Stockholm Doctoral Course Program in Economics

Topics in Applied Microeconometrics: Using GIS

Lecture 3

Buffer

Masayuki Kudamatsu

IIES, Stockholm University

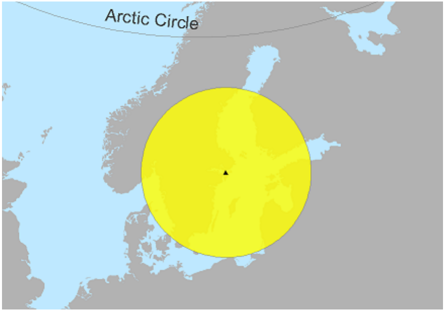
11 September, 2013

What does the *Buffer* tool do?

Create a polygon of the geographic neighborhood of an input feature

(i) For a point, create a circle of *x*-meter/feet radius

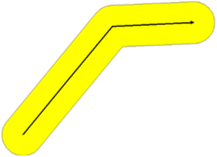
e.g. 500km buffer around Stockholm



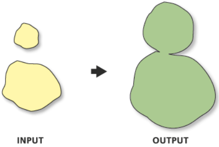
What does the *Buffer* tool do?

Create a polygon of the geographic neighborhood of an input feature

(ii) For a polyline...



(iii) For polygons...



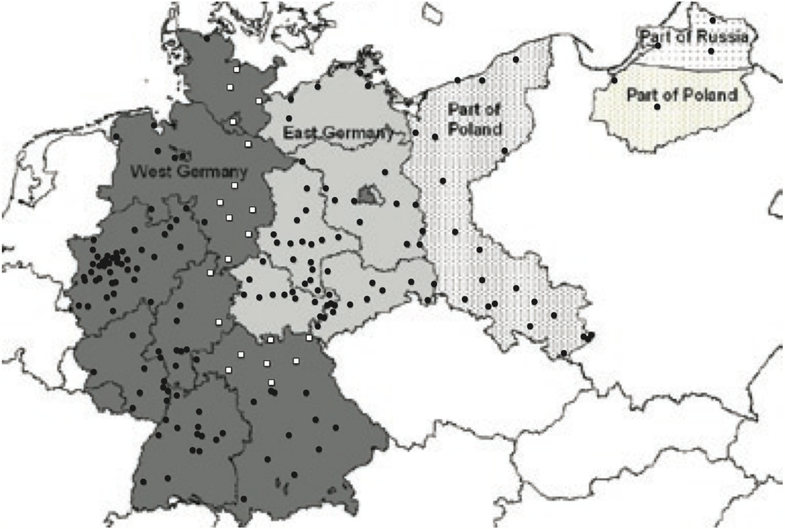
- Combined w/ the Spatial Join tool (cf. Lecture 2), the Buffer tool helps identify other observations in the neighborhood of each observation

The Buffer tool is econometrically useful for (at least) three reasons

(1) Estimate the spatial externality of a treatment
    - Miguel & Kremer (2004)
    - de Mel et al. (2008)

(2) Mitigate omitted variable bias in the estimation of peer effects
    - Conley & Udry (2010)

(3) Generate a treatment variable
    - Redding & Sturm (2008)



MAP 1. THE DIVISION OF GERMANY AFTER THE SECOND WORLD WAR

Taken from Map 1 of Redding & Sturm (2008)

Outline

1. Miguel & Kremer (2004)

2. Conley & Udry (2010)

3. UTM projections

4. Using the Buffer tool in ArcGIS

5. Loop over files in Python

1. Miguel & Kremer (2004)

- Estimating the spill-over effect of randomly assigned treatment

1.1 Research Questions

1. Does deworming school children reduce the infection for untreated children?
    - Interesting?
        - If yes, cost-benefit analysis calculation changes
    - Original?
        - No experimental evidence on the spill-over effects
    - Feasible?
        - Need to identify the surrounding area of treated schools

      ⇒ The Buffer tool helps.

1.1 Research Questions (cont.)

2. Does deworming increase school attendance?
    - Interesting?
        - If yes, very cost-effective intervention to increase attendance
        - Indeed, they find the cost is USD3.50 per 1 more year of schooling
    - Original?
        - No experimental evidence on health-education relationship
    - Feasible?
        - Health interventions took place at schools

1.2 Experiment Design

- 75 primary schools in western Kenya
  - Nearly all schools in the area
  - Over 30,000 pupils aged 6 to 18
- Randomly assigned to 3 groups
  - Group 1: Treated since 1998
  - Group 2: Treated since 1999
  - Group 3: Control (treated since 2001)

  by “alphabetization” (see Deaton (2010, p. 446) for critique) after stratified by admin zones & other NGO involvements

1.2 Experiment Design (cont.)

A bundle of treatments:

(1) Every six month or annually, deworm all boys & girls under 13 at school

- Those absent on the day of treatment: not treated

(2) Worm prevention education

- Wash hands, wear shoes, do not swim in infected water

Digression: Treatment bundle

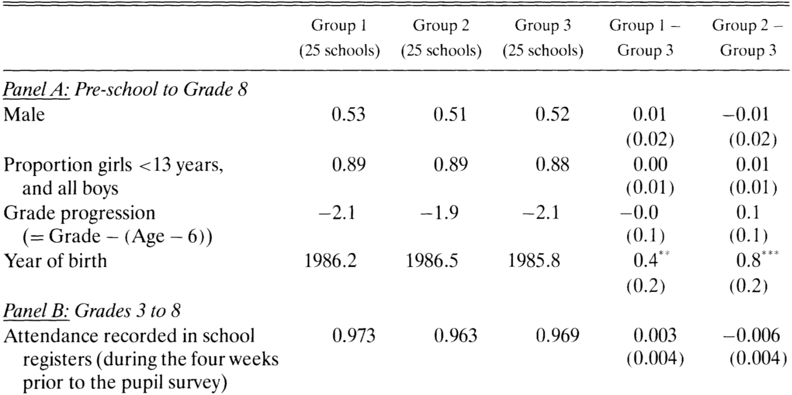
- Treatment often consists of several components
- Make sure to show which component drives the results
    - In Miguel & Kremer (2004), no change in prevention behavior (Table V Panel C)

