GIS FOR ECONOMICS RESEARCH

Masayuki Kudamatsu

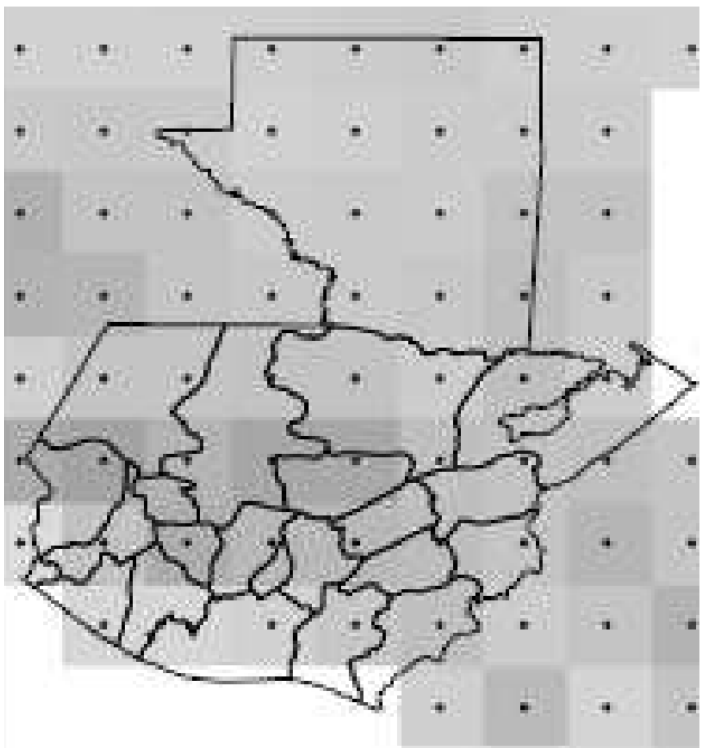
14-15 June, 2016

WHY GIS FOR ECONOMICS?

Reason 1: GIS makes more research feasible

GIS MAKES RESEARCH FEASIBLE (1/3)

BY MERGING DATA BY LOCATION



(Figure 2.3 of Dell 2009)

GIS MAKES RESEARCH FEASIBLE (2/3)

BY SCANNED OLD MAPS

e.g. Ethnic homelands in Africa by Murdock (1959)

Digitized by Nunn (2008)

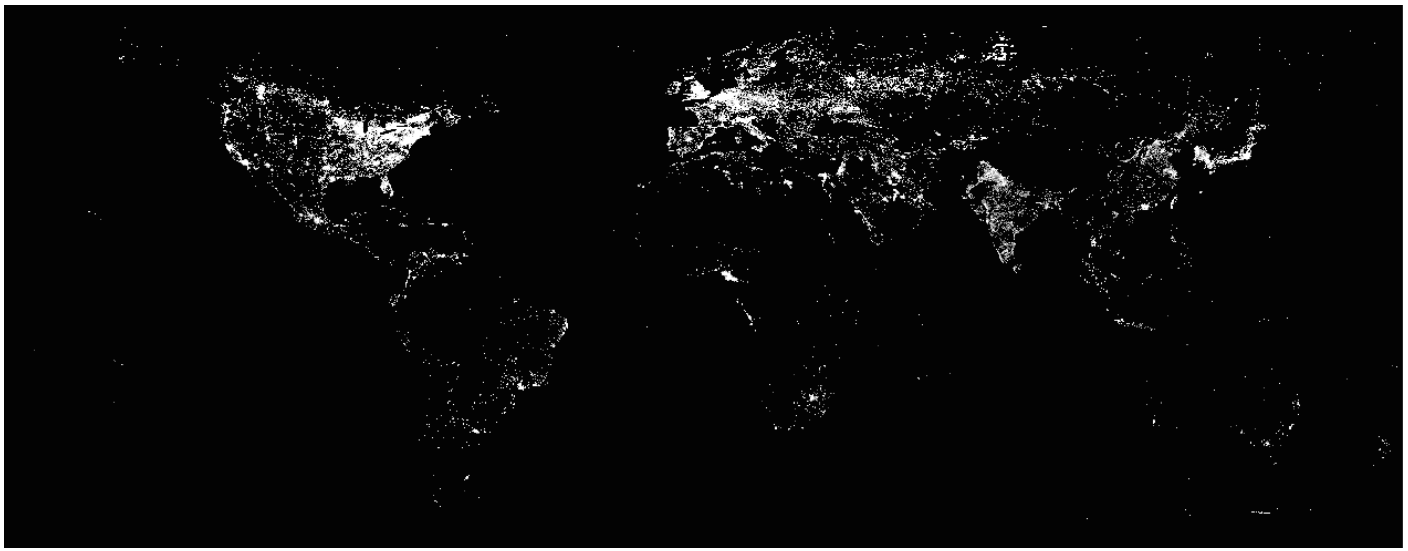
Used by Nunn & Wantchekon (2011), Michalopoulos & Papaioannou (2013, 2014, 2015), Alsan (2015), Alesina et al. (2016), etc.

(Figure 5A of Alsan (2015))

GIS MAKES RESEARCH FEASIBLE (3/3)

BY SATELLITE IMAGES

e.g. DMSP-OLS Nighttime Lights



Used by Henderson et al (2012), Pinkovskiy & Sala-i-Martin (2016), Hodler & Raschky (2014), Michalopoulos & Papaioannou (2013, 2014), Alesina et al (2016), etc.

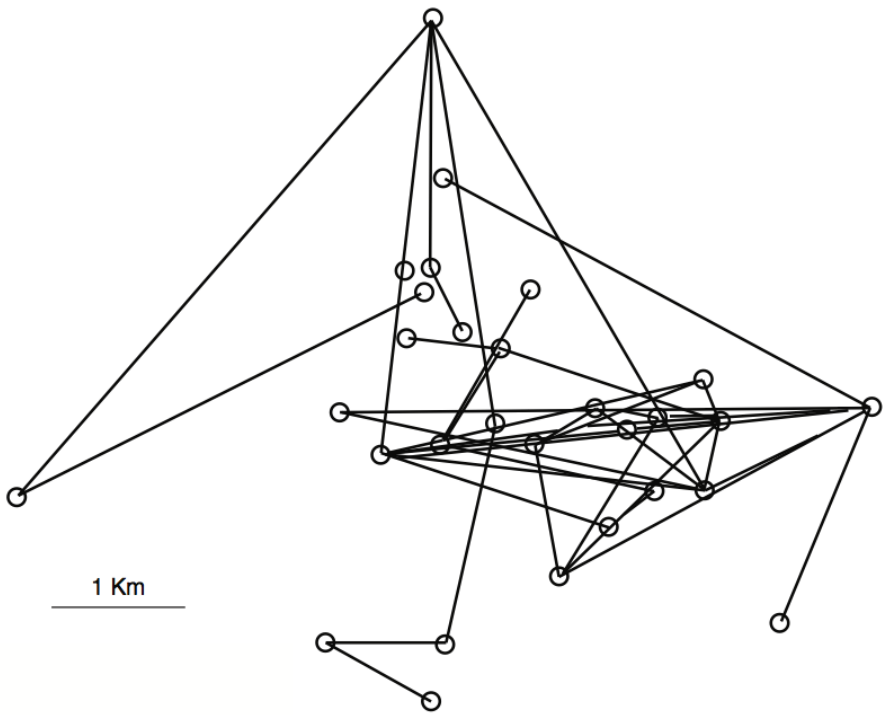
WHY GIS FOR ECONOMICS?

Reason 2: GIS makes identification more credible

GIS MAKES IDENTIFICATION CREDIBLE (1/4)

BY CONTROLLING FOR MORE COVARIATES

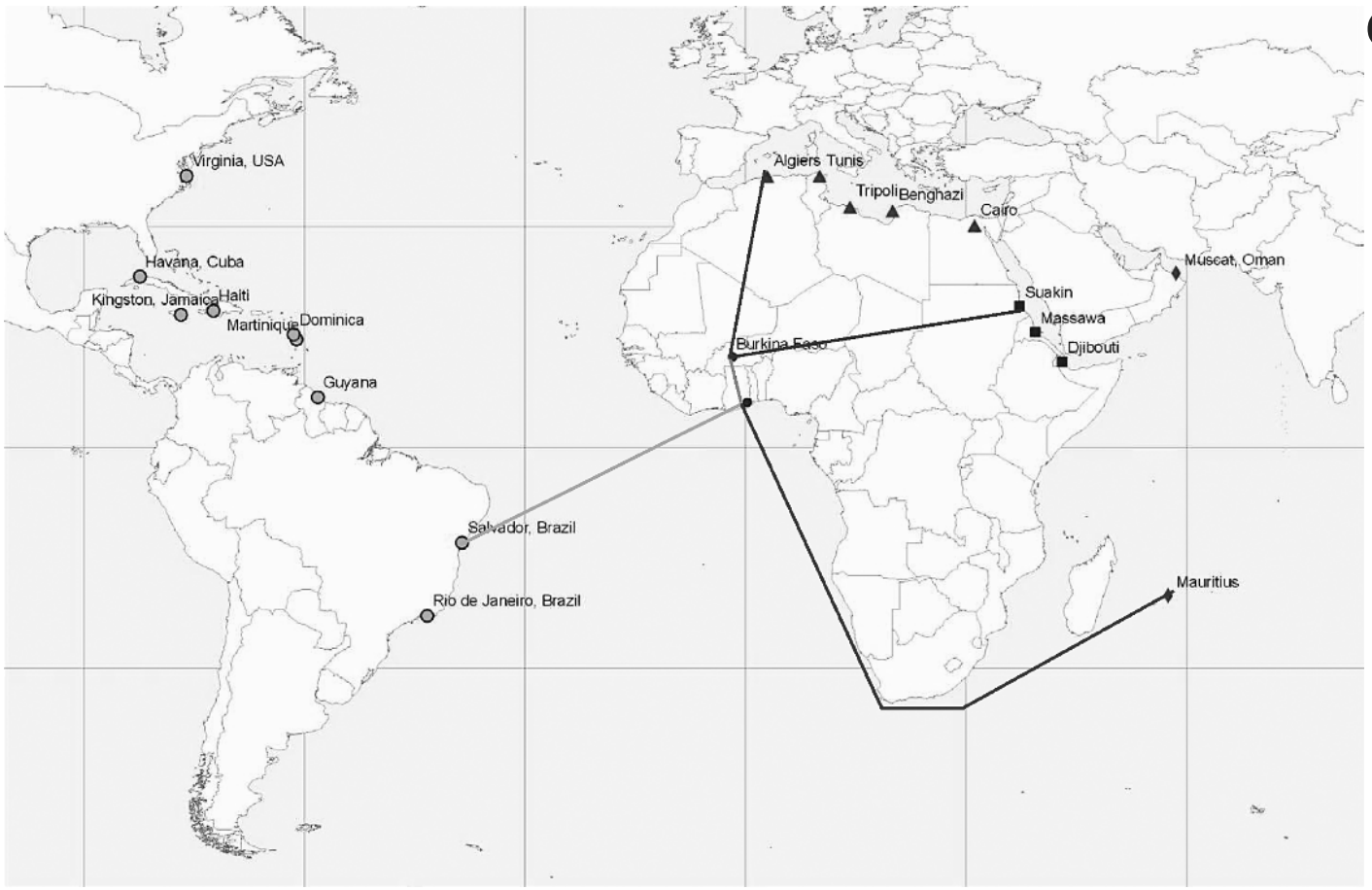
e.g. Peer effect estimation



(Figure 4 of Conley & Udry (2010))

GIS MAKES IDENTIFICATION CREDIBLE (2/4)

BY CONSTRUCTING INSTRUMENTS

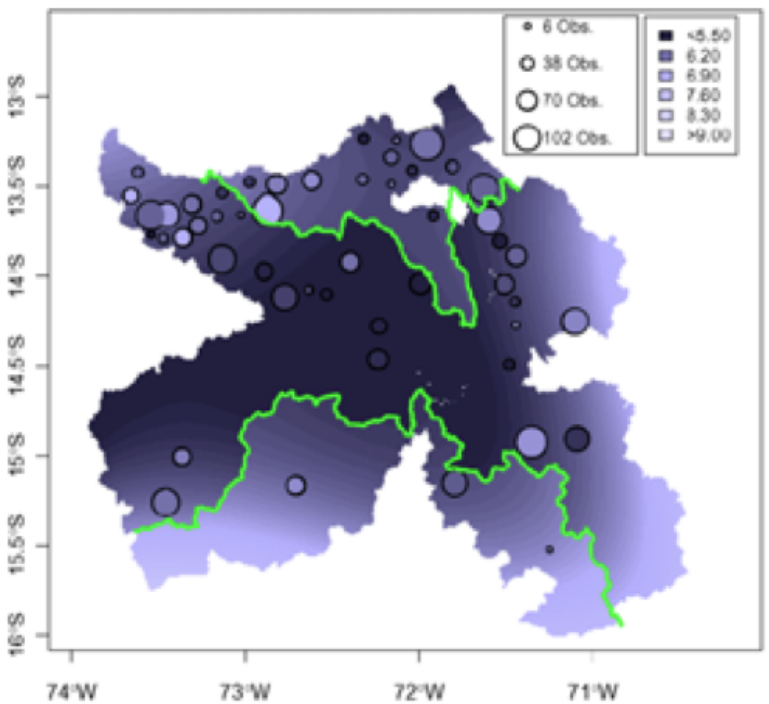


e.g. Distance

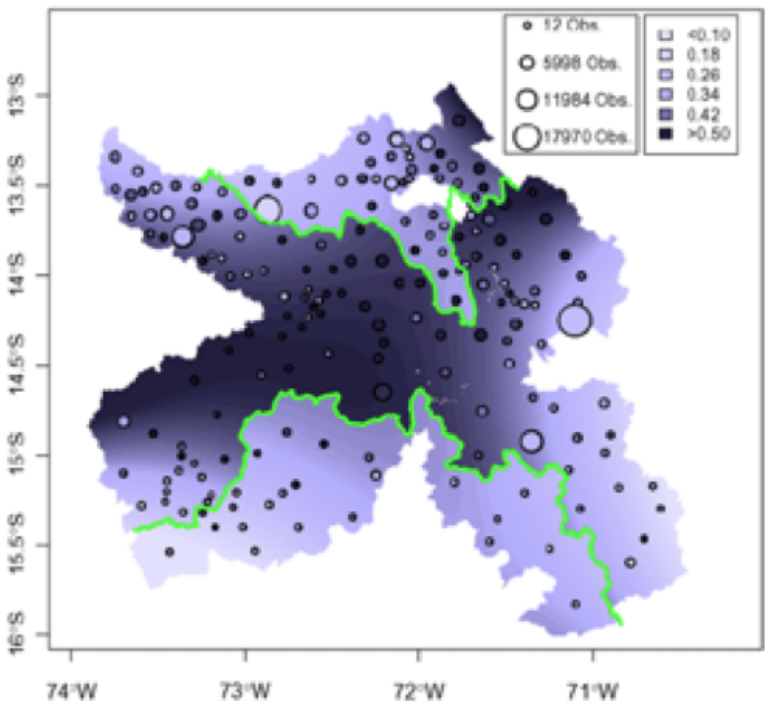
(Figure 5 of Nunn 2008)

GIS MAKES IDENTIFICATION CREDIBLE (3/4)

BY REGRESSION DISCONTINUITY



(a) Consumption (2001)

