Table to convert the mita district list in the Ascii format to a dBASE table
  - ⇐ Join Field works only with dBASE table...
- Join Field to attach the mita district indicator to district polygons


StudyDistricts.shp
Copy Features
StudyDistricts
districts.txt
Table to Table
districts
lecture7.gdb
Join Field
StudyDistricts (2)

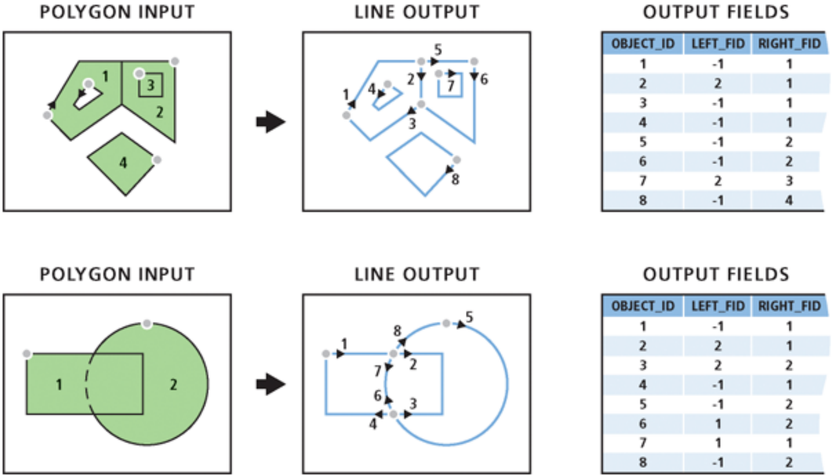
Then create boundary polylines:

- Dissolve by mita district indicator
- Polygon To Line to convert mita polygons to boundary polylines
- Select polylines that divide the study area

# Polygon To Line

- Check “Identify and store polygon neighboring information”
  - This creates two fields: LEFT_FID and RIGHT_FID
  - If a polyline is the edge of one polygon in the input data, LEFT_FID is -1

  ⇒ Allow us to identify which polylines are the boundary between two polygons in the input data



| OBJECT_ID | LEFT_FID | RIGHT_FID |
|---|---|---|
| 1 | -1 | 1 |
| 2 | 2 | 1 |
| 3 | -1 | 1 |
| 4 | -1 | 1 |
| 5 | -1 | 2 |
| 6 | -1 | 2 |
| 7 | 2 | 3 |
| 8 | -1 | 4 |

| OBJECT_ID | LEFT_FID | RIGHT_FID |
|---|---|---|
| 1 | -1 | 1 |
| 2 | 2 | 1 |
| 3 | 2 | 2 |
| 4 | -1 | 1 |
| 5 | -1 | 2 |
| 6 | 1 | 2 |
| 7 | 1 | 1 |
| 8 | -1 | 2 |

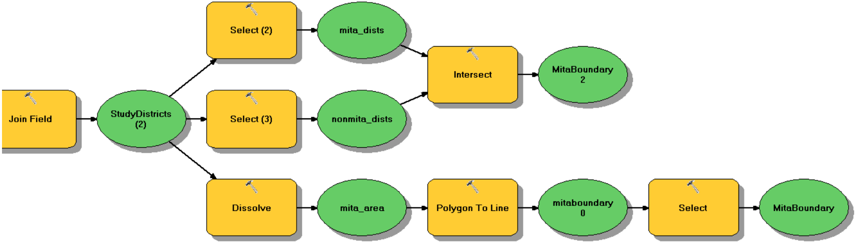
(Taken from Desktop Help for Polygon To Line)

# Select

- Expression: “LEFT_FID” <> -1

# Exercise 1 (cont.)

- An alternative is to use Select for creating mita zone polygon and non-mita zone polygon separately
- And then use Intersect with the Output Type option being LINE
  - If INPUT is chosen, the output file will be an empty polygon feature class


Select (2)
mita_dists
Intersect
MitaBoundary 2
Join Field
StudyDistricts (2)
Select (3)
nonmita_dists
Dissolve
mita_area
Polygon To Line
mitaboundary 0
Select
MitaBoundary

# Do you remember...

What tools should be used for obtaining the following two variables?