  ⇒ Any effect should be via deworming, not health education

1.3 Empirical pre-analysis (1)

Check if pre-treatment outcomes (& covariates) are balanced across groups

- Table I



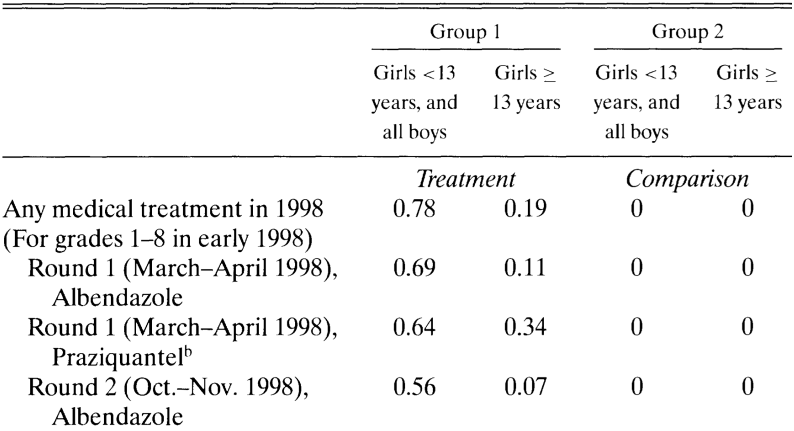
	Group 1 (25 schools)	Group 2 (25 schools)	Group 3 (25 schools)	Group 1 − Group 3	Group 2 − Group 3
Panel A: Pre-school to Grade 8					
Male	0.53	0.51	0.52	0.01 (0.02)	−0.01 (0.02)
Proportion girls <13 years, and all boys	0.89	0.89	0.88	0.00 (0.01)	0.01 (0.01)
Grade progression (= Grade − (Age − 6))	−2.1	−1.9	−2.1	−0.0 (0.1)	0.1 (0.1)
Year of birth	1986.2	1986.5	1985.8	0.4^{**} (0.2)	0.8^{***} (0.2)
Panel B: Grades 3 to 8					
Attendance recorded in school registers (during the four weeks prior to the pupil survey)	0.973	0.963	0.969	0.003 (0.004)	−0.006 (0.004)

Taken from Table I of Miguel & Kremer (2004)

1.3 Empirical pre-analysis (2)

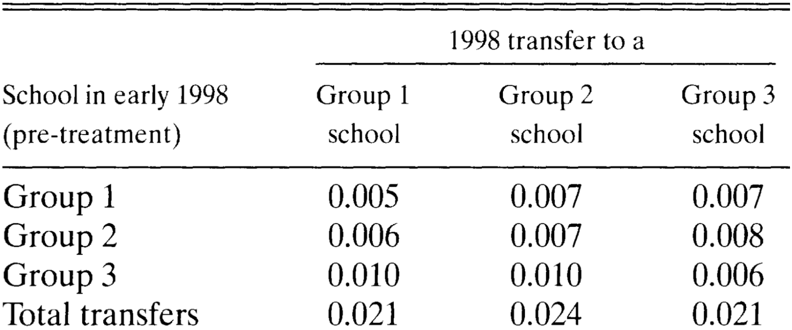
Check compliance of treatment

- Table III
- Table IV: see if students move to treated schools



	Group 1		Group 2	
	Girls <13 years, and all boys	Girls ≥ 13 years	Girls <13 years, and all boys	Girls ≥ 13 years
	Treatment		*Comparison*	
Any medical treatment in 1998 (For grades 1–8 in early 1998)	0.78	0.19	0	0
Round 1 (March–April 1998), Albendazole	0.69	0.11	0	0
Round 1 (March–April 1998), Praziquantel[b]	0.64	0.34	0	0
Round 2 (Oct.–Nov. 1998), Albendazole	0.56	0.07	0	0

Taken from Table III of Miguel & Kremer (2004)



	1998 transfer to a		
School in early 1998 (pre-treatment)	Group 1 school	Group 2 school	Group 3 school
Group 1	0.005	0.007	0.007
Group 2	0.006	0.007	0.008
Group 3	0.010	0.010	0.006
Total transfers	0.021	0.024	0.021

Taken from Table IV of Miguel & Kremer (2004)

Digression: Partial compliance

- Estimate Intention-To-Treat (ITT)
- Do not use IV if treatment spill-over is expected

cf. Section 6.2 of Duflo et al. (2008)

1.4 Estimating externality

$$
\begin{aligned}
Y_{ijt} &= \alpha + \beta T_{it} + \mathbf{X}'_{ijt}\delta \\
&\quad + \sum_{d}(\gamma_d N^{T}_{dit} + \phi_d N_{dit}) + u_i + e_{ijt}
\end{aligned}
$$

Y_{ijt} : outcome for pupil *j* in school *i* in year *t*

T_{it} : treatment dummy for school *i* in year *t*

1.4 Estimating externality

$$
\begin{aligned}
Y_{ijt} &= \alpha + \beta T_{it} + \mathbf{X}'_{ijt}\delta \\
&\quad + \sum_{d} (\gamma_d N^{T}_{dit} + \phi_d N_{dit}) + u_i + e_{ijt}
\end{aligned}
$$

N^{T}_{dit} : # of pupils in treated schools within distance *d* from school *i* in year *t* (*d*: 0-3km, 3-6km)