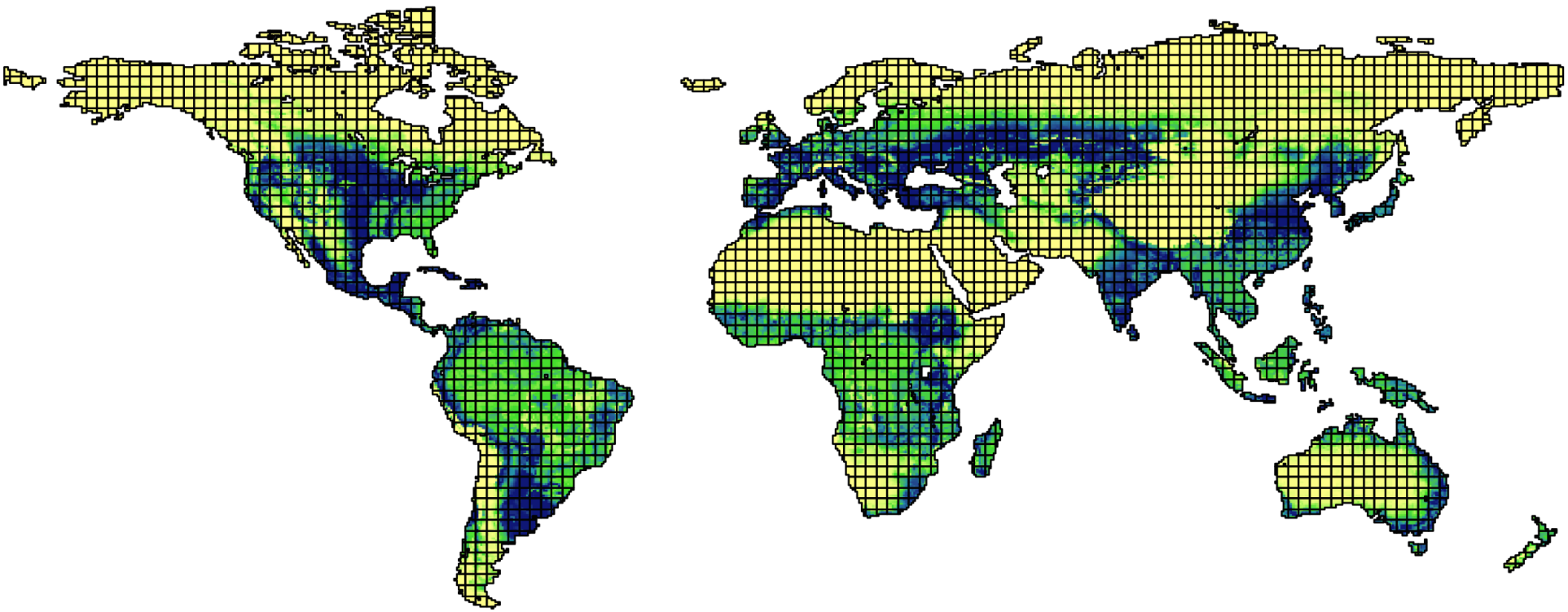


(b) Stunting (2005)

(Figure 2 of Dell (2010))

GIS MAKES IDENTIFICATION CREDIBLE (4/4)

BY EXOGENOUS BOUNDARIES



(image source)

ROAD MAP

1. GIS basics
2. Create spatial datasets on your own
3. Merge spatial datasets
4. Elevation
5. Distance
6. Spatial regression discontinuity
7. Surface area
8. Map Algebra

1. GIS BASICS

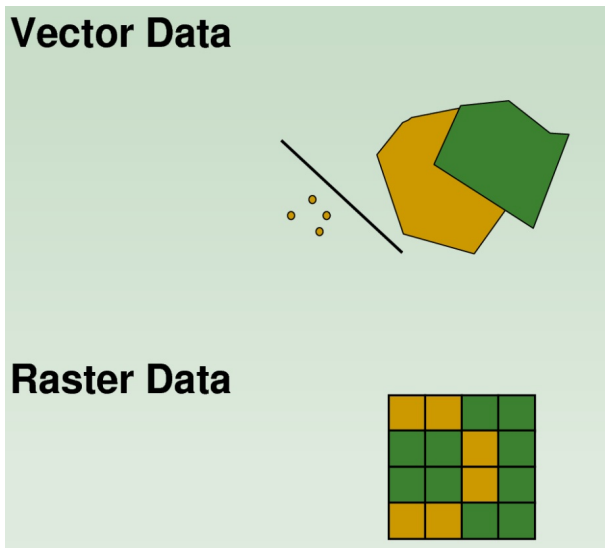
Data type

Coordinate systems

GIS software

1.1 DATA TYPE

Spatial data comes in two different formats: **Vector** & **Raster**



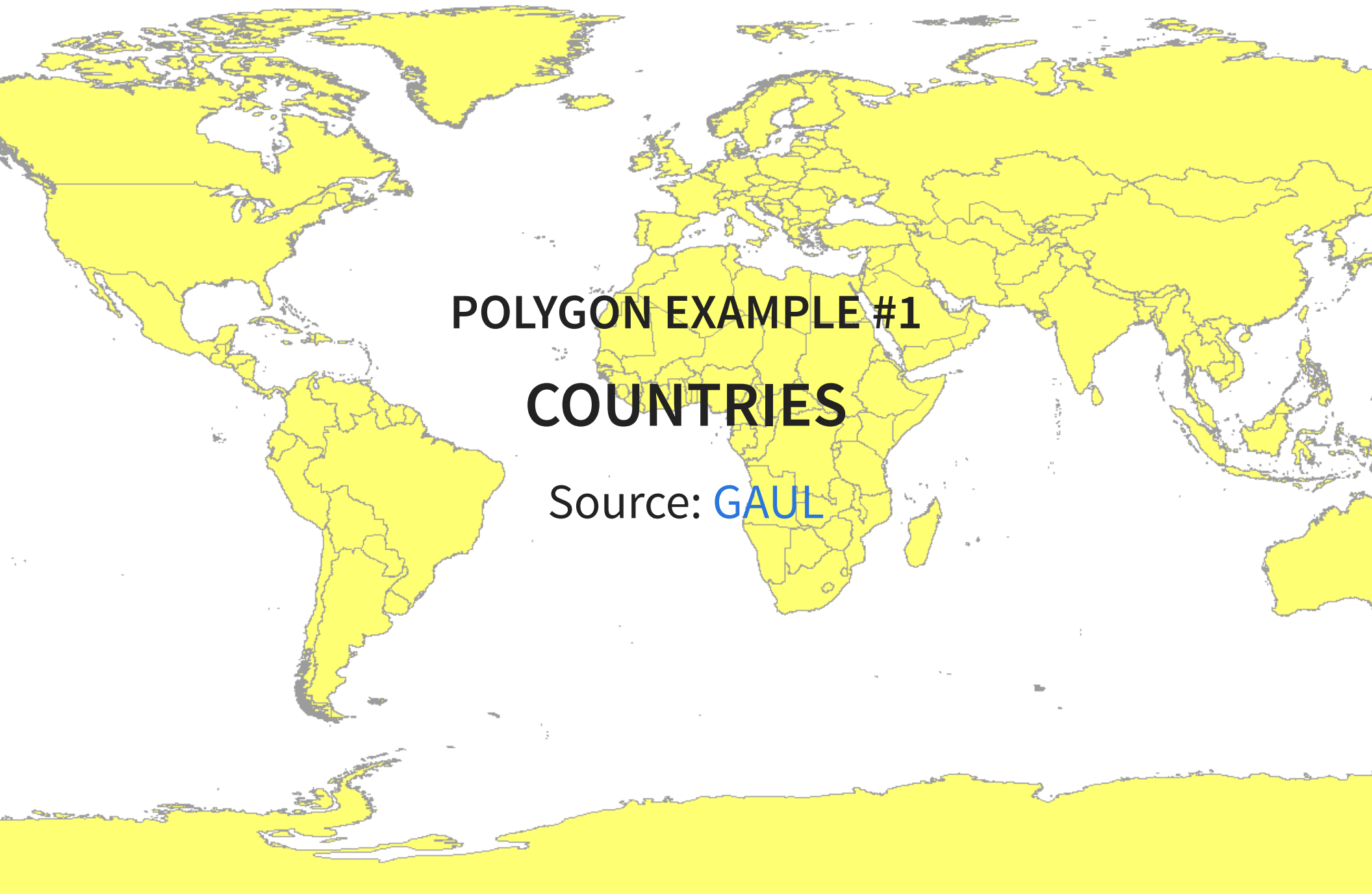
How to edit data differs a lot between them

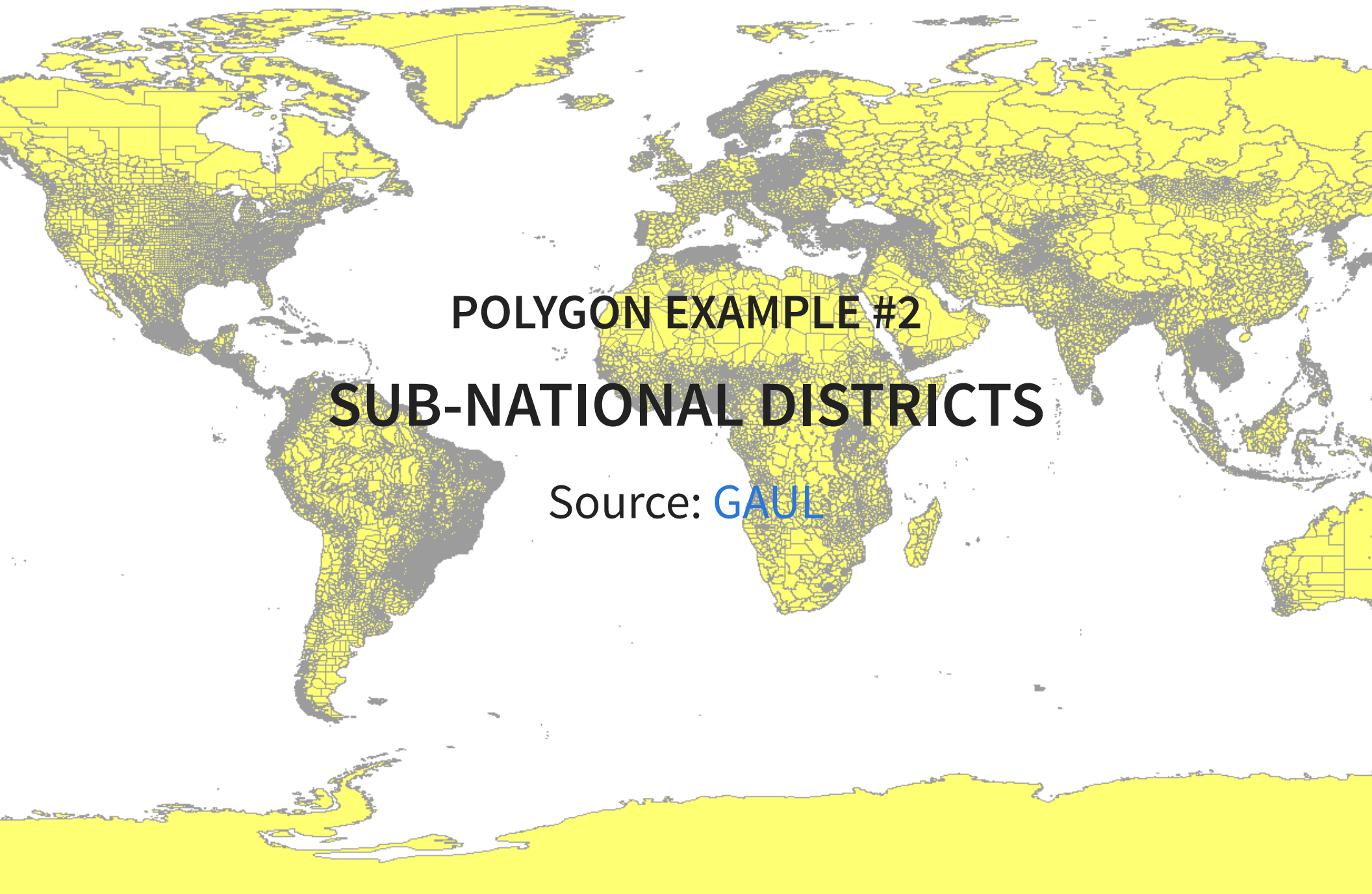
VECTOR DATA

Comes in three formats:

- Polygons
- Polylines
- Points

File format: **Shapefile** (.shp)

