3rd PODER Summer School
New Data in Development Economics

Spatial methods in development economics

Mariaflavia (Nina) Harari
The Wharton School, University of Pennsylvania

28 June 2016
University of Namur

Goals for this lecture

- Selected contributions to the development literature that use spatial data / methods
  - Data sources
  - Share insights on the research process
  - Examples of own research

Roadmap

All empirical data is "spatial" in some sense.

Why should we care about the spatial dimension of the data?

1. Spatial correlation/ dependence in the data affects inference

2. Spatial dimension of the data can be exploited for the identification

3. Spatial patterns are the object of interest

Roadmap

What do economists use spatial data for?

1. Spatial correlation/ dependence in the data affects inference
   - Mapping phenomena in space
2. Spatial dimension of the data can be exploited for the identification
   - Spatial data as controls or instruments
3. Spatial patterns are the object of interest
   - Spatial data as outcomes / proxies
     - Especially in development: a way around the lack of traditional data sources

Roadmap

Outline

- Intro: spatial data
  - Types
  - Sources
- What to do with spatial data
  - Spatial correlation / dependence
  - Spatial data & identification
  - Spatial data as outcomes

Roadmap

Outline

- **Intro: spatial data**
  - **Types**
  - **Sources**
- What to do with spatial data
  - Spatial correlation / dependence
  - Spatial data & identification
  - Spatial data as outcomes

Spatial data

Types of spatial data

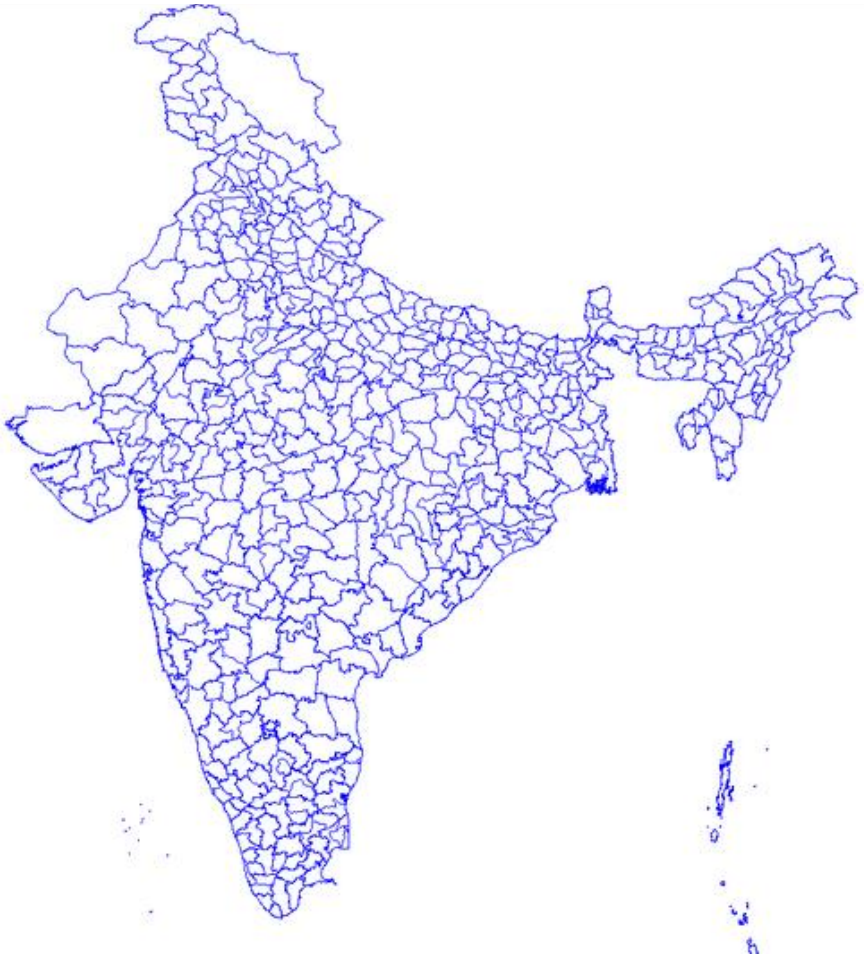
Two ways of organizing spatial information