1.4 Estimating externality

$$\begin{aligned} Y_{ijt} &= \alpha + \beta T_{it} + \mathbf{X}'_{ijt}\delta \\ &\quad + \sum_d (\gamma_d N^T_{dit} + \phi_d N_{dit}) + u_i + e_{ijt} \end{aligned}$$

N_{dit} : # of pupils in all schools within distance d from school i in year t

- N^T_{dit}: exogenous conditional on N_{dit}

1.4 Estimating externality

$$\begin{aligned} Y_{ijt} &= \alpha + \beta T_{it} + \mathbf{X}'_{ijt}\delta \\ &\quad + \sum_{d} (\gamma_d N^T_{dit} + \phi_d N_{dit}) + u_i + e_{ijt} \end{aligned}$$

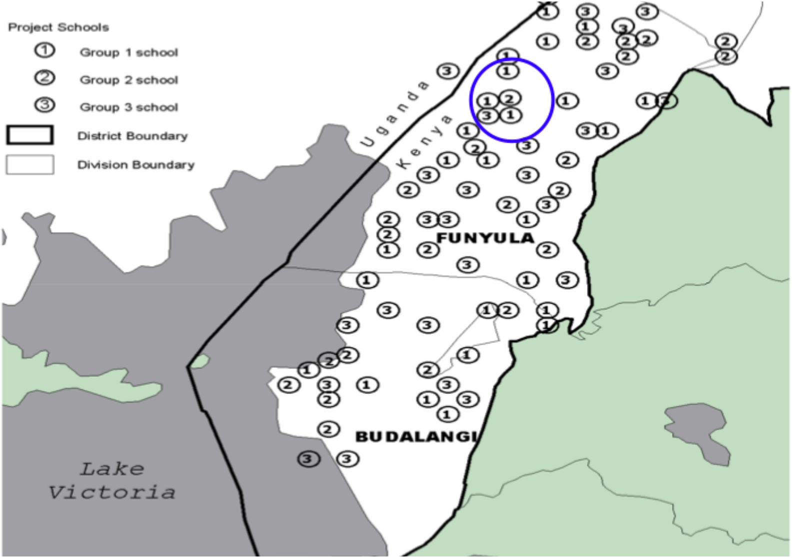
β : Direct effect of deworming for treated schools (incl. those not treated)

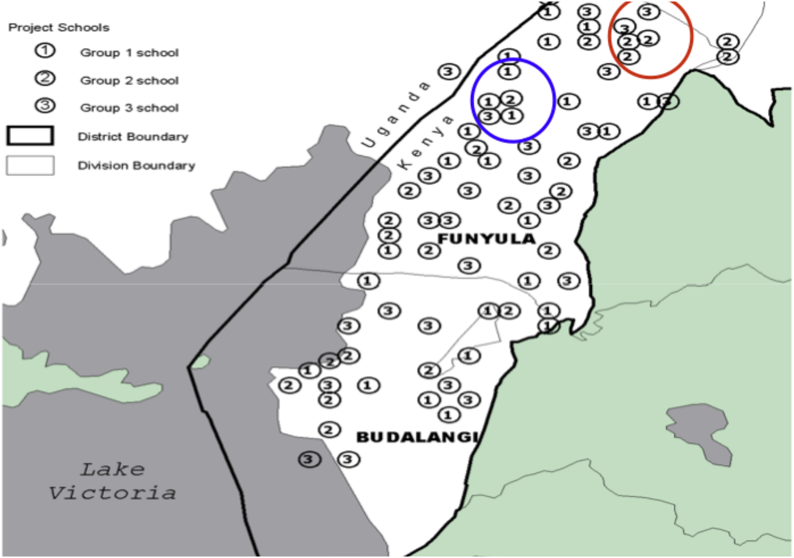
γ_d : Spill-over effect of deworming

1.4 Estimating externality

$$\begin{aligned} Y_{ijt} &= \alpha + \beta T_{it} + \mathbf{X}'_{ijt}\delta \\ &+ \sum_{d} (\gamma_d N^T_{dit} + \phi_d N_{dit}) + u_i + e_{ijt} \end{aligned}$$

- ArcGIS's Buffer and Spatial Join tools will help calculating N^T_{dit} & N_{dit}


Project Schools
① Group 1 school
② Group 2 school
③ Group 3 school
District Boundary
Division Boundary
Uganda
Kenya
FUNYULA
BUDALANGI
Lake Victoria


Project Schools
Group 1 school
Group 2 school
Group 3 school
District Boundary
Division Boundary
Uganda
Kenya
FUNYULA
BUDALANGI
Lake Victoria

- De Mel et al. (2008) use a similar specification to check if spill-over biases the treatment coefficient estimate.

1.5 Results

- Table VII Column (1): moderate-heavy helminth infection

Indicator for Group 1 (1998 Treatment) School	-0.25^{***}
	(0.05)
Group 1 pupils within 3 km (per 1000 pupils)	-0.26^{***}
	(0.09)
Group 1 pupils within 3–6 km (per 1000 pupils)	-0.14^{**}
	(0.06)
Total pupils within 3 km (per 1000 pupils)	0.11^{***}
	(0.04)
Total pupils within 3–6 km (per 1000 pupils)	0.13^{**}
	(0.06)

Gauging the size of the average spill-over effect

- Average # of pupils in treated schools:
  454 within 3km, 802 within 3-6km

⇒ 23% points lower infection due to spillover

1.5 Results (cont.)

- Table IX Column (3): school participation rate

Treatment school (T)	0.051***		
	(0.022)		
First year as treatment school (T1)		0.062***	0.060***
		(0.015)	(0.015)
Second year as treatment school (T2)		0.040*	0.034*
		(0.021)	(0.021)
Treatment school pupils within 3 km (per 1000 pupils)			0.044**
			(0.022)
Treatment school pupils within 3–6 km (per 1000 pupils)			–0.014
			(0.015)
Total pupils within 3 km (per 1000 pupils)			–0.033**
			(0.013)
Total pupils within 3–6 km (per 1000 pupils)			–0.010
			(0.012)

1.6 Follow-up

- Baird, Hicks, Kremer, & Miguel (2011) track these kids in the sample 5+ years later.
- Find a substantial impact on labor market participation and income earnings

2. Conley & Udry (2010)

- Show measuring peers as geographic neighbors could cause severe omitted variable bias in the estimation of peer effects
- Use standard error estimation that takes into account spatial correlations (Conley 1999)

2.1 Research Question

Do farmers learn from their peers regarding input use (fertilizer) for new technology (pineapple)?