a. Distance from each district capital to mita boundary
b. Coordinate of the nearest point on the mita boundary for each district capital

Assume we have the coordinates of district capitals (see Appendix p. A-3)

# Answer: the Near tool

- Make XY Event Layer and Copy Features to create district capital point features (cf. lecture 1)
- Project district capital points and mita boundary polylines to UTM Zone 18S (cf. lecture 3)
  - Mita boundary is already in UTM Zone 18S
- Then use Near with the option LOCATION ticked (cf. lecture 4)

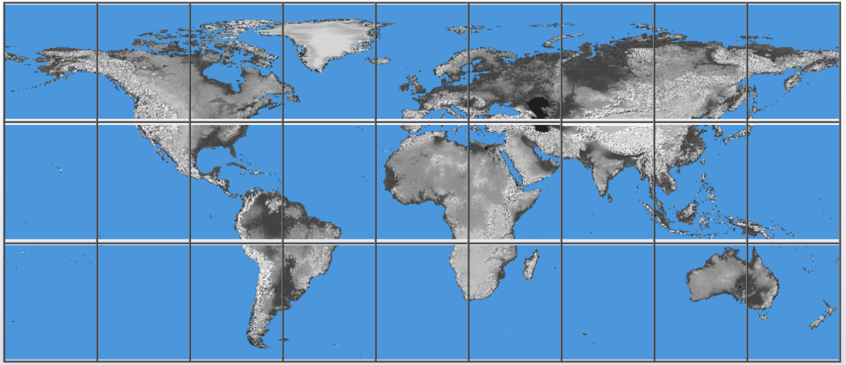
- One of the fields created by the Near tool is NEAR_DIST
- This contains the distance from each feature to the near feature
- Since input features are projected to UTM, the values in this field are proper distances in meters and can be used for analysis.

# Exercise 2

Obtain each district's mean elevation/slope from:

1. District boundary polygons in UTM Zone 18S (StudyDistricts.shp)
2. SRTM30 in WGS 1984 (w100n10c)

# SRTM30 raster tiles



Source: www.dgadv.com/srtm30/

If the study area spans two tiles or more, use Mosaic To New Raster to combine them

# Do you remember...

- What tool should be used for obtaining district-mean elevation, slope, or whatever raster data value?

# Do you remember...

- What tool should be used for obtaining district-mean elevation, slope, or whatever raster data value?
- Yes, it’s Zonal Statistics as Table

# Do you remember...

- What tool should be used for obtaining district-mean elevation, slope, or whatever raster data value?
- Yes, it’s Zonal Statistics as Table
- But the coordinate system of elevation raster is different between that of district polygons.

⇒ Use Project Raster

# Project Raster

- Coordinate system: UTM Zone 18S
- Cell size: 1000m
  - $\Leftarrow$ 30 arc-sec $\approx$ 1km
- Resampling: Bilinear / Cubic
  - $\Leftarrow$ Input raster values are not integer

# Project Raster (cont.)

- Projected raster cells $\neq$ raster cells in geographic coordinate system

$\Rightarrow$ Need to interpolate

  - Bilinear: distance-weighted average of surrounding cell values
  - Cubic: fitting a smooth curve through surrounding cells
  - Nearest: take the value from the nearest cell (good for discrete raster values; quick)
  - Majority: take the majority value from surrounding cells (good for discrete raster values)

- Now Dell (2010) calculate the mean elevation/slope over the non-water areas of a district

⇒ We will clip the elevation raster by land mass of Peru (“peru_nw” in mita.gdb)

- To create “peru_nw”, use Erase to clip out water bodies from a Peru country boundary polygon

# Clip out water bodies from country polygons

- For water body polygons, we can use WWF's Global Lakes and Wetland Database (GLWD) (worldwildlife.org/pages/global-lakes-and-wetlands-database)

    - An alternative is SRTM Water Body Data (SWBD) (dds.cr.usgs.gov/srtm/version2_1/SWBD/)

- Select Peru from Natural Earth country boundary shapefile (ne_10m_admin_0_countries.shp)
- Erase GLWD Level 1 (lakes of area >50 km$^2$) and Level 2 (0.1-50 km$^2$)
- Project to UTM Zone 18S