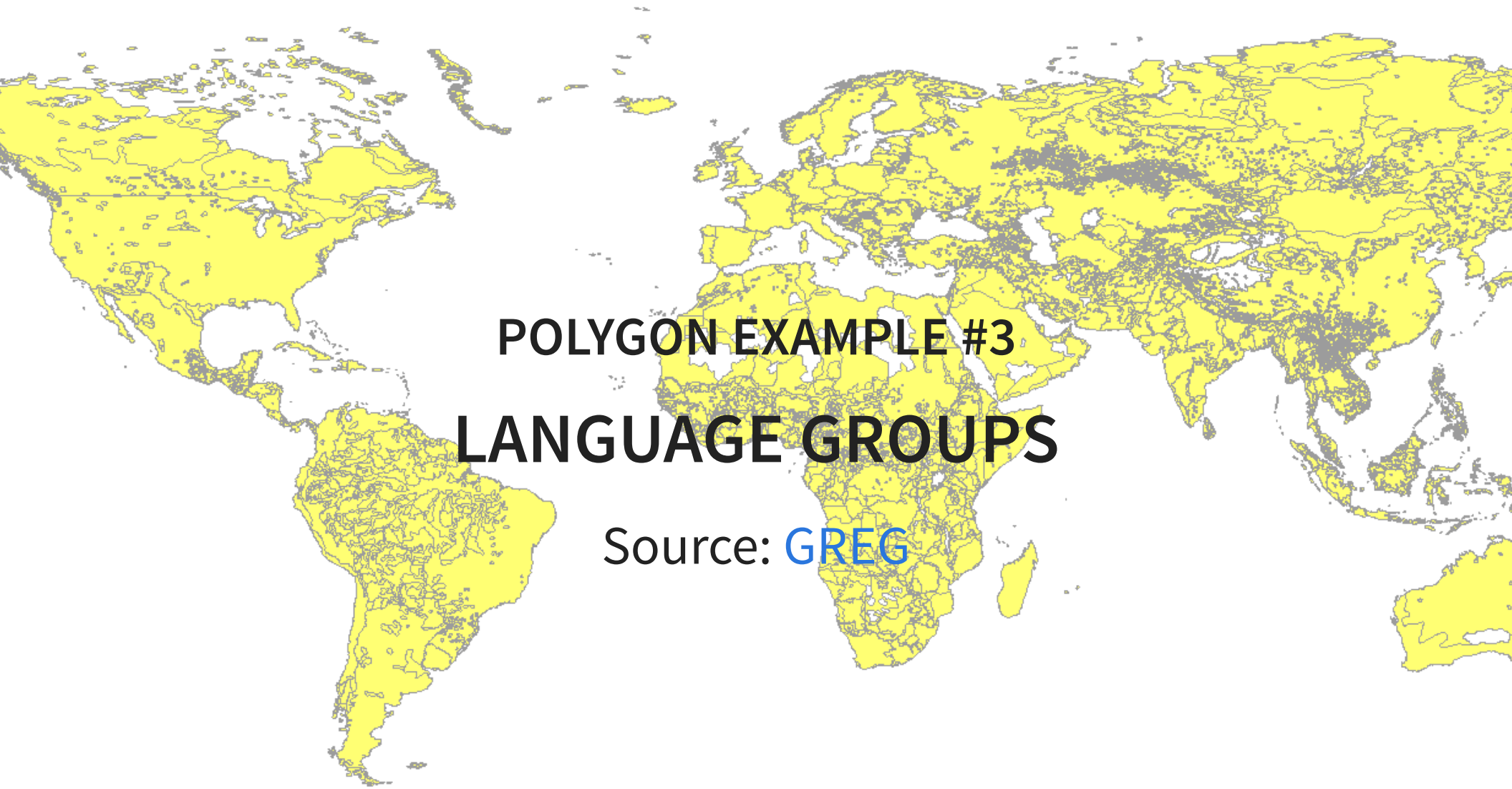

POLYGON EXAMPLE #1
COUNTRIES
Source: GAUL



POLYGON EXAMPLE #2

SUB-NATIONAL DISTRICTS

Source: GAUL



POLYGON EXAMPLE #3

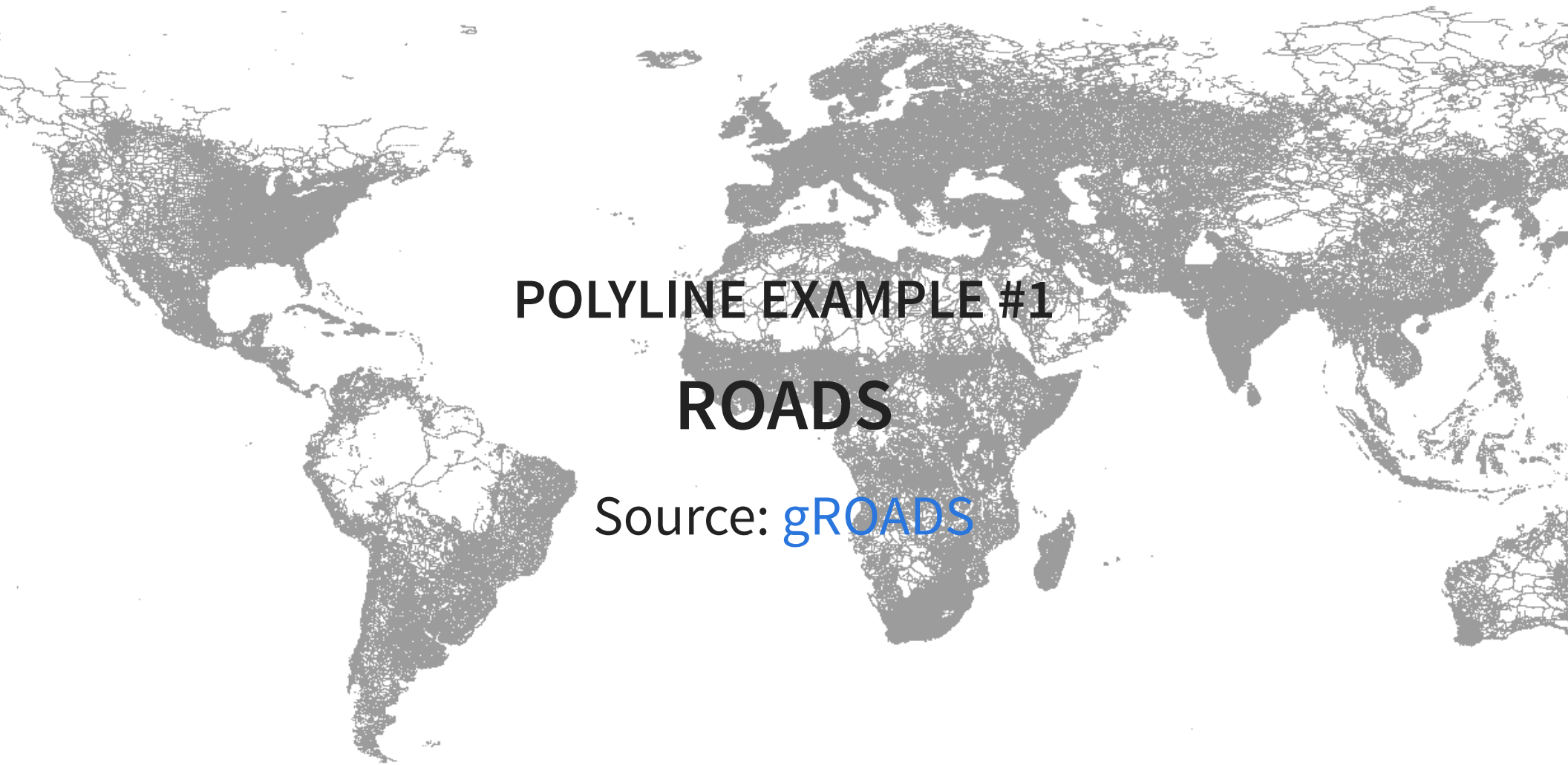
LANGUAGE GROUPS

Source: GREG

POLYGON EXAMPLE #4

LAKES & RESERVOIRS

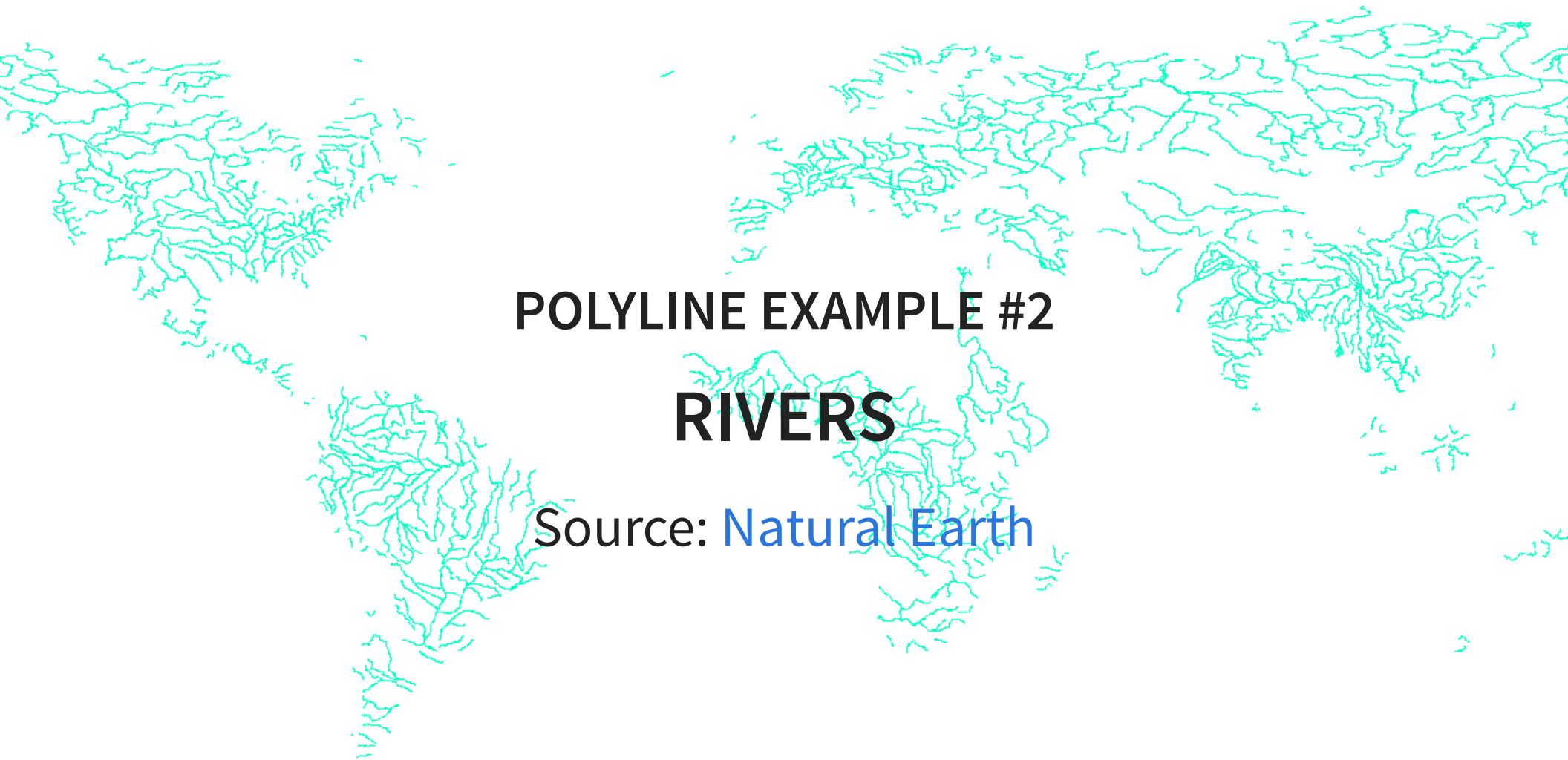
Source: Natural Earth



POLYLINE EXAMPLE #1

ROADS

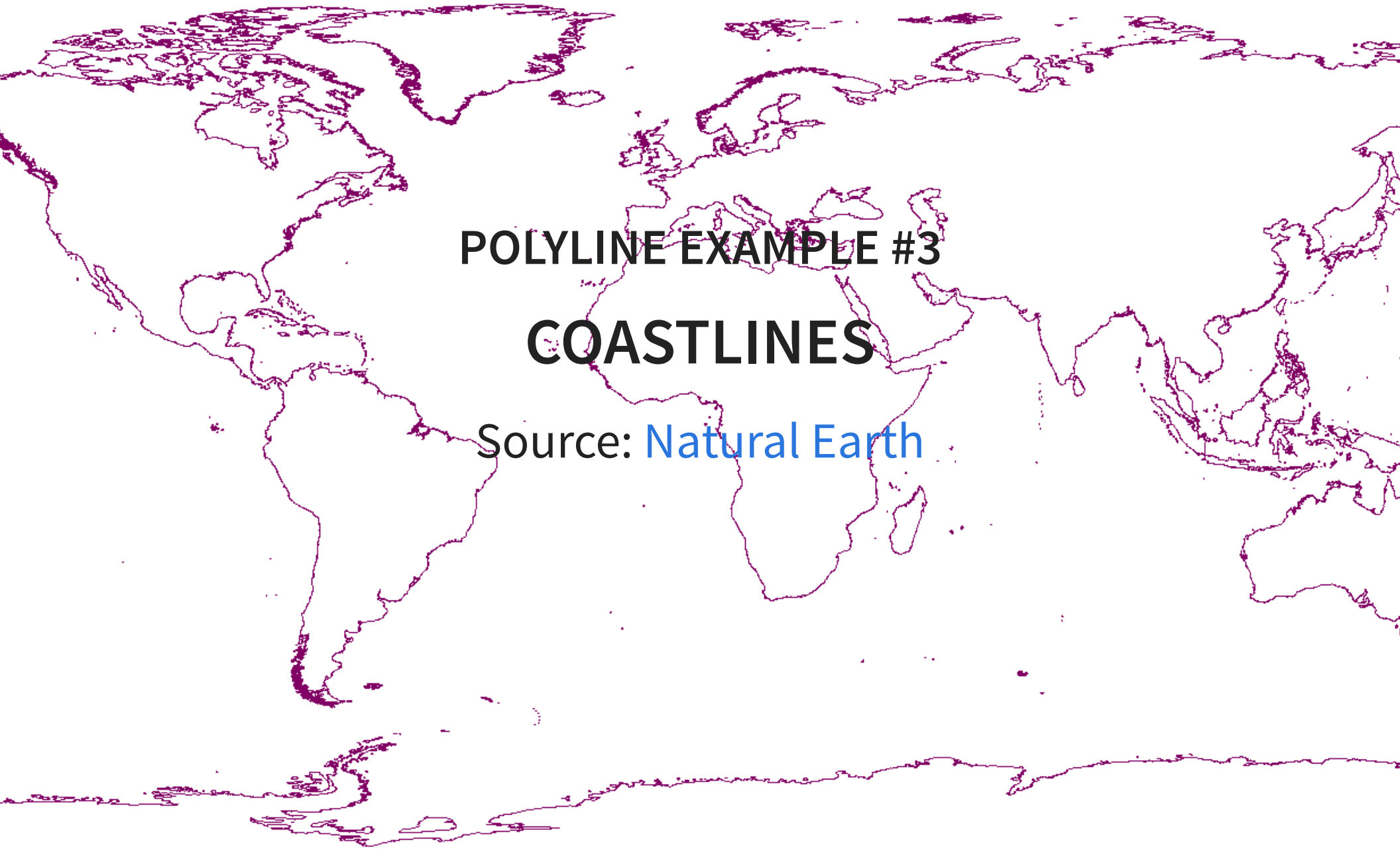
Source: gROADS



POLYLINE EXAMPLE #2

RIVERS

Source: Natural Earth


POLYLINE EXAMPLE #3
COASTLINES
Source: Natural Earth

POINT EXAMPLE

CONFLICT LOCATIONS (1997-2015)

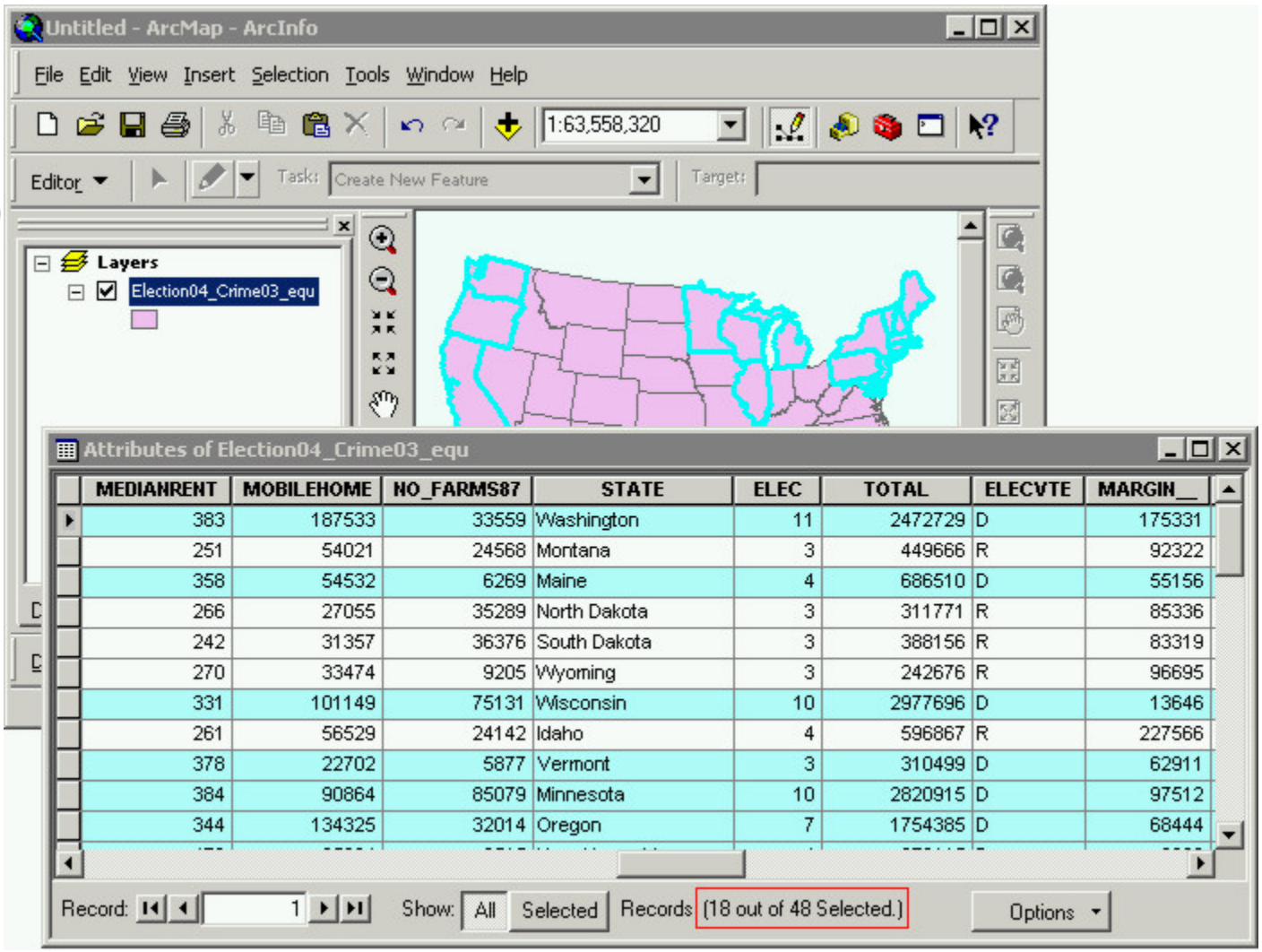
Source: ACLED

VECTOR DATA (CONT.)