- Interesting?
  - If yes, only a few farmers need to be subsidized for universal adoption
- Original?
  - Overcome econometric issues of identifying the impact of peer behavior
- Feasible?
  - Very detailed data collection

Empirical Challenge

- In the previous literature, peers are approximated by geographic proximity

  e.g. Everyone else in the same village

- This makes it difficult to disentangle the peer effect from common shocks in the same geographic area

2.2 Data

- Panel household surveys (every six week in 1996-98) in 3 villages of southern Ghana
- Outcome variable: Changes in amount of fertilizer used
- Sample: 107 plantings by 47 pineapple plots whose previous planting is also observed
    - Other observed plantings are also used for constructing regressors

2.2 Data (cont.)

- Each farmer’s info neighbors: obtained by asking
    - Among 7 other farmers randomly chosen from the sample,
    - Whom they turn to for advice on their farm
- Location of all plots: collected by GPS receivers

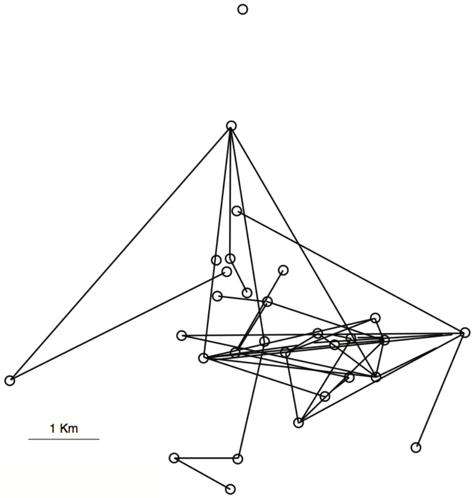


Figure 4. Roster Connections and Average Pineapple Plot Coordinates, Village 3

2.3 Theoretical predictions

Implication 3:

- Good news $\Rightarrow \ \Delta x_{it} = x_{j,s-5} - x_{i,t_p}$
- Otherwise, $\Delta x_{it} = 0$

x_{it} Amount of fertilizer chosen by farmer *i* in period *t*

t_p *i*'s previous period of cultivation

j *i*'s information neighbor

s Period between t_p and *t* in which *i* observes *j*'s profit

Implication 4:

- The effect should diminish with farmer *i*’s experience

2.4 Testing implication 3

Therefore, define

$$M_{i,t} \equiv \frac{GoodNews(x_{j,s-5}) \times (x_{j,s-5} - x_{i,t_p})}{Experience_{it}}$$

- $GoodNews(x_{j,s-5})$: dummy for $\pi_{j,s}(x_{j,s-5})$ above *i*’s expectation
- $Experience_{it}$: How many plantings *i* experienced up to time *t*

2.5 Estimation method

OLS estimation of

$$\Delta x_{it} = \beta_1 M_{it} + \beta_2 \Gamma_{it} + \mathbf{z}'_{it}\boldsymbol{\beta}_3 + \nu_{it}$$

- Γ_{it}: Changes in growing conditions for farmer *i* at time *t*

$\Rightarrow$ Common geographic shocks in the info. neighborhood controlled for

- $\mathbf{z}_{it}$: other controls

Changes in growing conditions

$$\Gamma_{it} \equiv x_{it}^{close} - x_{it_p}$$

x_{it}^{close}: Average of x_{ks} where:

- k: plots within 1km of plot i
- $s \in \{t-3, t-2, t-1, t\}$

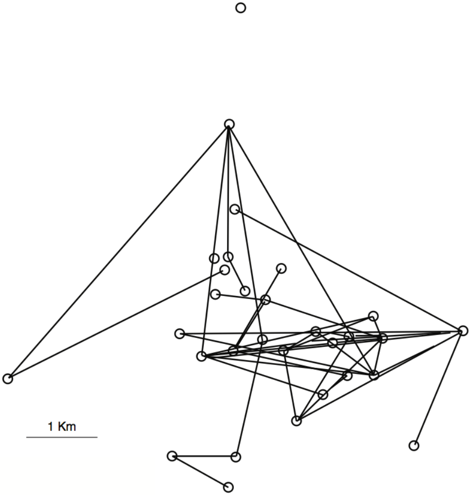


Figure 4. Roster Connections and Average Pineapple Plot Coordinates, Village 3

2.5 Estimation method (cont.)

$$\Delta x_{it} = \beta_1 M_{it} + \beta_2 \Gamma_{it} + \mathbf{z}'_{it}\beta_{\mathbf{3}} + \nu_{it}$$

- S.E.: Calculated by Conley's (1999) method
  - Which is now the standard procedure for cross-sectional regression with spatial data

Digression: Conley (1999)

- Denote a vector of coefficients estimated by OLS by **b**
- In general, we have

$$Var(\mathbf{b}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E(\varepsilon\varepsilon')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