⇒ Create “peru_nw”

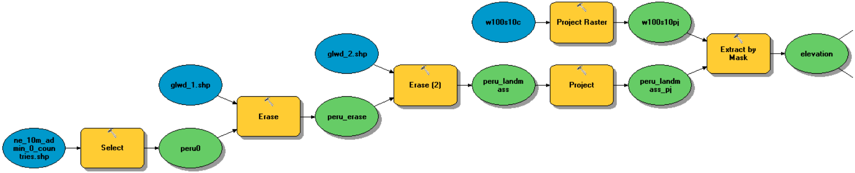
- Then we can clip elevation raster by Peru land mass polygon by...

# Extract By Mask

Throw away water areas (& areas outside Peru) from elevation raster

- Input raster: the output from Project Raster
- Input raster or feature mask data: peru_nw

($\Rightarrow$ creates “el_pe” in mita.gdb)


ne_10m_ad min_0_coun tries.shp
Select
peru0
glwd_1.shp
Erase
peru_erase
glwd_2.shp
Erase (2)
peru_landm ass
Project
peru_landm ass_pj
w100s10c
Project Raster
w100s10pj
Extract by Mask
elevation

# Calculating slope...

(cf. Lecture 6)

- Slope to obtain slopes
  - Output measurement: DEGREE
  - Z factor: 1
    (Elevation data is already projected to UTM, thus x-y units are meters, the same as z unit)

⇒ Creates “slope_nw” in mita.gdb

# Calculating mean elevation and slope...

(cf. Lecture 5)

- Zonal Statistics as Table to obtain mean elevation and slope
    - Zone Data: StudyDistricts.shp
        - It has the unique integer ID field (dist_id)
    - Value Raster: “el_pe” for elevation, “slope_nw” for slope
    - Statistics Type: MEAN

# Exercise 3

Calculate road length by district from

- Road network polylines (“roads_pj” in mita.gdb)
- District boundary polygons (StudyDistricts.shp)

# Calculating length of polylines

- Usually, Add Field and Calculate Field is enough for obtaining length of polylines (cf. Lec 6)
- But Peru is a hilly country. Length of roads should take elevation into account

⇒ Use Add Surface Information with elevation data as an additional input

- We will obtain surface length by road type.
- So we first need to use...

# Select

- Input features: roads_prj
- Expression
  - Local: "RUTA" = ’Vecinal’ AND "SUPERFICIE" >= 1 AND "SUPERFICIE" <= 4
  - Regional: "RUTA" = ’Departamental’ AND "SUPERFICIE" >= 1 AND "SUPERFICIE" <= 4
  - Paved/Gravel Regional: "RUTA" = ’Departamental’ AND ("SUPERFICIE" = 1 OR "SUPERFICIE" = 2)

(SUPERFICIE is 5 if under construction;

9 if planned)

Then we need to divide roads into different districts by ...

Intersect (cf. Lecture 4)

- Input features: StudyDistricts.shp and the output from the Select tool above

Dissolve

- Dissolve Field: dist_id
  ⇐ We need the total length of roads by district

Now we’re ready for calculating length

# Add Surface Information

(For the previous versions of ArcGIS, it was called

Surface Length)

- To calculate the length of polyline features with elevation taken into account
- If cannot be launched, click Tool > Extensions in the menu bar & check 3D analyst.

# Add Surface Information (cont.)

- Input Surface: “elev_pe” in mita.gdb
- Input Feature Class: the output from Dissolve tool above
- Check “Surface Length”
- Z factor: 1 ($\Leftarrow$ elevation data was projected into UTM and thus in meters in x-y dimensions)

Note: This tool overwrites the input feature class

- Then use Table Select, to export a dBASE table.