Each unit: called a **feature**

Comes w/
attribute table

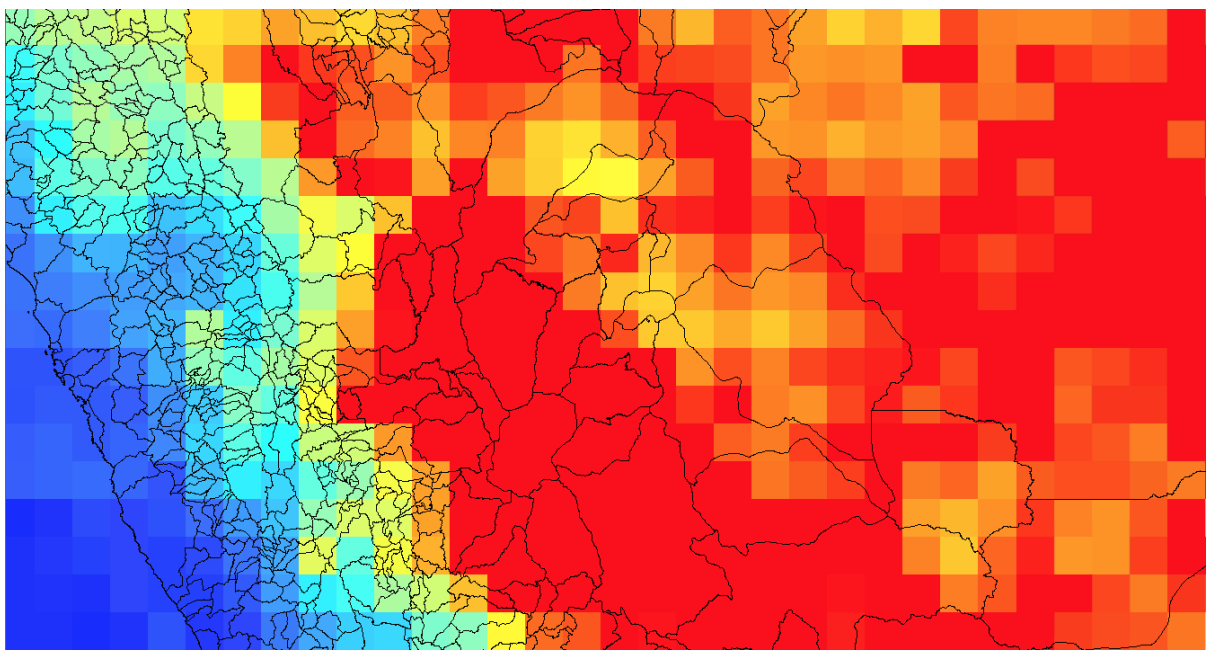
(image source)

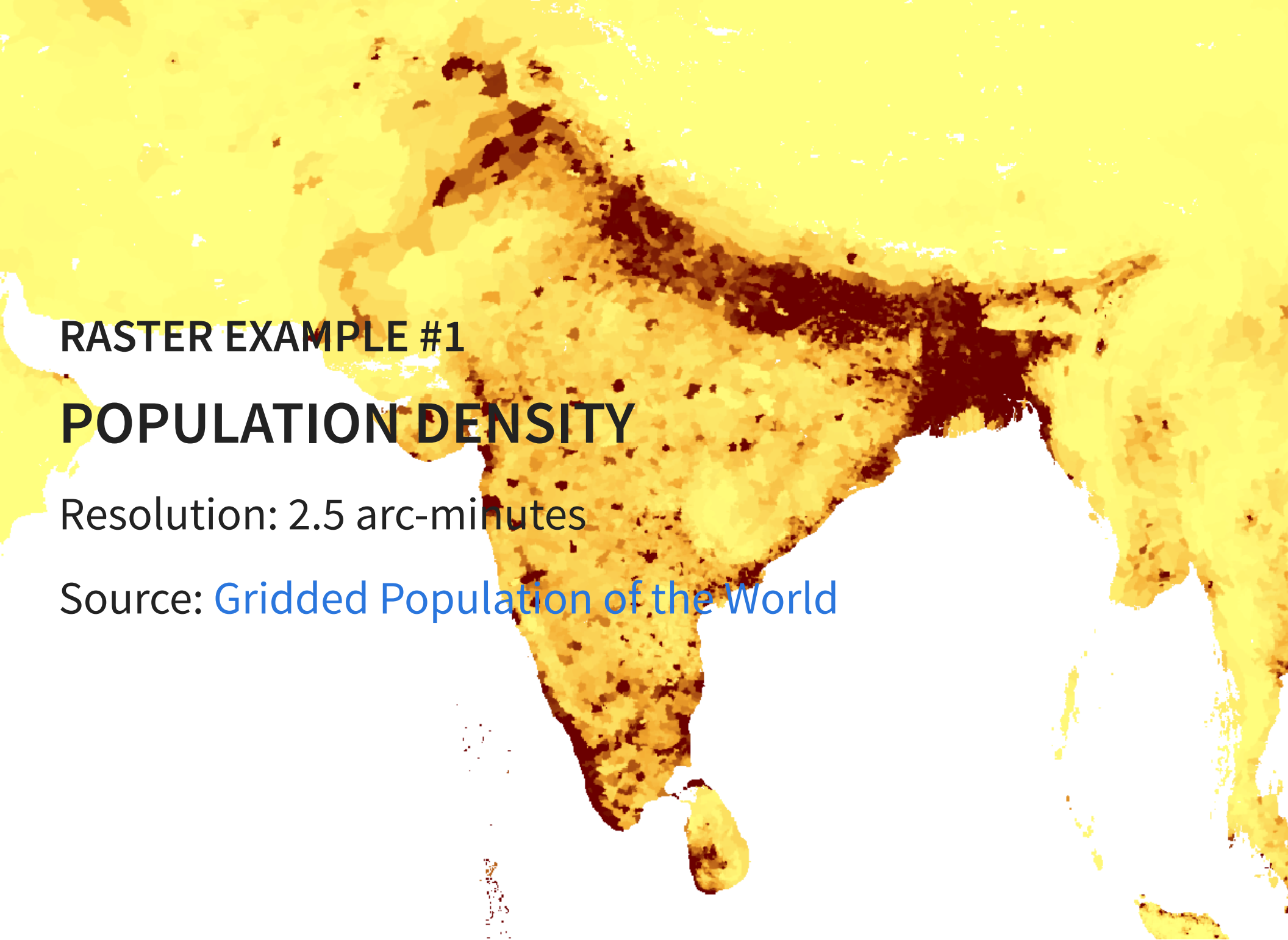


RASTER DATA

Divides the earth surface into many "square" cells (or pixels)

Each cell contains one value



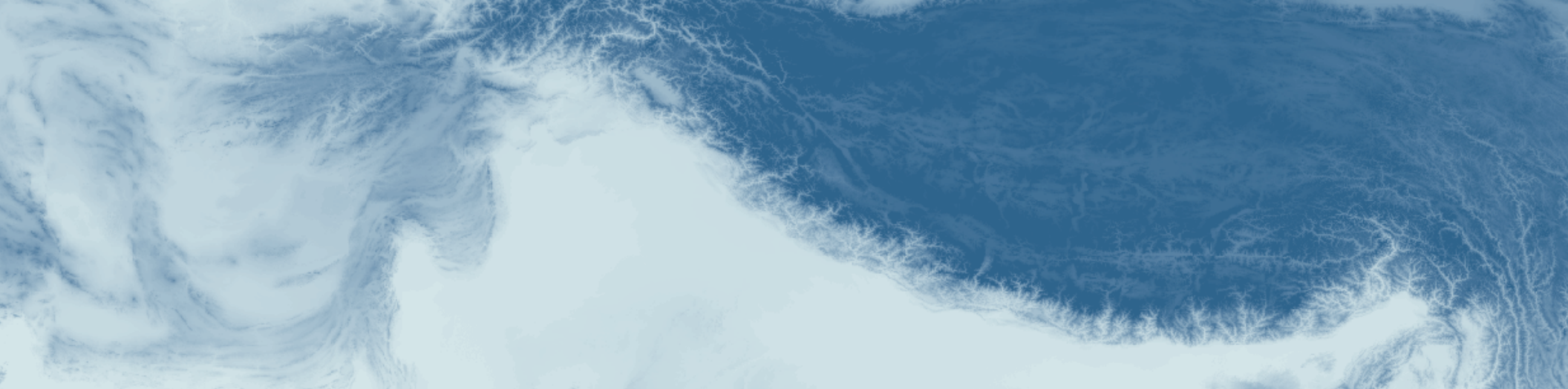


RASTER EXAMPLE #1

POPULATION DENSITY

Resolution: 2.5 arc-minutes

Source: Gridded Population of the World

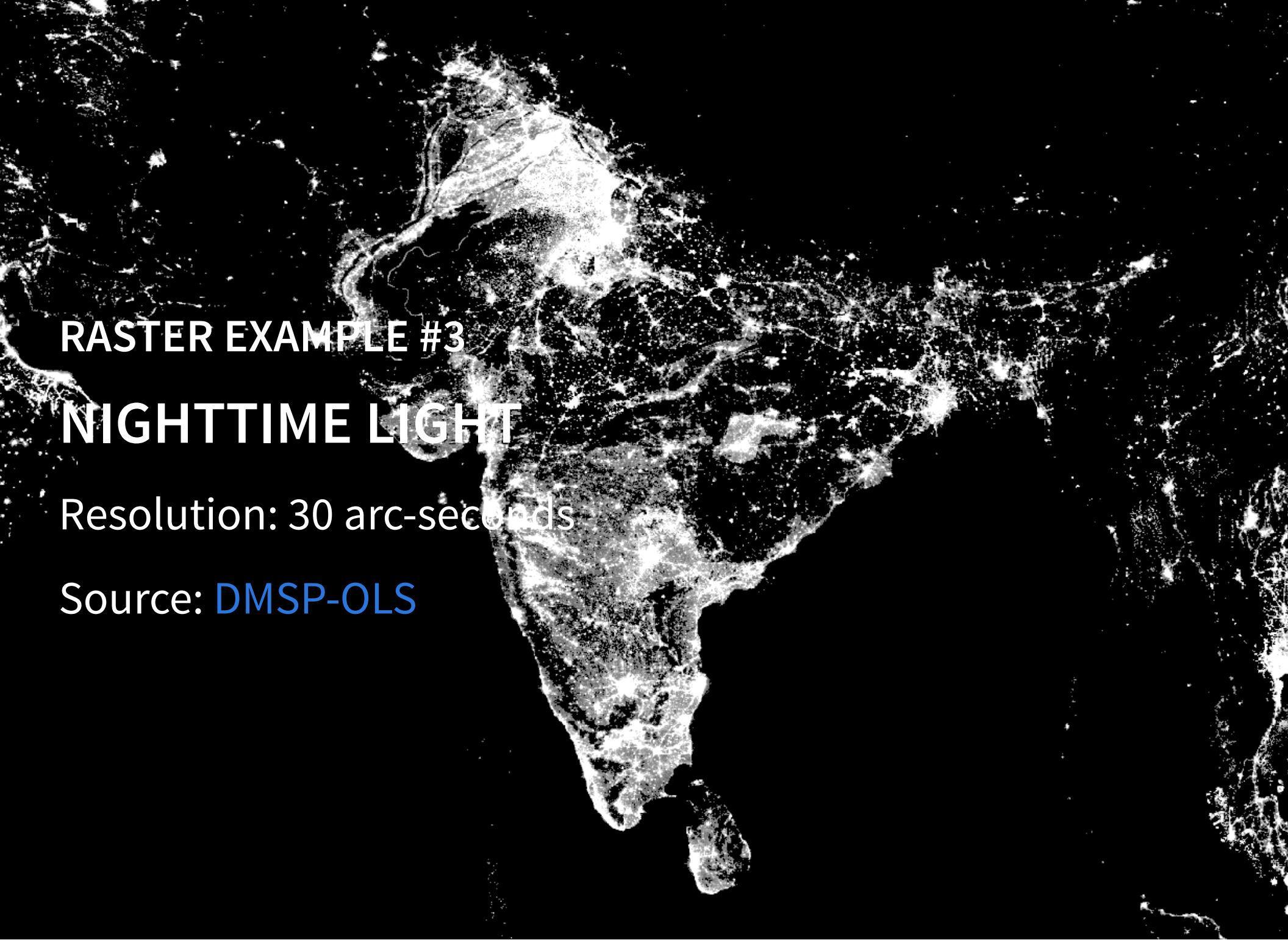


RASTER EXAMPLE #2

ELEVATION

Resolution: 30 arc-seconds

Source: STRM30


RASTER EXAMPLE #3
NIGHTTIME LIGHT
Resolution: 30 arc-seconds
Source: DMSP-OLS

1.2 COORDINATE SYSTEMS

Earth is a sphere (approximately)

Various ways to *two-dimensionally* represent Earth

Each way corresponds to a **coordinate system**

- Also called "spatial reference" or "map projection"

WHY IMPORTANT?

To merge different spatial datasets accurately

cf. Apple Map did this wrong when it was launched in 2012

WHY IMPORTANT? (CONT.)

To calculate distance and surface area properly

GEOGRAPHIC COORDINATE SYSTEMS

Each location is coded by angle from earth center

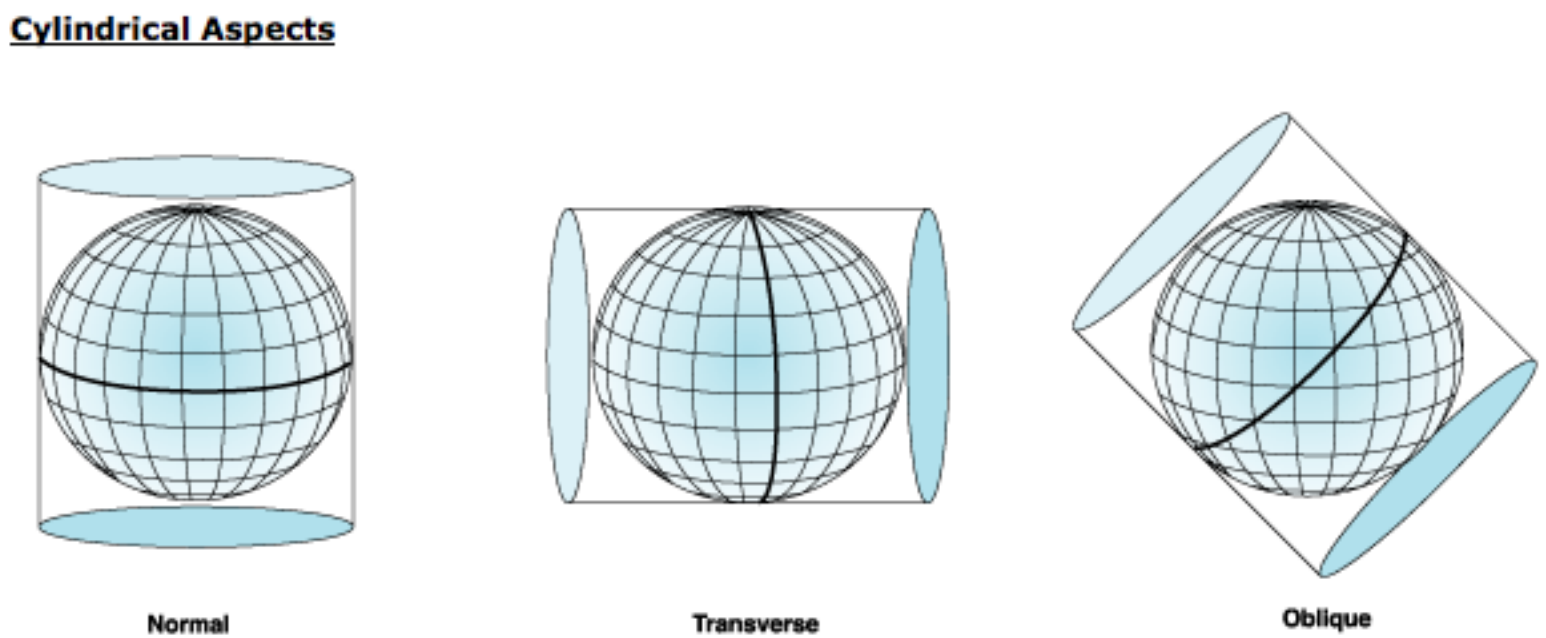
e.g. Stockholm: 59.3293° N / 18.0686° E

Most popular: **WGS 1984**

PROJECTED COORDINATE SYSTEMS

Earth surface is projected by "light" from earth center on:

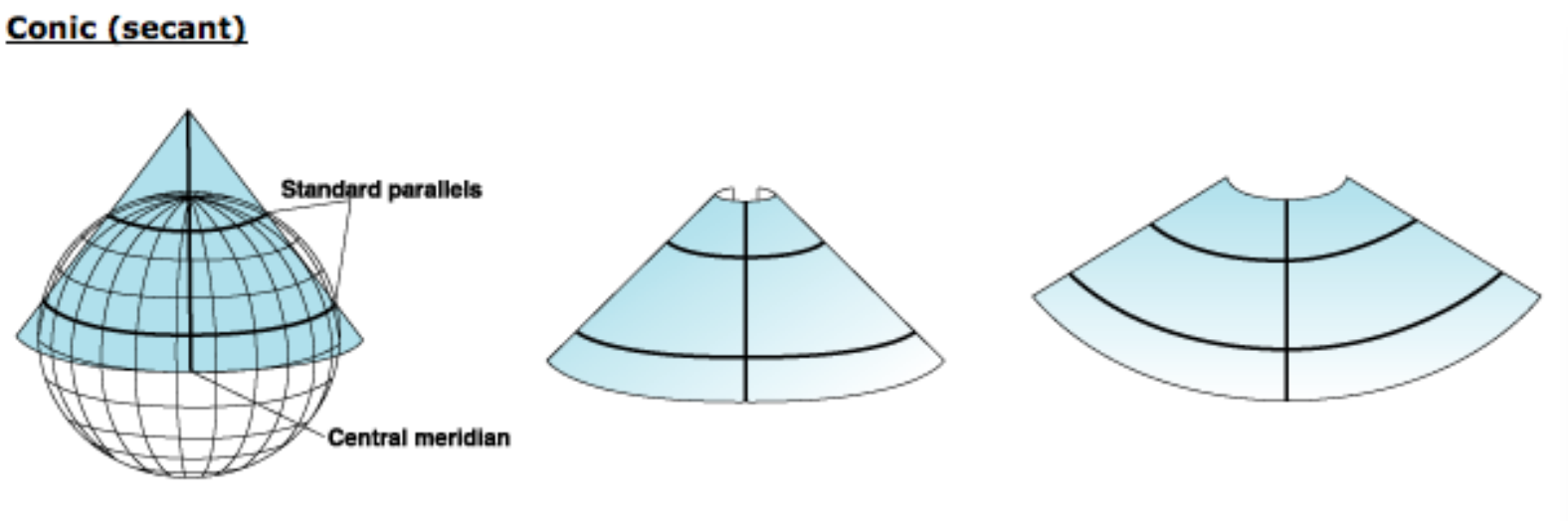
Cylinder



PROJECTED COORDINATE SYSTEMS

Earth surface is projected by "light" from earth center on:

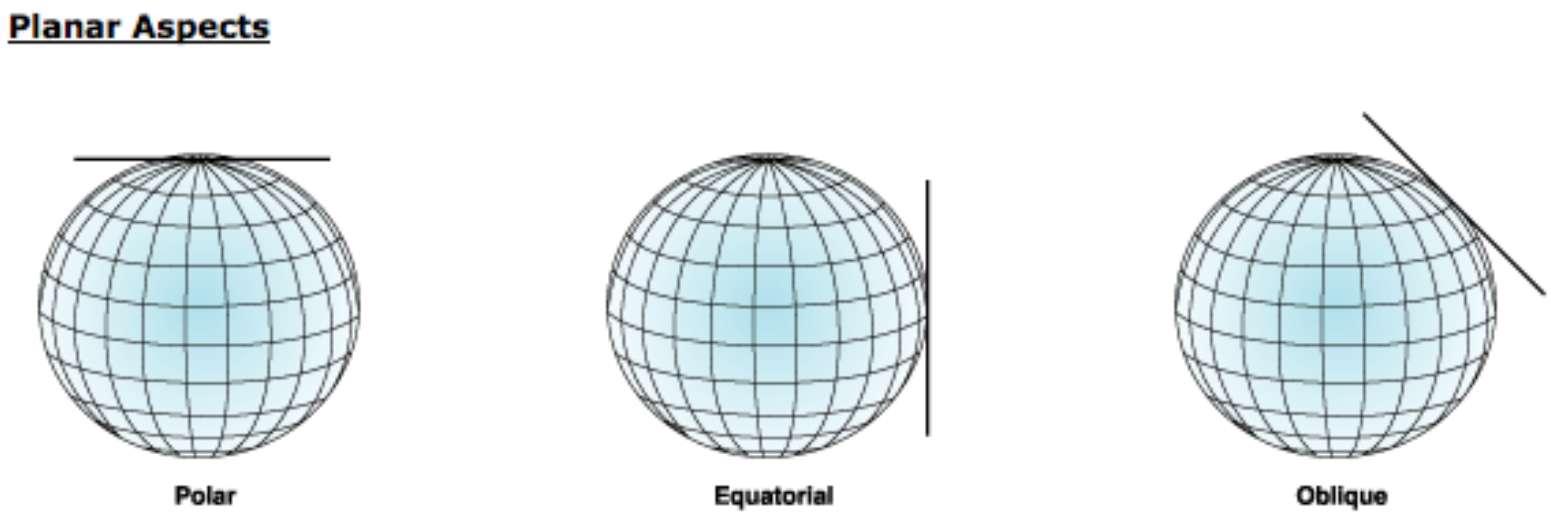
Cone



PROJECTED COORDINATE SYSTEMS

Earth surface is projected by "light" from earth center on:

Plane



PROJECTED COORDINATE SYSTEMS (CONT.)

Each location: coded in *meters* from a certain origin

village	hhn	plot	Xcoord	Ycoord
1	17	1	787877.6	644354.5
1	17	2	788041.2	644449.1
1	17	3	788041.3	644207.7
1	23	5	788942.7	644312.3
1	23	7	788942.7	644312.3
1	27	2	787733	644638.4
1	31	5	788631.2	644102.7
1	34	2	788020.9	644384.6
1	34	4	788008.4	644230.5
1	44	2	786556.3	643995.7
1	44	3	786452.8	644015.9
1	59	2	787416.6	644368.8
1	63	5	787217.3	644308.3
1	63	6	786832.4	644340.4
2	1	1	807386.6	645925.6

PROJECTED COORDINATE SYSTEMS (CONT.)

Examples (relevant for social scientists):

- UTM projections
- Equal Area projections

We will learn these projections later.

IF YOU WANT TO KNOW MORE:

Map Projections: A Working Manual, by John P. Snyder (U.S. Geological Survey, 1987) (Downloadable for free)

1.3 GIS SOFTWARE

ArcGIS

- Python-friendly
- Buggy; tricky to create map images; Windows only

QGIS

- Free; easy to create map images; compatible with any OS
- Python-unfriendly
- Tutorial: www.qgistutorials.com

⇒ ArcGIS recommended for the ease of use of Python (for replication), at least for now

1.3 GIS SOFTWARE (CONT.)

R

- Textbook: Brunsdon & Comber (2015)
- Tutorial by Nick Eubank

Geopandas

- A Python extension to work on spatial data
- Still under development (as of May 2016)

2. CREATE SPATIAL DATA ON YOUR OWN

Satellite images

Scanned old maps

Point locations

Grid cells

2.1 SATELLITE IMAGES

Images consist of pixels

Map each pixel's "color" into raster value

- By using statistical learning methods

A lot of time (and money to hire experts) needed, though.

2.1 SATELLITE IMAGE DATA (CONT.)

Some satellite images: freely available

- See "15 Free Satellite Imagery Data Sources" by GIS Geography

Examples of constructing data from satellite images

- Measuering Yields from Space
- Deforestation

APPLICATION: BURGESS ET AL. 2012

of districts w/i province ↑ ⇒ Deforestation ↑

Theory:

- Each district govt official engages in Cournot competition in selling (illegal) logging permits
- More districts ⇒ More supply of illegal permits

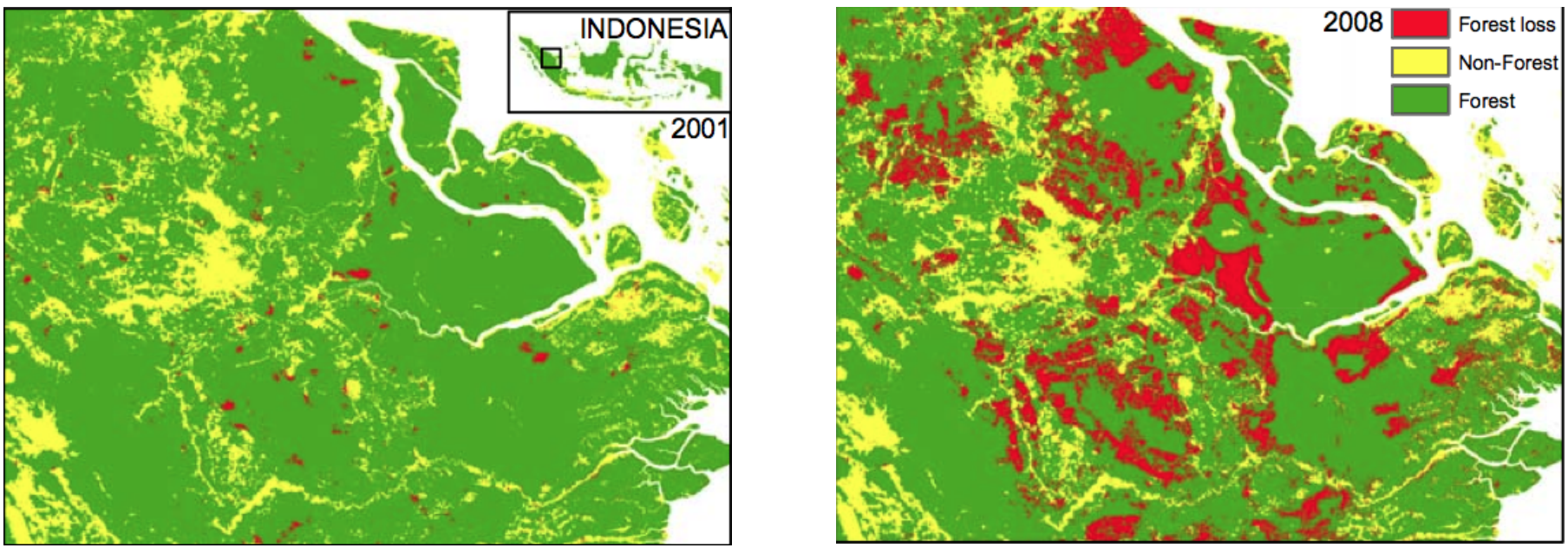
APPLICATION: BURGESS ET AL. 2012 (CONT.)

Cannot rely on official stats of logging

⇒ Use satellite images

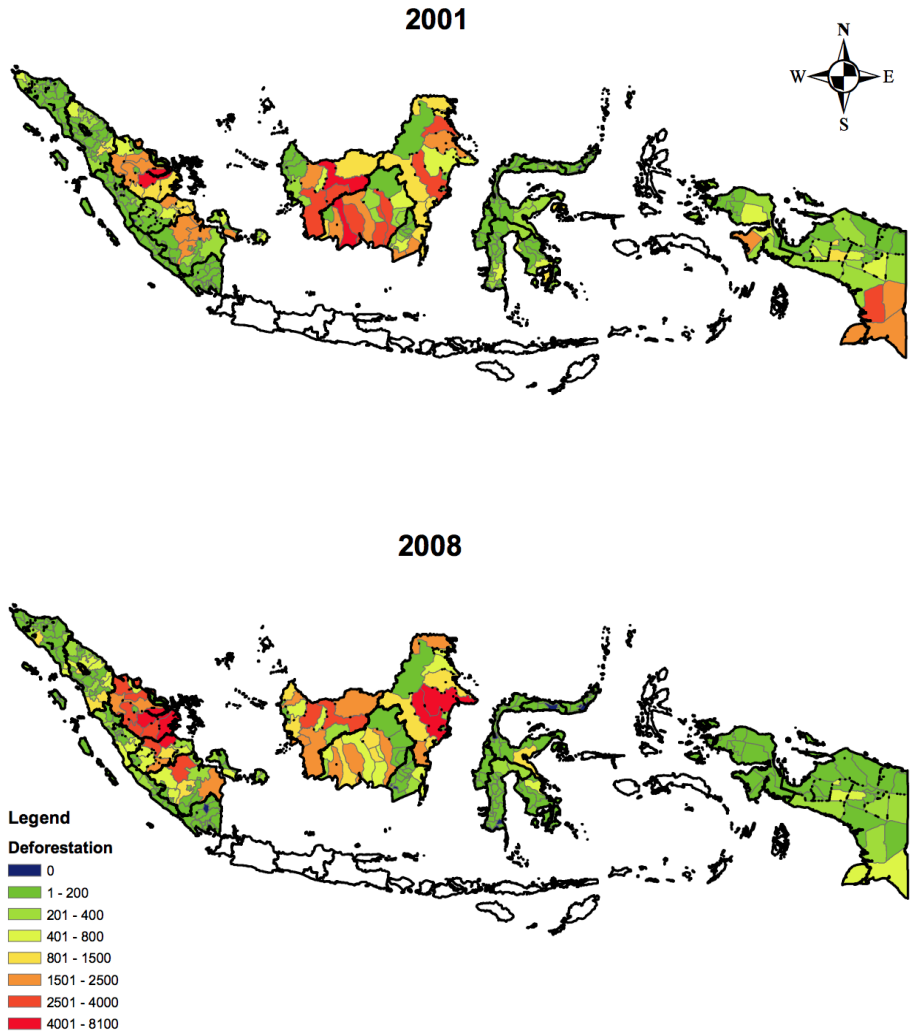
- Spatial resolution: 250m x 250m pixel
- Data: electromagnetic radiation strength in 36 bands of spectrum
- Develop algorithm to convert radiation patterns to forest coverage

PIXEL-LEVEL DATA ON DEFORESTATION



(Figure I of Burgess et al. 2012)

DISTRICT-LEVEL DATA ON DEFORESTATION



(Figure II of Burgess et al. 2012)

2.2 SCANNED OLD MAPS

First, geo-reference the map

- See Yale Map Collection (2009) (pp. 8-10)

Then, create vector data by tracing lines w/ mouse

- See ArcGIS 10: Editing & Creating Your Own Shapefiles (Parts 3-6)

Also time-consuming but feasible with patience

APPLICATION: BURGESS ET AL. 2015

Did Kenyan presidents build more roads for their co-ethnics?

Digitize Michelin maps for Kenya since 1961

Track road network expansion over time

DIGITIZING OLD MAPS

Michelin map in 1961

Digitization and Standardization in GIS

(source: Remi Jedwab's presentation slide)

2.3 POINTS

First, create a table in text format, where:

- Each row: location
- Column 1: longitude (x value)
- Column 2: latitude (y value)
- Other columns: attributes of each location
  - Name
  - Statistics
  - Key (unique ID)
  - Foreign keys (for merging with other data)

HOW TO OBTAIN LONGITUDE & LATITUDE?