$\mathbf{X}$: *n* (# of obs.) by *k* (# of regressors) data matrix

ε : *n*-dimensional vector of the error term

- Denote *ij*-th element of $E(\varepsilon\varepsilon')$ by ρ_{ij} (error correlation between *i* & *j*)
- Conley (1999) imposes $\rho_{ij} = K_{ij}\hat{\varepsilon}_i\hat{\varepsilon}_j$ where K_{ij} is

$$\begin{cases} (1-\frac{x_{ij}}{\bar{x}})(1-\frac{y_{ij}}{\bar{y}}) & \text{if } x_{ij} < \bar{x} \text{ \& } y_{ij} < \bar{y} \\ 0 & \text{otherwise} \end{cases}$$

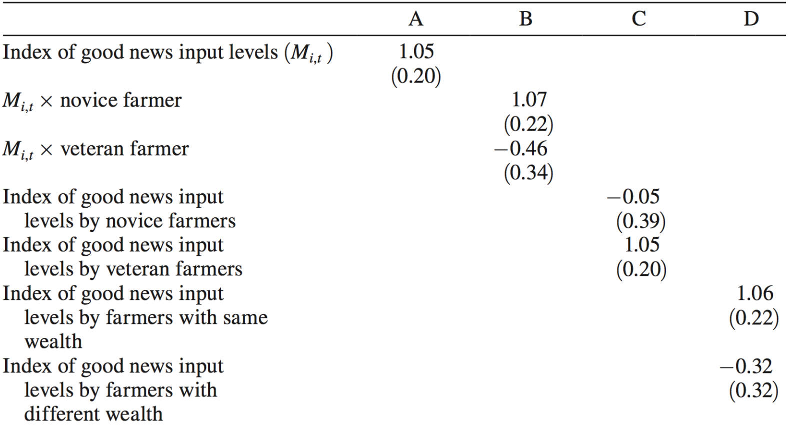
  - x_{ij}, y_{ij}: distance in x,y dimension
  - $\bar{x}, \bar{y}$: cut-off value chosen by researcher (1.5km in Conley-Udry 2010)

- Tim Conley's website (economics.uwo.ca/faculty/conley) offers Stata ado files for implementing Conley (1999) for OLS, GMM, and Logit
- In this ado program, x_{ij}, y_{ij} is calculated from the coordinates of *i* and *j*

$\Rightarrow$ Important to have coordinates in meters, not in degrees

- For a cross-sectional regression using spatial data, Conley standard errors are by now the industry standard.
- This method cannot be used for panel regressions...

2.6 Results (Table 5)



	A	B	C	D
Index of good news input levels ($M_{i,t}$)	1.05			
	(0.20)			
$M_{i,t}$ × novice farmer		1.07		
		(0.22)		
$M_{i,t}$ × veteran farmer		−0.46		
		(0.34)		
Index of good news input levels by novice farmers			−0.05	
			(0.39)	
Index of good news input levels by veteran farmers			1.05	
			(0.20)	
Index of good news input levels by farmers with same wealth				1.06
				(0.22)
Index of good news input levels by farmers with different wealth				−0.32
				(0.32)

2.7 Additional Results

- Impact on labor use: similar result for pineapple while no learning for maize-cassava (Table 7 A-B)
  - ⇐ Nice placebo test
- Good news in geographic neighborhood: misleading results (Table 7 C)
  - ⇒ Measuring the ACTUAL network: crucial

Crop	Pineapple (labor and cost in cedis per plant) A	Maize-cassava (labor cost in 1000 cedis per hectare) B	Maize-cassava (labor cost in 1000 cedis per hectare) C
Index of good news input levels: M^{labor}	1.96 (0.86)	0.02 (0.16)	
Index of good news input levels in the the geographic neighborhood			0.32 (0.12)
Average deviation of lagged use from geographic neighbors' use $[\,\Gamma_{i,t}^{labor}\,]$	0.49 (0.20)	0.74 (0.09)	

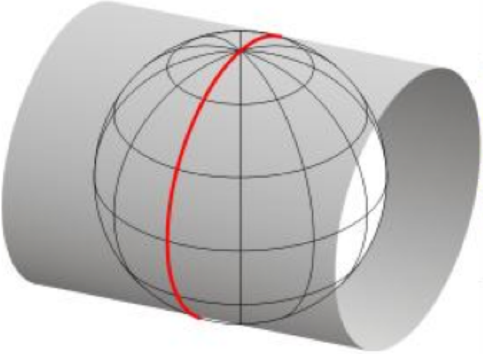
- Launch ArcMap now.
- Download the folder “Lecture3” from:

Network > INK00001 > teacher > Economics

3. UTM projections

- The buffer tool requires the proper calculation of distance
- With geographic coordinate system, this is not feasible (with one exception; see below)
- UTM projections should be used for calculating distances (& areas) for a small study area such as the one for Miguel-Kremer and Conley-Udry

In UTM projections, earth surface is projected onto a cylinder tangent on a meridian (called the *standard meridian*)



- Father away from standard meridian, more distortion

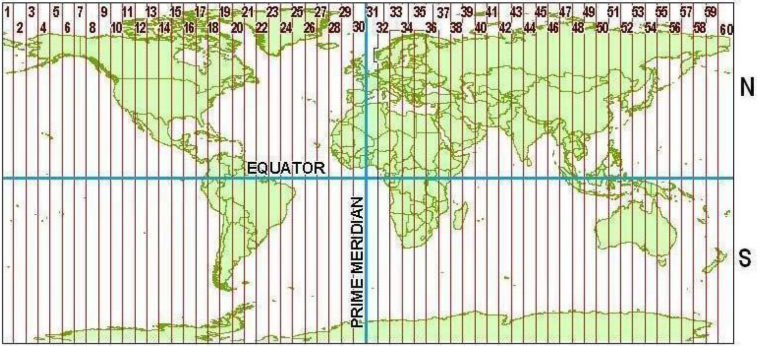
⇒ To minimize distortions:

1. Earth is divided into 60 zones
   ⇒ Each zone spans 6 degrees in longitude
2. For each zone, the standard meridian is set in the middle
3. The scale factor is 0.9996 on the standard meridian

Scale factor

- is how much the distance along the standard meridian is scaled up
- 0.9996 means the distance on the standard meridian is slightly shorter than its real distance
- 0.9996 is chosen to minimize the overall distortion within the 6-degree expanse in longitude.