# 2.3 Replicate Table I

What we need is a table of 20x20 km grid cells with columns

a. Centroid coordinates
    - to calculate Conley (1999) standard errors (cf. Lecture 3)

b. Distance to mita boundary
    - For sample selection

c. Mean elevation/slope
    - Outcomes to compare

d. Mita area indicator
    - Treatment dummy

# 2.3a Centroid coordinates

- Create Fishnet cf. Lectures 2 & 5, but now in UTM projection
  - Template Extent: peru_nw
  - Cell width/height: 20000 (meters)
- Define Projection to UTM Zone 18S
- Add Field (make sure to check “Field Is Nullable”) and Calculate Field (use !OID! instead of !FID!+1) to assign unique positive integer ID

⇒ Create “grid”

# 2.3a Centroid coordinates (cont.)

(cf. Lectures 2, 4)

- Feature To Point to obtain grid cell centroids ($\Rightarrow$ “cent”)
- Add XY Coordinates to obtain centroid coordinates
- Clip (Analysis) by district boundary polygons

# Clip (Analysis) tool (exercise 4)

- Input features: cent
- Clip features: StudyDistricts
  - Dell (2010) uses the dissolved study district polygon (see geogrid.py), but you don't have to.

* Don't confuse it w/ Clip (Data Management) tool, which clips a raster data by a rectangle that you specify

# Digression: An ArcGIS tip

To drop spatial data that's outside the study area,

- First, obtain the study area polygon(s)
- Then use
  - Extract By Mask tool for raster data (cf. Exercise 2 above)
  - Clip (Analysis) tool for vector data

## 2.3b Distance to Mita boundary

Same as before. The input features for Near are now cell centroid points instead of district capital points

## 2.3c Mean elevation/slope

Same as before. The input zone features for <u>Zonal Statistics as Table</u> are now cell polygons instead of district polygons

# 2.3d Mita area indicator

- Intersect grid cell centroid points w/ Mita district polygons (“in_prj”)
- Intersect grid cell centroid points w/ non-Mita district polygons (“out_prj”)

## 2.4 Replicate Figure 2

- You need to learn programming with R...

# 2.5 Reading a Python script

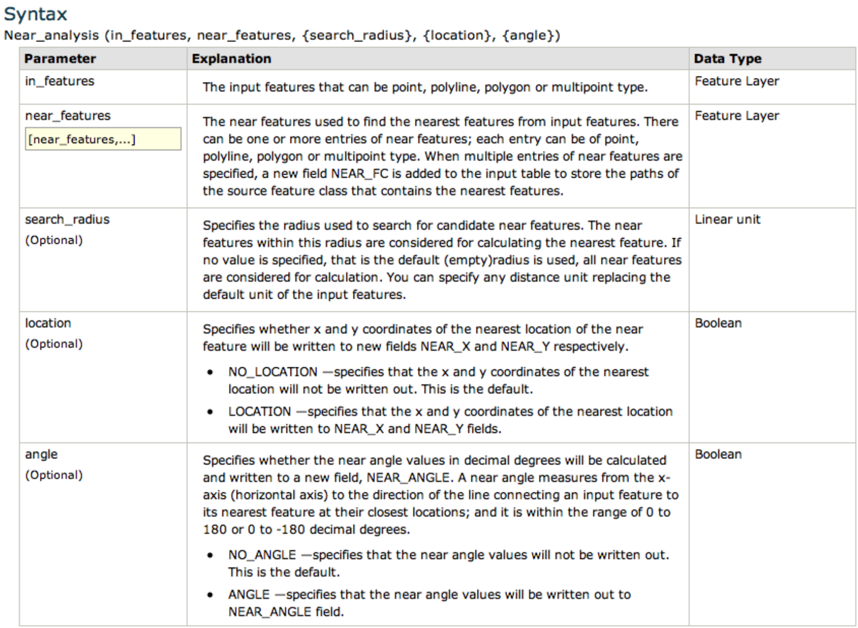
- Open gis_prep.py
- This script was written before ArcGIS 10 became available
- Instead of the arcpy object, it uses arcgisscripting object
- Slight difference, but the syntax is more or less the same

- To understand the syntax for each geoprocessing tool, look up ArcGIS Desktop Help.
- For example, gis_prep.py contains the following command:

```
Near_analysis(d2bnd, ph_bnd, “100000 Meters”,
“LOCATION”, “NO_ANGLE”)
```

- ArcGIS Desktop Help describes which arguments refer to what:

# Desktop Help for Near



## Syntax

Near_analysis (in_features, near_features, {search_radius}, {location}, {angle})

| Parameter | Explanation | Data Type |
|---|---|---|
| in_features | The input features that can be point, polyline, polygon or multipoint type. | Feature Layer |
| near_features<br>[near_features,...] | The near features used to find the nearest features from input features. There can be one or more entries of near features; each entry can be of point, polyline, polygon or multipoint type. When multiple entries of near features are specified, a new field NEAR_FC is added to the input table to store the paths of the source feature class that contains the nearest features. | Feature Layer |
| search_radius<br>(Optional) | Specifies the radius used to search for candidate near features. The near features within this radius are considered for calculating the nearest feature. If no value is specified, that is the default (empty)radius is used, all near features are considered for calculation. You can specify any distance unit replacing the default unit of the input features. | Linear unit |
| location<br>(Optional) | Specifies whether x and y coordinates of the nearest location of the near feature will be written to new fields NEAR_X and NEAR_Y respectively.<br>• NO_LOCATION —specifies that the x and y coordinates of the nearest location will not be written out. This is the default.<br>• LOCATION —specifies that the x and y coordinates of the nearest location will be written to NEAR_X and NEAR_Y fields. | Boolean |
| angle<br>(Optional) | Specifies whether the near angle values in decimal degrees will be calculated and written to a new field, NEAR_ANGLE. A near angle measures from the x-axis (horizontal axis) to the direction of the line connecting an input feature to its nearest feature at their closest locations; and it is within the range of 0 to 180 or 0 to -180 decimal degrees.<br>• NO_ANGLE —specifies that the near angle values will not be written out. This is the default.<br>• ANGLE —specifies that the near angle values will be written out to NEAR_ANGLE field. | Boolean |

# 3. What we’ve learned for ArcGIS

- File geodatabase
- Attach a new variable to shapefile attribute table
- Create boundary polylines
- Project raster data
- Clip vector and raster datasets
- Calculate length of polylines on a hilly area
- Create fishnet in UTM projection

# Appendix: Review of geoprocessing tools

The tools we learned could be categorized into five groups:

1. Convert into ArcGIS formats
2. Projection
3. Create new spatial data
4. Merge two spatial datasets
5. Add variables to spatial data

1-3: edit inputs; 4-5: create outputs

# 1 Convert into ArcGIS formats

XY data (Lec 1)

- Make New XY Layer & Copy Features

Raster in non-grid format (Lec 1)

- Ascii To Raster
- Copy Raster / Raster To Other Formats

# 1 Convert into ArcGIS formats (cont.)

Ascii tables for Join Field (Lec 7)

- Table To Table

# 2 Projection

- Define Projection (Lec 1)
    - if projection unassigned
- Project (Lec 1)
    - to change projection for vector data
- Project Raster (Lec 7)
    - to change projection for raster data

# What projections to use

- UTM (Lec 3,6,7)
  - length of line features
  - distance / area in small regions (within 6 degrees in longitude)
- Sinusoidal or any other equal area projections (Lec 4)
  - Area of large regions

# 3 Create new spatial data

a. Grid cell polygon features (Lec 2,5,7)
  - Create Fishnet (& Define Projection)

b. Centroid point features (Lec 4)
  - Feature To Point
  - Add XY Coordinates
    - For using Stata's globdist or Conley (1999) standard errors

c. Neighborhood polygon feature (Lec 3)
  - Buffer

# 3 Create new spatial data (cont.)

d. Intersections of features
   - Intersect (Lec 4)
   - Union (if non-intersected inputs need to be kept) (Lec 5)

e. Grouping features into one feature
   - Dissolve (Lec 5,6)

f. A subset of features
   - Select (by attributes) (Lec 5,6)
   - Clip (Analysis) (by extent of other features) (Lec 7)
   - Erase (the negative of Clip) (Lec 7)

# 3 Create new spatial data (cont.)

g. Boundary polylines (Lec 7)
    - Polygon To Line + Select

h. Raster cell centroid point features (Lec 5)
    - Raster To Point

# 3 Create new spatial data (cont.)

i. New raster data

- Reclassify (Lec 6)
  - To create a dummy variable
- Slope (Lec 6)
  - Could be used for non-elevation raster data
- Extract By Mask (Lec 7)
  - For sample selection
- Mosaic to New Raster (Lec 7)
  - To append raster tiles