GPS receiver

- If you conduct your own survey

Online gazetteer

- If location names are available, search at:
  - Geonames
    - Geonames Tools toolbox automates search
  - Global Gazetteer Version 2.3
  - JRC Fuzzy Gazetteer
- If address is available, use Google Geocoder
  - Stata ado `geocode3` automates search

2.3 POINTS (CONT.)

To convert the text file table into point vector data in ArcGIS, use:

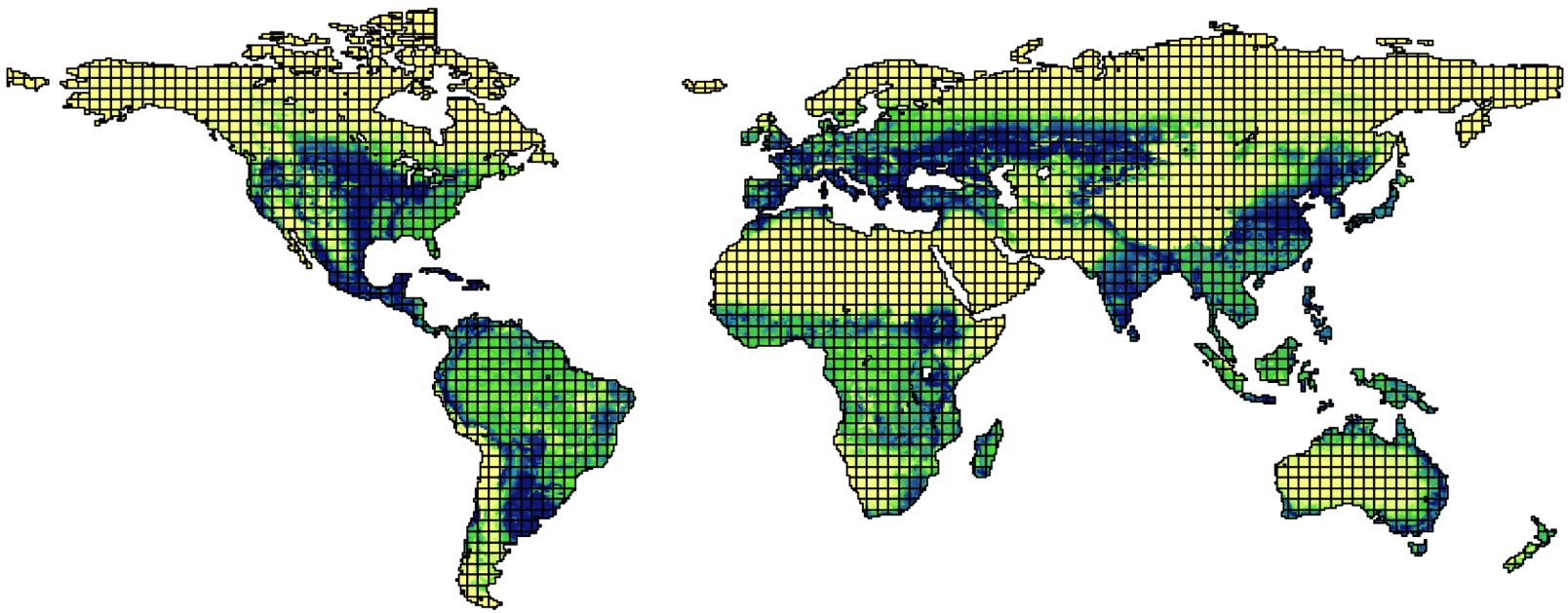
- Make XY Event Layer
- Copy Features

These are the examples of **geo-processing tools**

PYTHON CODE FOR CREATING POINT FEATURES

```
import arcpy

input_table = "coordiates.txt"
output_shp = "points.shp"

varname_x = "longitude"
varname_y = "latitude"

coordinate_system = arcpy.SpatialReference(4326)

arcpy.MakeXYEventLayer_management(
        input_table, varname_x, varname_y,
        "Layer", coordinate_system)
arcpy.CopyFeatures_management("Layer", output_shp)
```

2.4 GRID CELLS



Any size of grid cell polygons can be created in ArcGIS

2.4 GRID CELLS (CONT.)

Useful for:

- Merging weather data
  - WorldClim (Dell, Jones, & Olken 2009)
  - GCPC (Miguel et al. 2004)
  - TOMS air pollution index (Jayachandran 2009)
- Exogenous boundaries

2.4 GRID CELLS (CONT.)

To create grid cell polygons in ArcGIS, use:

- Create Fishnet
- Define Projection

These are also the examples of **geo-processing tools**

PYTHON CODE FOR CREATING POINT FEATURES

```
import arcpy

output_shp = "gridcells25.shp"
cellsize = "2.5"
bottom_left = "-180 -65"
top_right = "180 85"
y_axis = "-180 -55"
coordinate_system = arcpy.SpatialReference(4326)

arcpy.CreateFishnet_management(
        output_shp, bottom_left, y_axis,
        cellsize, cellsize, "0", "0", top_right,
        "NO_LABELS", "", "POLYGON")
arcpy.DefineProjection_management(
        output_shp, coordinate_system)
```

GEO-PROCESSING TOOLS

Master ArcGIS = Know which geo-processing tools to use

Takes vector/raster data as inputs

Most will create new vector/raster data

- Some tools just overwrite the input data

Can be executed in Python

- Should be, for replication

PLAN FOR REST OF THIS LECTURE

Introduce each geo-processing tool

Demonstrate how it's used by economists

3. MERGE SPATIAL DATASETS

1. Spatial Join
2. Intersect + Dissolve
3. Zonal Statistics as Table

3.1 SPATIAL JOIN

Add new variables from a second vector data

Based on **location**

- Not on key variables as in Stata's `merge`

3.1 SPATIAL JOIN (CONT.)

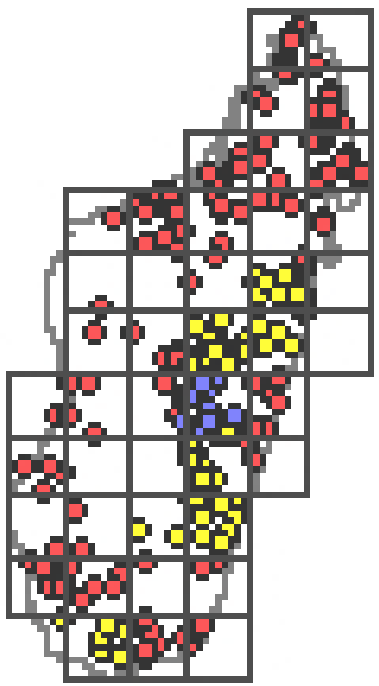
One useful application: merge with weather data

Weather data: available at grid cell level

- No information on country, province, district etc.

3.1 SPATIAL JOIN (CONT.)

⇒ Spatial Join specifies which grid cells are relevant for each observation



PYTHON CODE FOR SPATIAL JOIN

```
import arcpy

target = "cities.shp"
join = "weather_data_cells.shp"
output = "cities_with_weather_data.shp"

arcpy.SpatialJoin_analysis(
        target, join, output)
```

APPLICATION 1: FEYRER & SACERDOTE (2009)

European colonization ⇒ Economic development?

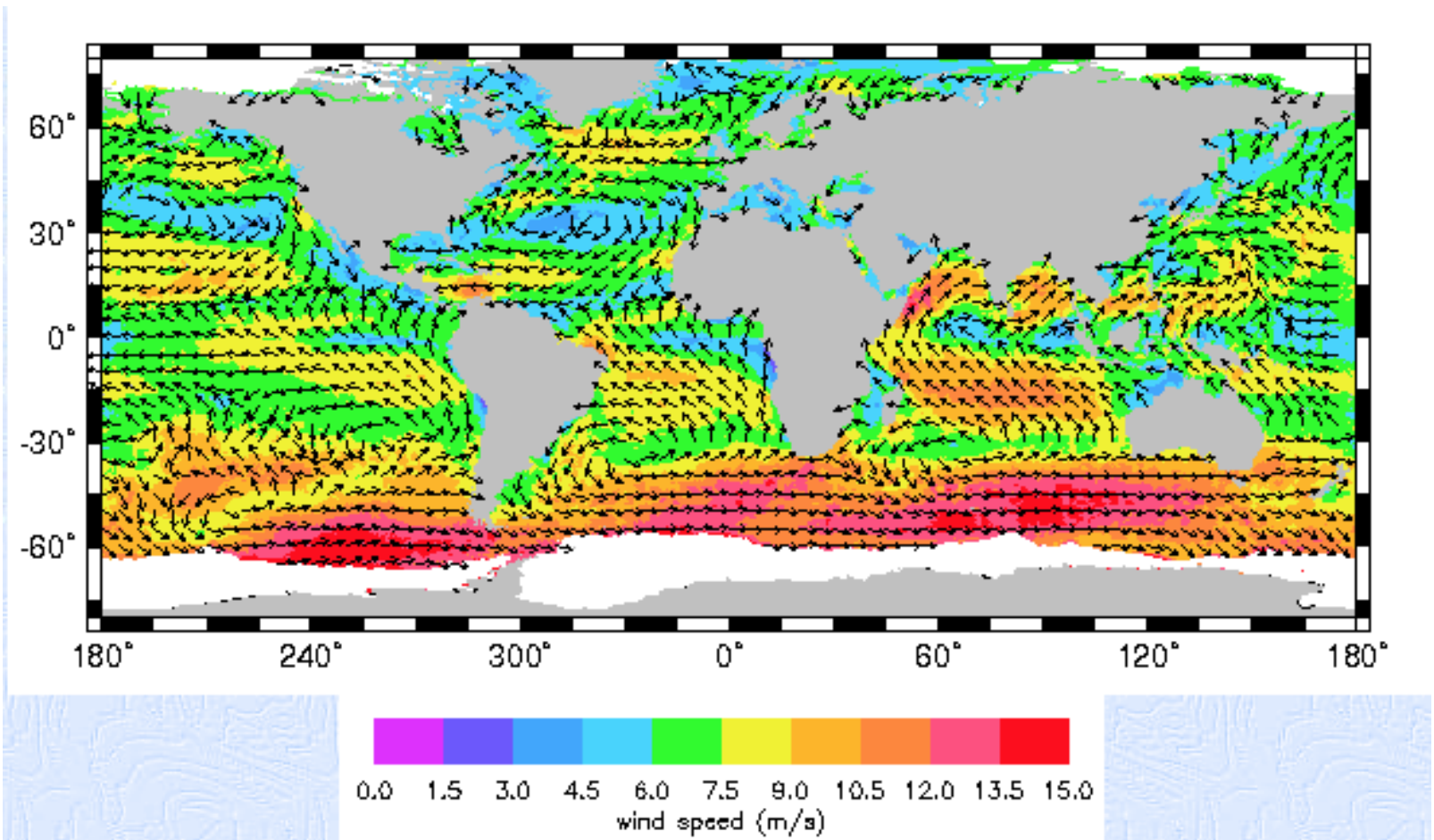
Sample: islands (geo-referenced)

IVs for duration of colonization:

- Mean east-west wind speed
- SD of east-west wind speed

APPLICATION 1: FEYER & SACERDOTE (2009) (CONT.)

Wind data: CERSAT (1° x 1°)



APPLICATION 2: ALSAN (2015)

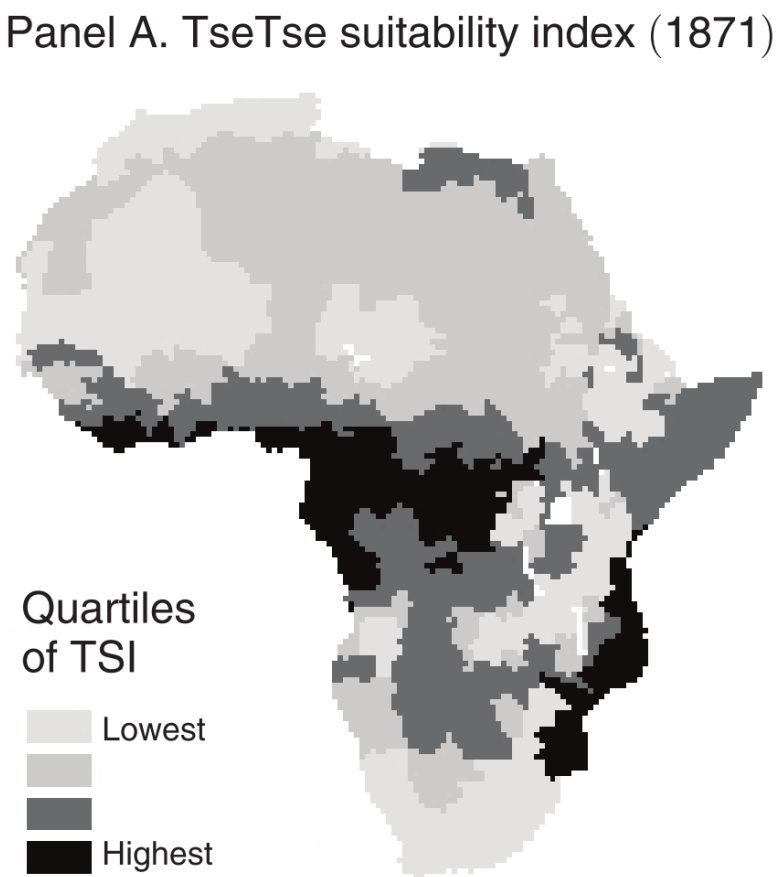
Tsetse flies ⇒ Africa's underdevelopment?

Weather in 1871 at 2° x 2° resolution

Temperature & humidity fed into a model to predict Tsetse fly survival

APPLICATION 2: ALSAN (2015) (CONT.)

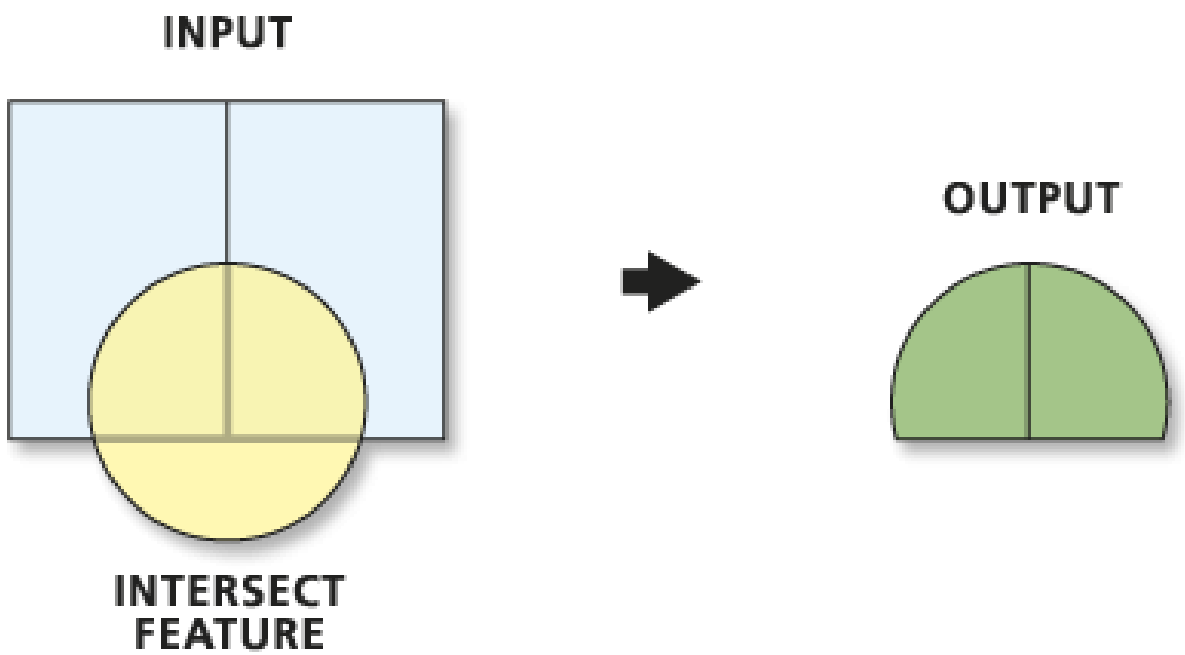
Matched with Ethnographic Atlas



(Figures 5A and 3A of Alsan 2015)

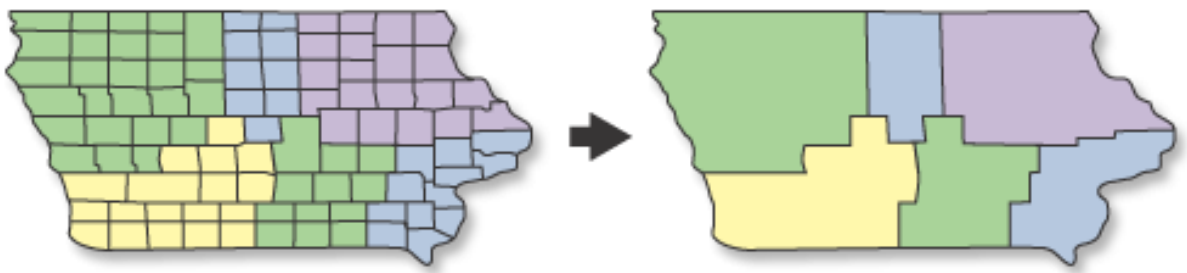
3.2 INTERSECT + DISSOLVE

Intersect: Creates intersection features



3.2 INTERSECT + DISSOLVE (CONT.)

Dissolve: Combines features by key variables



Can also create summary statistics

- Stata's `collapse`

3.2 INTERSECT + DISSOLVE (CONT.)

Can calculate # of polygons/polylines w/i zone polygon

```
import arcpy

inFeatures = ["counties.shp", "streams.shp"]
intersectOutput = "streams_by_country.shp"
dissolve_field = "county_id"
outFeatures = "counties_number_of_streams.shp"

arcpy.Intersect_analysis(
        inFeatures, intersectOutput)
arcpy.Dissolve_management(
        intersectOutput, outFeatures,
        dissolve_field, "COUNT")
```

APPLICATION 1: HOXBY (2000)

Competition ⇒ School quality ↑?