UTM ZONE NUMBERS

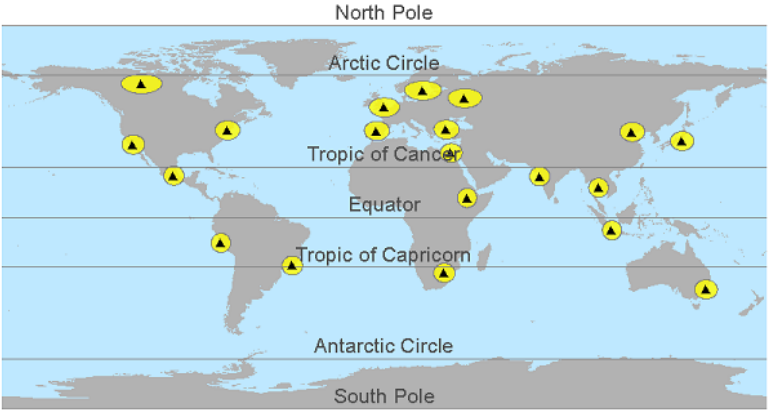


- ArcGIS provides the projection for each of the 120 UTM zones (the 60 zones divided by the equator)
- Pick the appropriate zone for your study area
    - In Conley and Udry (2010), it should be UTM 30 North.

cf. In lecture 6, we use several UTM zones covering India

- The Buffer tool works with the geographic coordinate system, though, if the input is a point feature class.

e.g. 500km (geodesic) buffer from several points on the earth


North Pole
Arctic Circle
Tropic of Cancer
Equator
Tropic of Capricorn
Antarctic Circle
South Pole

4. Using the Buffer tool in ArcGIS

Exercise 1: Identify the geographic neighbors of each plot (within 1km radius) in Conley & Udry (2010) data

- First, see the plot location data (udry2010.txt)
    - If you do a survey with GPS receiver, you will have a XY data like this one
- This is a panel data: each plot is observed whenever pineapples are grown
- XY coordinate is in meters: UTM projections

- Conley and Udry (2010) construct the growing condition variable as

$$\Gamma_{it} \equiv x_{it}^{close} - x_{it_p}$$

x_{it}^{close}: Average of x_{ks} where:

  - k: plots within 1km of plot i
  - $s \in \{t-3, t-2, t-1, t\}$

⇒ Output table should include the following columns:

  - ID of plot (*i*)
  - Time (*t*) of planting for plot *i*
  - ID of another plot (*j*) in the neighborhood of *i*
  - Time (*s*) of planting for plot *j*

- Read this in Stata and keep observations if $t - 3 \leq s \leq t$

  (see fn. 20 of Conley and Udry 2010)

- Merging fertilizer input data for (j, s)
- Collapse by (i, t) to obtain average

Geoprocessing tools to be used:

1. Make XY Event Layer
2. Copy Features
3. Buffer (Analysis)
4. Spatial Join

- 1 & 2 convert GPS data on plot location into a point feature class shape file (cf. Lec 1 Ex 3; Lec 2 Ex 2)
- 3 & 4 identify geographic neighbors for each plot

Make XY Event Layer

- XY Table: udry2010.txt
- X Field: Xcoord
- Y Field: Ycoord
- Spatial Reference: WGS 1984 UTM Zone 30N (in the folder: Projected Coordinate Systems > UTM > WGS 1984 > Northern Hemisphere)

Then use Copy Features to make the output permanent

How to match each plot with all the plots within 1km

- By Buffer, create a 1km radius circle polygon for each plot
- By Spatial Join, join to each circle polygon all the plots within the polygon

Buffer

- Distance: Linear unit, 1, Kilometers
- Dissolve Type: NONE

- Now run the model
- See the attribute tables of the plot point features and the buffer polygon features
- Notice field names are the same between the two tables
- This creates one problem for Spatial Join (see below)

Spatial Join

(cf. Lec 2 Ex 1-2)

- Target Features:
  the output from Buffer
- Join Features:
  the output from Copy Features
- Join Operation: JOIN ONE TO MANY
- Keep All Target Features: checked
- Field Map of Join Features: see next slide
- Match Option: INTERSECT or CONTAINS

Field Map of Join Features

- As we see in Lec 2 Ex 1, we should delete everything as long as we run the Spatial Join in Python
- If we run the Spatial Join in the Model Builder, deleting everything doesn’t work
  - ⇐ When both target and join features have the same field names, the join features’ fields won’t be merged

- The Spatial Join in the Model Builder automatically assigns different field names to the join feature class (by adding "_1" at the end) if the join features are already created
- It's a good idea to run the Model before you use Spatial Join in the Model Builder.

To summarize...

- Don’t run Spatial Join in the Model Builder.
- Always delete fields from Field Map of Spatial Join
- Run Spatial Join in Python
- If you still need to run Spatial Join in the Model Builder, run the model before adding Spatial Join to the model.

It took three years for me to figure this out...

- Your Model should now look like the same as the one saved in L3model.tbx in the “solutions4exercises” folder in the downloaded data set

- Now export the model as a Python script (cf. Lec 2 Ex 3)
- Edit the script by using the template included in the downloaded folder (template4L3.py)

- Now let’s review the steps to edit and run a Python script for geoprocessing.
    1. Try-Except statement
    2. Local variables for file names
    3. Print commands
    4. Run the script

(1) 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.