# 4 Merge two spatial datasets

- Spatial Join (Lec 2,3,5,6)
  - features with features
- Extract Values To Point (Lec 5)
  - point features with raster
  - Combined w/ Raster To Point, can be used to merge raster with raster
- Zonal Statistics as Table (Lec 5,6,7)
  - polygon/polyline features with raster

# 5 Add variables to spatial data

a. Add Field & Calculate Field

- !FID!+1 (or !OID! if using file geodatabase) for unique identifier necessary for Zonal Statistics as Table (Lec 5)
- float(!shape.area!) for polygon areas (Lec 4)
- float(!shape.length!) for polyline lengths (Lec 6)

b. Add Surface Information

- Length of polyline features w/ elevation taken into account (Lec 7)

# 5 Add variables to spatial data (cont.)

c. Near (Lec 4,7)
   - Distance from point features to features
   - XY coordinates of the nearest point on polyline/polygon features from point features (with the LOCATION option on)

d. Join Field (Lec 6,7)
   e.g. To add the zonal statistics table to the polygon/polyline features

# 5 Add variables to spatial data (cont.)

- These tools overwrite the attribute table of the input shapefile

⇒ Use the Copy Features tool if the input shapefile is original.

# Appendix: RD design (1/5)

- Ideally, local linear regressions preferred (Imbens & Lemieux 2008)

  $$y = \alpha + \tau D + \beta_l (X - c) + \beta_r D (X - c) + \varepsilon$$

  with the sample restricted to
  $X \in [c - h, c + h]$

  - See Section 4.3.1 of Lee & Lemieux (2010) for choosing *h* optimally

- But requires lots of data around *c* for precisely estimating $\tau$

# Appendix: RD design (2/5)

- An alternative: Controlling for polynomials in *X*
- Drawbacks:
  - Sensitive to observations far away from *c*
  - Sensitive to the polynomial orders
- Solutions:
  - Restrict the sample to those close to *c*
  - Check robustness to different orders of polynomials

# Appendix: RD design 3/5

- Other recommended practices for polynomial approach (section 4.3.2 of Lee & Lemieux (2010))
  - Allow polynomials to differ btw. both sides of the cutoff
  - To pick the optimal order of polynomials...
    - Use Akaike Information Criterion
    - Add dummies for bins in *X* as regressors & test for joint significance; increase the order until becoming insignificant

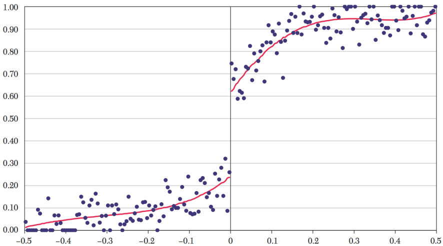
# Appendix: RD design 4/5

Validity checks (Imbens & Lemieux 2008, p. 633):

- See if covariates (incl. pre-treatment outcomes) jump at the cutoff
  - cf. Table 1 in RCTs
- See if density distribution jumps at the cutoff
  - cf. Compliance check in RCTs
- See if no jump in outcome at other values of *X* (e.g. median for either side of the cutoff)

# Appendix: RD design 5/5

- Graph the binned local averages of the outcome
- Superimpose the predicted values from a benchmark polynomial specification



Dell, Melissa. 2010. “The Persistent Effects of Peru's Mining Mita.” *Econometrica* 78(6): 1863-1903.

Imbens, Guido W., and Thomas Lemieux. 2008. “Regression discontinuity designs: A guide to practice.” *Journal of Econometrics* 142(2): 615-635.

Lee, David S. 2009. “Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects.” The Review of Economic Studies 76(3): 1071 -1102.

Lee, David S, and Thomas Lemieux. 2010. “Regression Discontinuity Designs in Economics.” *Journal of Economic Literature* 48(2): 281-355.

Manski, Charles F. 1990. “Nonparametric Bounds on Treatment Effects.” American Economic Review 80(2): 319-323.

Michalopoulos, Stelios, and Elias Papaioannou. 2012. “National Institutions and Subnational Development in Africa." Quarterly Journal of Economics, forthcoming. (NBER Working Paper no.18275.)