IV: **# of streams** w/i city

- More streams ⇒ More school districts

APPLICATION 2: BAI AND JIA (2016)

In early 20c, Imperial China abolished 1300-year-old civil service exams

Prefectures w/ higher quota ⇒ more uprisings during the 1911 Revolution

IV: **# of streams** w/i prefecture

- Quota depends on # of counties w/i prefecture

3.3 ZONAL STATISTICS AS TABLE

Calculates sum stat of raster values w/i zone

Zone: defined by polygon or polyline

- Stata's `collapse` by zone, executed on raster data

3.3 ZONAL STATISTICS AS TABLE (CONT.)

- Mean / Standard deviation
- Min / Max / Range
- Sum
- Count

For integer raster:

- Median
- Variety (# of unique values)
- Majority (most frequent value)
- Minority (least frequent value)

3.3 ZONAL STATISTICS AS TABLE (CONT.)

If unit of analysis is point, use either:

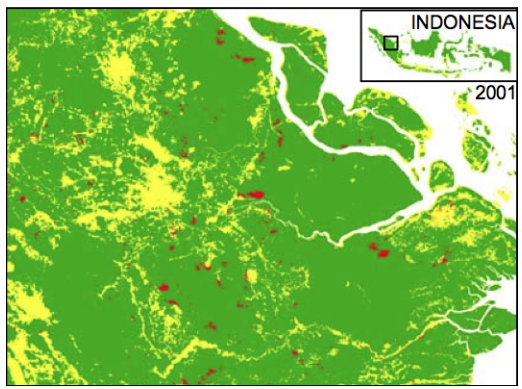
- **Extract Multi Values To Points**
- **Buffer** + Zonal Statistics as Table

PYTHON CODE FOR ZONAL STATISTICS

```
import arcpy
arcpy.CheckOutExtension("spatial")

# Inputs
elevation = "srtm30.dem"
districts = "district.shp"
# Intermediate
zonalstat = "zonalstat.dbf"
# Outputs
mean_elevation = "mean_elevation.xls"

arcpy.gp.ZonalStatisticsAsTable_sa(
        districts, "dist_id", elevation,
        zonalstat, "DATA", "MEAN")
arcpy.TableToExcel_conversion(
        zonalstat, mean_slope)
```

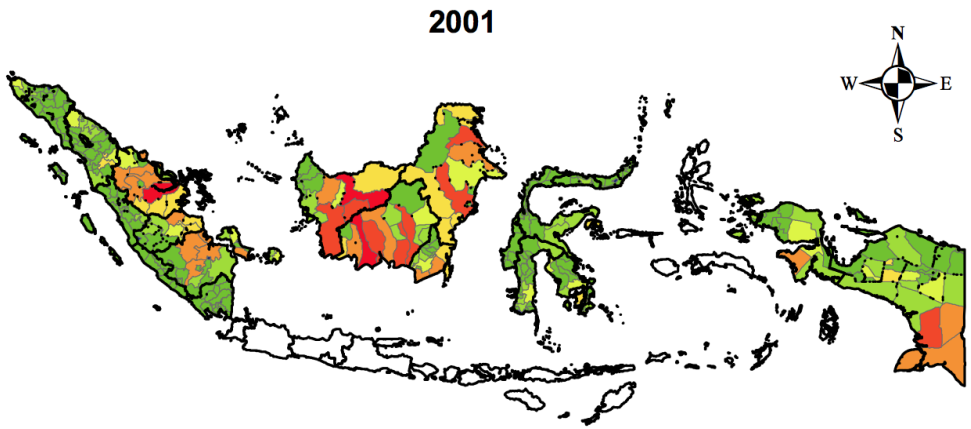
APPLICATION 1: BURGESS ET AL. 2012

Pixel-level



⇒

District-level



APPLICATION 2: NIGHTTIME LIGHT STUDIES

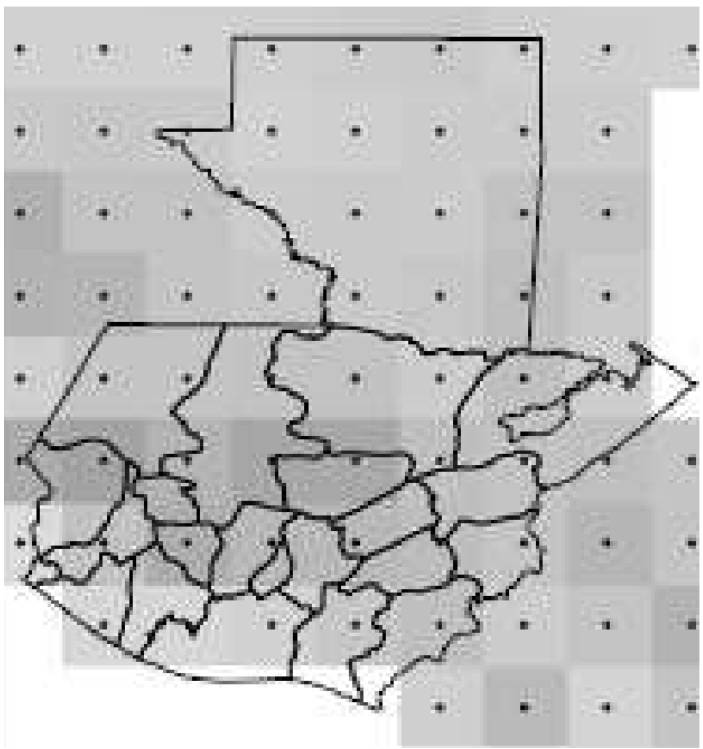
Obtain the mean cell value by:

- Country (Henderson et al. 2012)
- Provinces (Hodler & Raschky 2014)
- Electoral districts (Baskaran et al. 2015)
- Ethnic homelands (Michalopoulos & Papaioannou 2013 / 2014, Alesina et al. 2016)

APPLICATION 3: DELL ET AL. 2012

Temperature ↑ ⇒ Economic growth ↓?

Use population-weighted average temperature



(Figure 2.3 of Dell 2009)

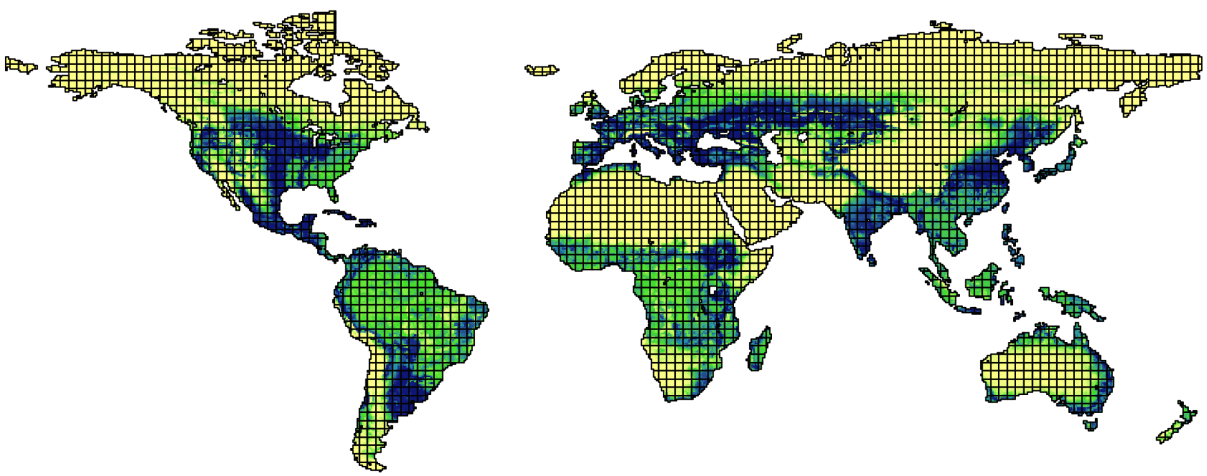
APPLICATION 4: MICHALOPOULOS (2012)

Geographic diversity ⇒ Ethnic diversity?

Empirical challenge:

- Endogeneity of geographic diversity by country formation

⇒ 2.5°x2.5° grid cells as units of analysis



APPLICATION 4: MICHALOPOULOS (2012) (CONT.)

How to measure ethnic & geographic diversity, then?

ArcGIS helps:

- Intersect + Dissolve ⇒ # of languages spoken
- Zonal Statistics ⇒ S.D. of elevation / land quality

APPLICATION 4: MICHALOPOULOS (2012) (CONT.)

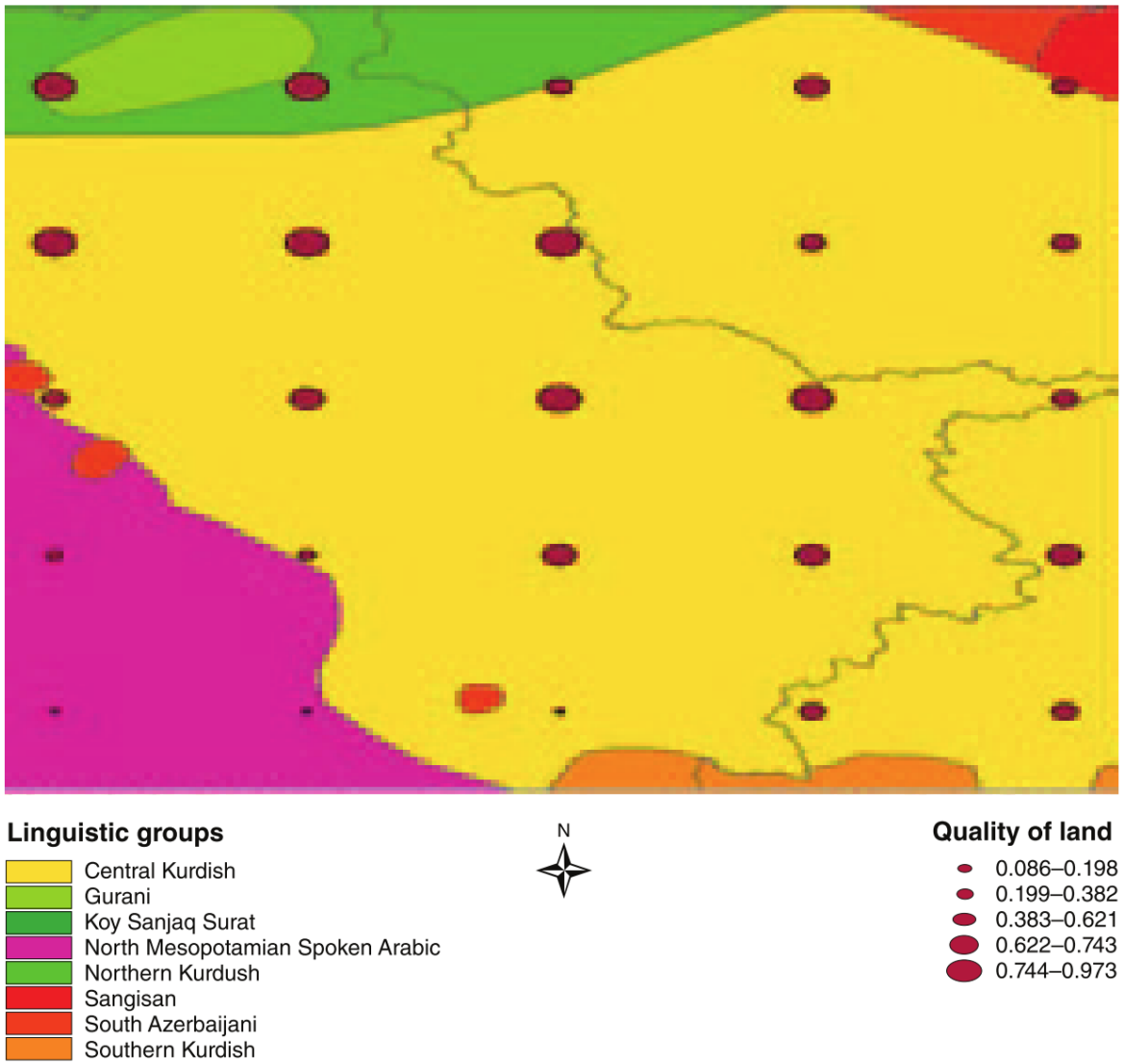


FIGURE 7. EXAMPLE OF A VIRTUAL COUNTRY

4. ELEVATION

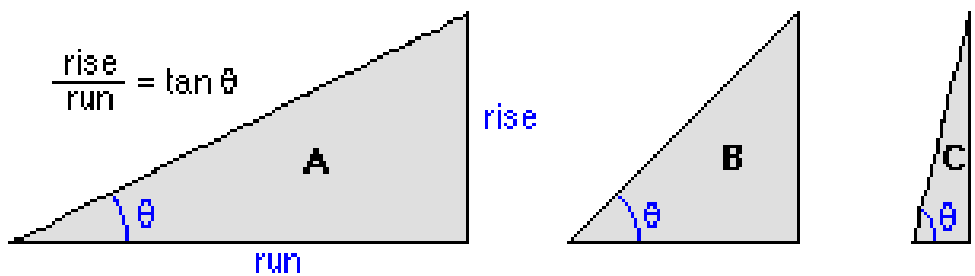
1. Slope
2. Slope + Reclassify
3. Irregular Terrain Model

4.1 SLOPE

Returns either θ or $\tan\theta$ in diagram below

Degree of slope = θ

Percent of slope = $\frac{\text{rise}}{\text{run}} * 100$



Degree of slope =	30	45	76
Percent of slope =	58	100	373

4.1 SLOPE (CONT.)

$\tan\theta$ for cell e is obtained by

a	b	c
d	e	f
g	h	i

$$\Rightarrow \quad \tan\theta = \sqrt{(dz/dx)^2 + (dz/dy)^2}$$

where

$$\frac{dz}{dx} = \left[\frac{c + 2f + i}{4} - \frac{a + 2d + g}{4}\right]/2$$

$$\frac{dz}{dy} = \left[\frac{a + 2b + c}{4} - \frac{g + 2h + i}{4}\right]/2$$

APPLICATION 1: DINKELMAN (2011)

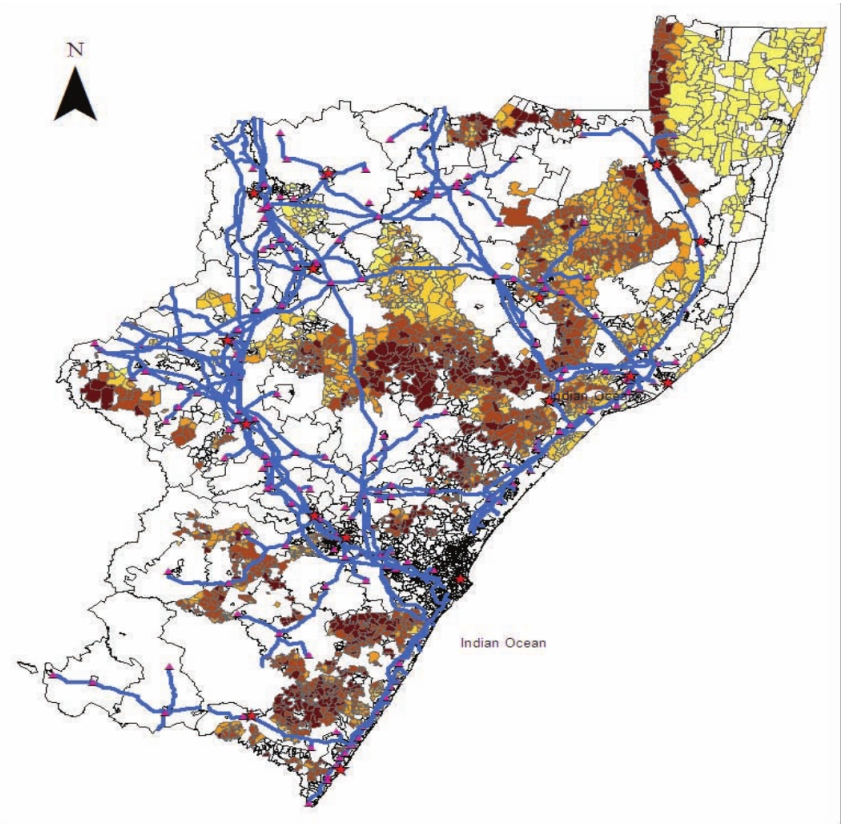
Electrification in South Africa (1996-2001)