(2) Edit local file names

- Copy local file names 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

(2) Edit local file names (cont.)

- Rename the final output shapefile to something 5 letter long or less
  - The attributes table of this shapefile will be read in Stata
  - Stata's odbc load command cannot read a dBASE table whose name is longer than 5 letters (cf. Lec 1, sec 4.2)

(3) Add print commands

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

(4) Run the script

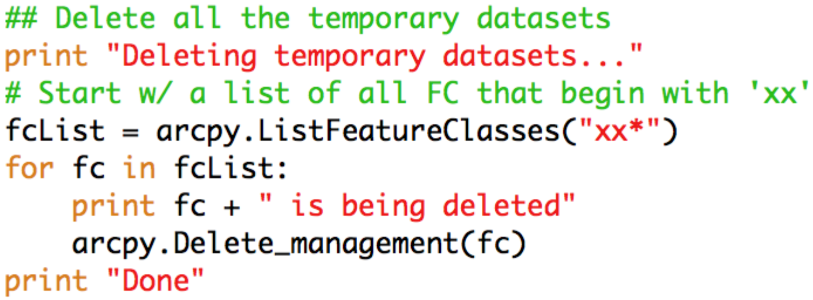
- Save the script (File > Save / Save As)
- Click “Run > Run Module”
- Cross your fingers
- Read output shapefiles in ArcMap (Lecture 1 Exercises 1-2) to see if everything worked out

5. Loop over files

Let’s learn one useful Python scripting trick.

- We may want to delete all the intermediate files at once
  - To create one particular output, ArcGIS often requires quite a few tools to use
  - And each tool creates an output
  - We don’t need these intermediate outputs
- Looping over files is useful for this purpose

- Name all the intermediate files in the form of “xx....”
- Then the set of commands in the next slide deletes all the files of name “xx...” in the working directory
    - In the template script you downloaded, these commands are commented out by ##
    - For the first run, you perhaps want to keep intermediate files so you can check where the script went wrong



```
## Delete all the temporary datasets
print "Deleting temporary datasets..."
# Start w/ a list of all FC that begin with 'xx'
fcList = arcpy.ListFeatureClasses("xx*")
for fc in fcList:
    print fc + " is being deleted"
    arcpy.Delete_management(fc)
print "Done"
```

Review: Object

- An *object* consists of properties and methods
- A *method* is a function
  - To use a method, type

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

Enumeration methods

- The arcpy object contains enumeration methods
  - ListDatasets
  - ListFeatureClasses
  - ListRasters
  - ListTables
- These methods return a list of the names of files in the workspace

  (as specified by arcpy.env.workspace("directory"))

- To obtain the list of shape files of the name starting with xx, type

```
fcList = arcpy.ListFeatureClasses(“xx*”)
```

*** “fcList” can be replaced with some other word

- For raster files,

```
rasList = arcpy.ListRasters(“xx*”)
```

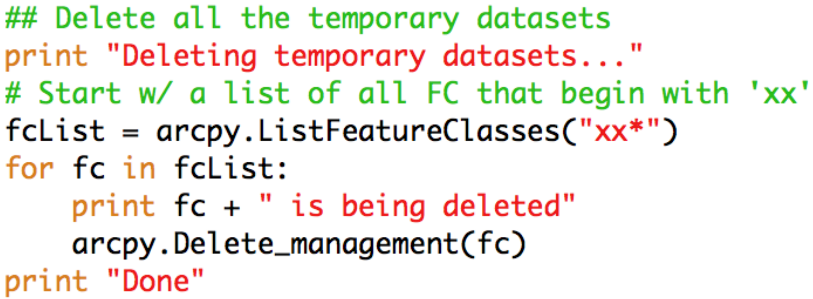
- For dBASE tables (.dbf),

```
dbfList = arcpy.ListTables(“xx*”,
        “dBASE”)
```

- Then use the following command

```
for fc in fcList:
```

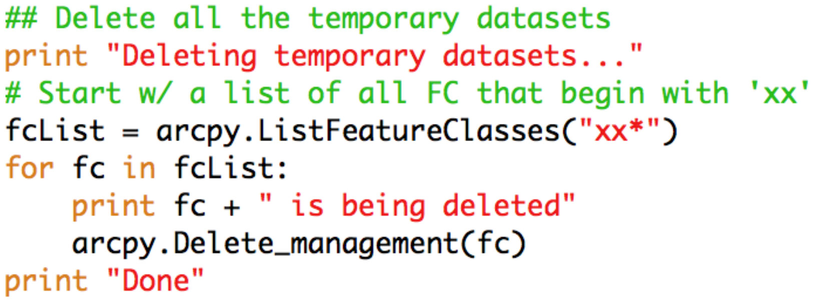
- This assigns the variable name “fc” to 1st file name in “fcList”.
- The indented commands that follow will be executed for this file name.
- Then Python comes back to this line and assign “fc” to 2nd file name in “fcList”
- And so forth, until the loop exhausts all the file names in “fcList”



```
## Delete all the temporary datasets
print "Deleting temporary datasets..."
# Start w/ a list of all FC that begin with 'xx'
fcList = arcpy.ListFeatureClasses("xx*")
for fc in fcList:
    print fc + " is being deleted"
    arcpy.Delete_management(fc)
print "Done"
```

How to concatenate strings

- Use +
- A new string should be enclosed by “ ” or ‘ ’
- A variable that contains a string can also be used



```
## Delete all the temporary datasets
print "Deleting temporary datasets..."
# Start w/ a list of all FC that begin with 'xx'
fcList = arcpy.ListFeatureClasses("xx*")
for fc in fcList:
    print fc + " is being deleted"
    arcpy.Delete_management(fc)
print "Done"
```

The Delete tool

- This tool simply deletes a file
- But only one at a time.
- That’s why we need a loop over files...