$\Rightarrow$ Female labor supply $\uparrow$

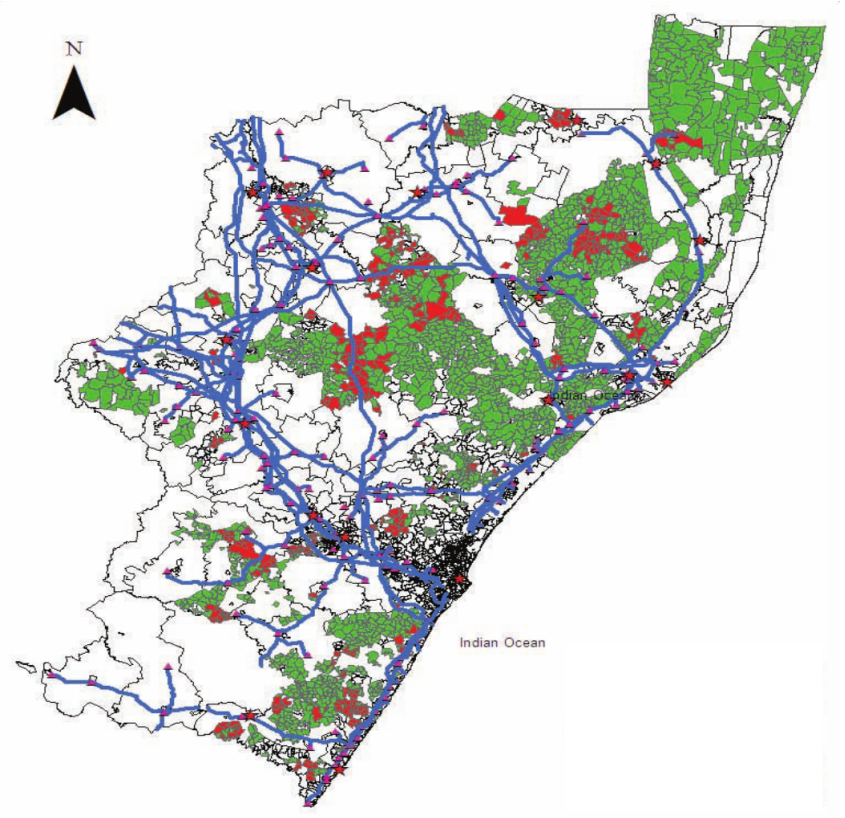
IV: **mean land slope**

- Flat terrain: cheap to lay power lines

APPLICATION 1: DINKELMAN (2011) (CONT.)



Slope (lighter = flatter)



Electrification (red)

APPLICATION 2: QIAN (2008)

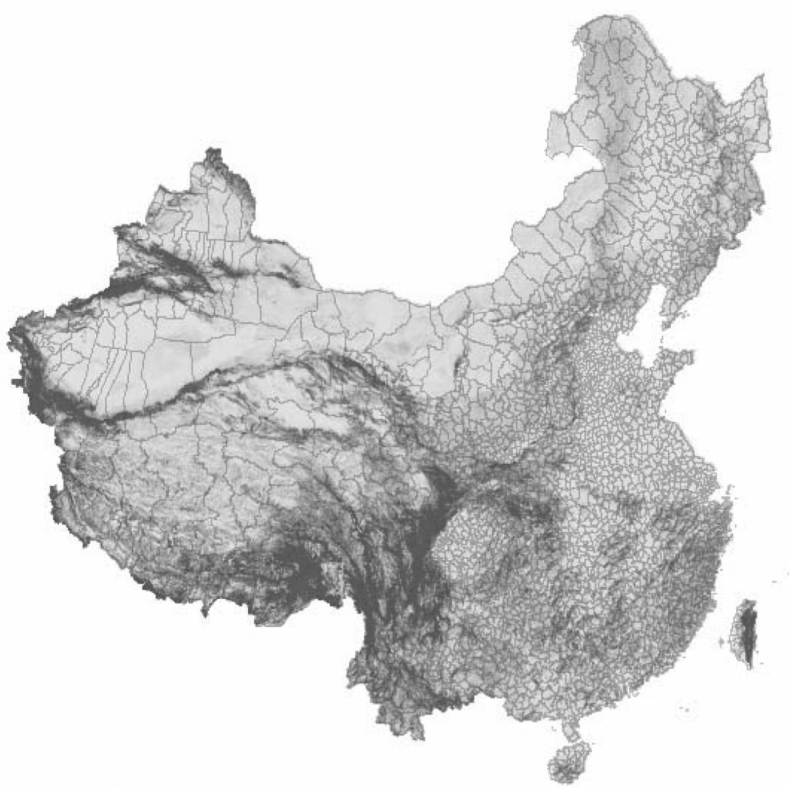
Tea production ↑ in China due to liberalization in 1979

⇒ Male-to-female ratio ↓

IV: **mean land slope**

- Tea grows in hilly terrain

APPLICATION 2: QIAN (2008) (CONT.)



Slope (darker = steeper)

Tea (darker = more)

PYTHON CODE FOR AVERAGE SLOPE

```
import arcpy
arcpy.CheckOutExtension("spatial")

# Inputs
elevation = "srtm30.dem"
districts = "district.shp"
# Intermediates
slope = "slope.tif"
zonalstat = "zonalstat.dbf"

arcpy.gp.Slope_sa(
        elevation, slope, "PERCENT_RISE", "0,000009")
arcpy.gp.ZonalStatisticsAsTable_sa(
        districts, "dist_id", slope,
        zonalstat, "DATA", "MEAN")
```

4.2 SLOPE + RECLASSIFY

Reclassify: Creates categorical raster data

Example: a dummy variable for slope 3-6%

Old values	New values
0 - 3	0
3 - 6	1
6 - 193.229706	0
NoData	NoData

PYTHON CODE FOR SLOPE CATEGORY

```
import arcpy
arcpy.CheckOutExtension("spatial")

# Inputs
elevation = "srtm30.dem"
# Outputs
slope = "slope.tif"
slope_3_6 = "slope3to6.tif"

arcpy.gp.Slope_sa(
        elevation, slope, "PERCENT_RISE", "0,000009")
arcpy.gp.Reclassify_sa(
        slope, "Value",
        "0 3 0;3 6 1;6 193,22999999999999 0",
        slope_3_6, "DATA")
```

APPLICATION 1: DUFLO & PANDE (2007)

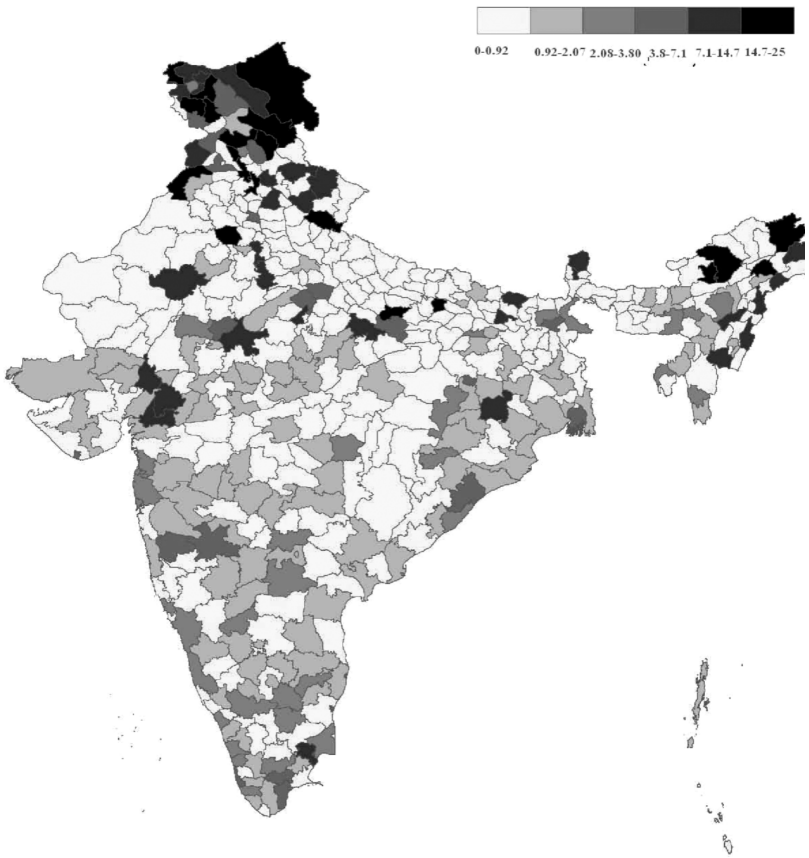
Irrigation dams $\Rightarrow$ Poverty $\downarrow$?

IV: Fraction of river areas in three slope ranges

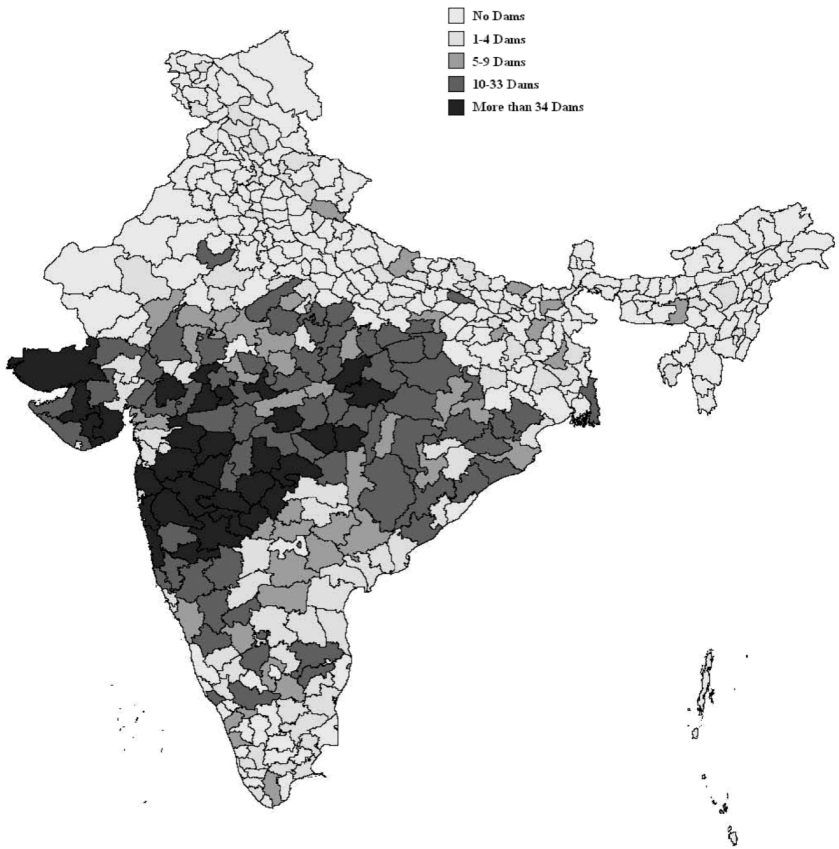
$\Leftarrow$ Easy to build if river slope is:

- Moderate (1.5-3%) for irrigation dams
- Very steep (6%+) for hydroelectricity dams

APPLICATION 1: DUFLO & PANDE (2007) (CONT.)



River slope (darker = steeper)



Dams (darker = more)

APPLICATION 1: DUFLO & PANDE (2007) (CONT.)

Intersect + Dissolve ⇒ River by districts

Slope + Reclassify ⇒ Indicator for each slope range

Zonal Statistics as Table ⇒ Fraction of river areas in each slope range by district

APPLICATION 2: SAIZ (2010)

Measures % of areas with slope > 15% as unsuitability for urban development

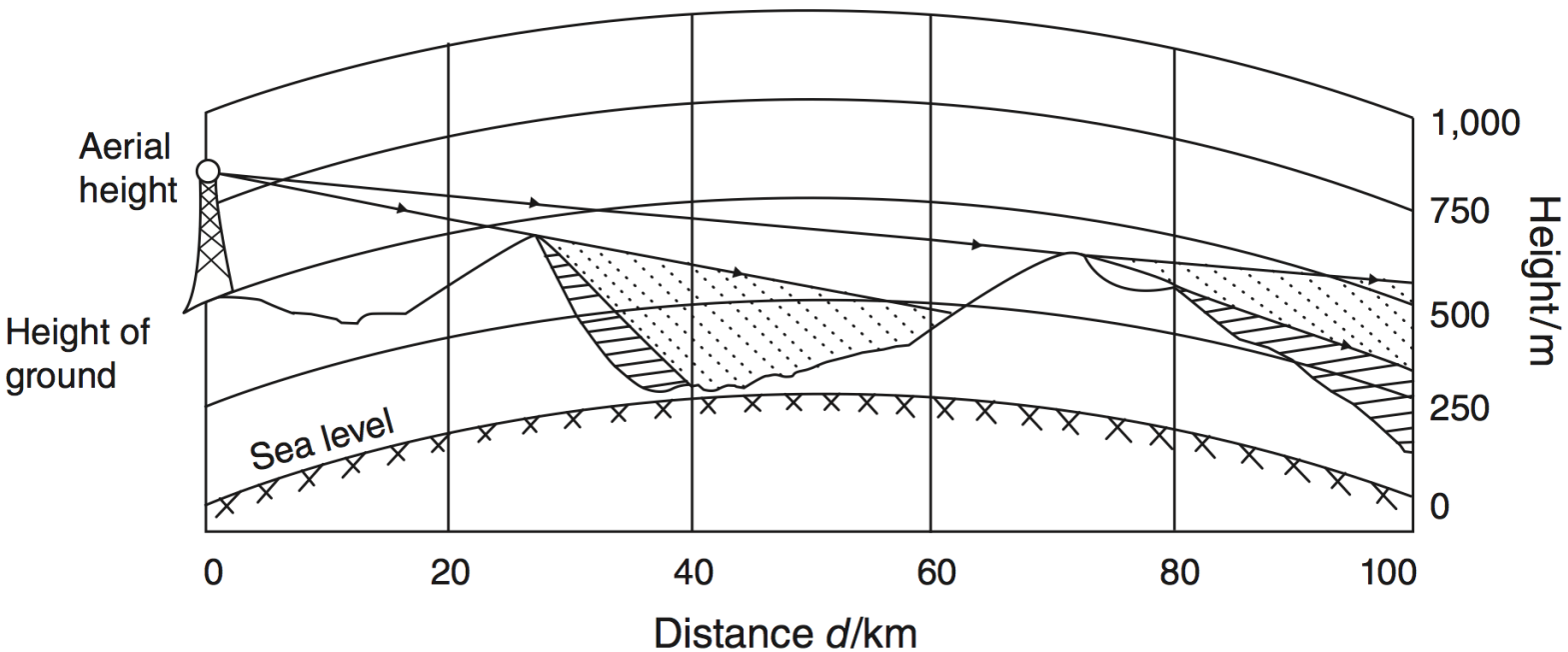
Finds housing supply inelastic in such areas

Slope of over 15%: now often used as geographic constraints to housing supply / urban development

- Diamond (2016), Hariri (2015), Chen & Kung (2013)

4.3 IRREGULAR TERRAIN MODEL

Used by radio/tv enginneers to predict signal reception



(Figure 2 of Olken (2009))

APPLICATION 1: OLKEN (2009)

of TV channels ↑ in Indonesia ⇒ Social capital ↓

IV: **TV signal strength**

APPLICATION 2: YANAGIZAWA-DROTT (2014)

Anti-Hutu radio ⇒ Rwandan genocide incidents ↑

IV: **radio signal strength**