Exercise 2: delete intermediate files

- Rename intermediate shapefiles to “xx***.shp”
  - Since we use variables for file names, we only need to change one line per file, not all the methods that use this file.
- Uncomment all the commands for deleting intermediates (cf. Lec 2 Ex 8)
- Save the script
- Run the script again (or press F5)

- Your Python script should now look like the same as “modelscript4L3.py” in the “solutions4exercises” folder in the downloaded data set

Other types of looping

- The loop over fields (ie. columns in the attribute table for a shapefile) can be done with the ListFields method.
- See ArcGIS Desktop Help for ListFields, for detail
- For looping over values, we will learn in Lec 6.

5. What we’ve learned on ArcGIS

- Match each point feature with its neighboring points (i.e. within a radius of a certain distance)

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

Appendix 1: How to randomize

Two issues

- At which level treatment should be randomized
- How to achieve ex post balancing of covariates

A1.1 Level of treatment

- Miguel & Kremer (2004): treatment at school level, not at pupil level
- Such group-level treatments are:
  - Useful if within-group spillover effects are expected
  - Often more feasible than individual-level (ethically & cost-wise)
  - Drawback: Need more observations to achieve power to detect effect

cf. Sections 4.2 & 5.1 of Duflo et al. (2008)

- SE should be clustered at the group-level
- If # of groups <50 (Bertrand et al. 2004), use randomization inference
    - Recent example: Cohen & Dupas (2010)

cf. Section 7.1 of Duflo et al. (2008)

Randomization inference

- Suppose the number of groups is n, of which k treated.
- Consider a hypothetical treatment assignment by randomly picking k groups out of n
  - There are $n!/k!(n-k)!$ possible assignments
- For each hypothetical assignment, estimate the treatment effect

$\Rightarrow$ Obtain the distribution of treatment effect coefficients

- Then conduct hypothetical testing of the actual treatment effect against this distribution
  - Randomization inference: fix outcome & redraw treatment dummy
  - Regression-based inference: fix treatment dummy & redraw outcome (ie. error term)
- Valid for any sample size
  - In principle, we can use this even with more than 50 groups

A1.2 Ex post balancing

- Purpose of randomization: balancing the mean characteristics across groups
- With a small sample (typical for RCTs in development), balancing may not be achieved ex post, reducing statistical power to detect the treatment effect

- Bruhn & McKenzie (2009) discuss what methods of randomization achieve balancing ex post more likely
  - Stratification
  - Pair-wise matching
  - Re-randomization
- They use actual micro panel datasets to see whether future outcomes are balanced by each of these methods

- For non-persistent outcomes (firm profit, school attendance) or a large sample ($n = 300$), not much difference across methods
- With small sample (less than 300) and persistent outcomes (e.g. test scores, height), stratification and pair-wise matching that balances on baseline outcome achieves balance more likely

Pair-wise matching:

- Hotly debated on whether it’s better than stratification
  - King et al. (2007): politically robust
  - Imbens (2011): difficult to estimate heterogeneous impacts etc.
- Stata code available as data appendix to Bruhn-McKenzie (2009)
- Takes quite a lot of computational power

- Which variables to balance on?
  - Baseline outcome (Bruhn-McKenzie 2009)
  - Dummies for geographic regions
  - Those correlated w/ future outcomes (if panel baseline data available)
  - Those over which heterogeneous treatment effects are expected (Duflo et al. 2008, Sec 4).
  - Cluster population size (Imbens 2011)

    $\Leftarrow$ If interested in average treatment effect over population, not over clusters

Appendix 2: Checking balance

- Always show if pre-treatment outcomes & covariates are balanced btw. treated & control groups
- Clearly indicate which variables were used in randomization (if stratification or pair-wise matching was used)

- If not balanced for some variables,
    - Mention the direction of bias
    - Compare results with & without such variables controlled for in the regression analysis
    - Regress outcomes on such covariates to see the size of bias

      e.g. Gine & Yang (2009)

References for Lecture 3

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. “How Much Should We Trust Differences-in-differences Estimates?.” *Quarterly Journal of Economics* 119: 249-275.

Bruhn, Miriam, and David McKenzie. 2009. “In Pursuit of Balance: Randomization in Practice in Development Field Experiments.” *American Economic Journal: Applied Economics* 1(4): 200-232.

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.

Cohen, Jessica, and Pascaline Dupas. 2010. “Free Distribution or Cost-Sharing? Evidence from a Randomized Malaria Prevention Experiment.” *Quarterly Journal of Economics* 125(1): 1-45.

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

Conley, Timothy G, and Christopher R Udry. 2010. “Learning about a New Technology: Pineapple in Ghana.” *American Economic Review* 100(1): 35-69.

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

de Mel, Suresh, David McKenzie, and Christopher Woodruff. 2010. “Returns to Capital in Microenterprises: Evidence from a Field Experiment.” *Quarterly Journal of Economics* 123(4): 1329-1372.

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.

Freedman, David A. 2008. “On regression adjustments to experimental data.” *Advances in Applied Mathematics* 40(2): 180-193.

Giné, Xavier, and Dean Yang. 2009. “Insurance, credit, and technology adoption: Field experimental evidencefrom Malawi.” *Journal of Development Economics* 89(1): 1-11.

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.

Miguel, Edward, and Michael Kremer. 2004. “Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities.” *Econometrica* 72(1): 159-217.

Redding, Stephen J, and Daniel M Sturm. 2008. “The Costs of Remoteness: Evidence from German Division and Reunification.” *American Economic Review* 98(5): 1766-1797.