- **Vector** data: each spatial unit is one of the following
  - Polygon
  - Polyline
  - Points

- **Raster** data: each spatial unit is a pixel in a regular grid, that the Earth surface is divided into
  - Multidimensional raster data: pixels over time

Spatial data

Examples: polygon data



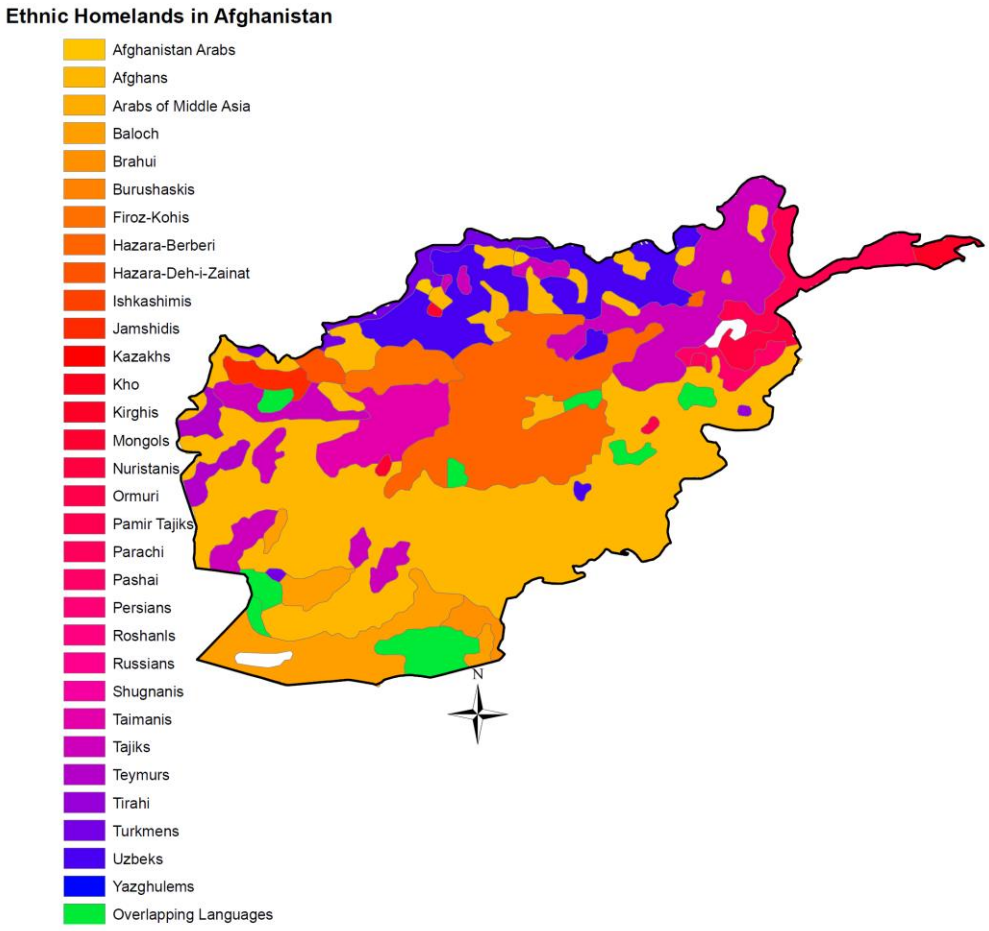
Administrative units

Each polygon =
a district in India

Source: Census of India

Spatial data

Examples: polygon data



Ethnographic boundaries

Each polygon = homeland of a a different ethnic group

Source: Georeferencing of Ethnic Groups (GREG) dataset, from Alesina and Michalopoulos (2105)

Spatial data

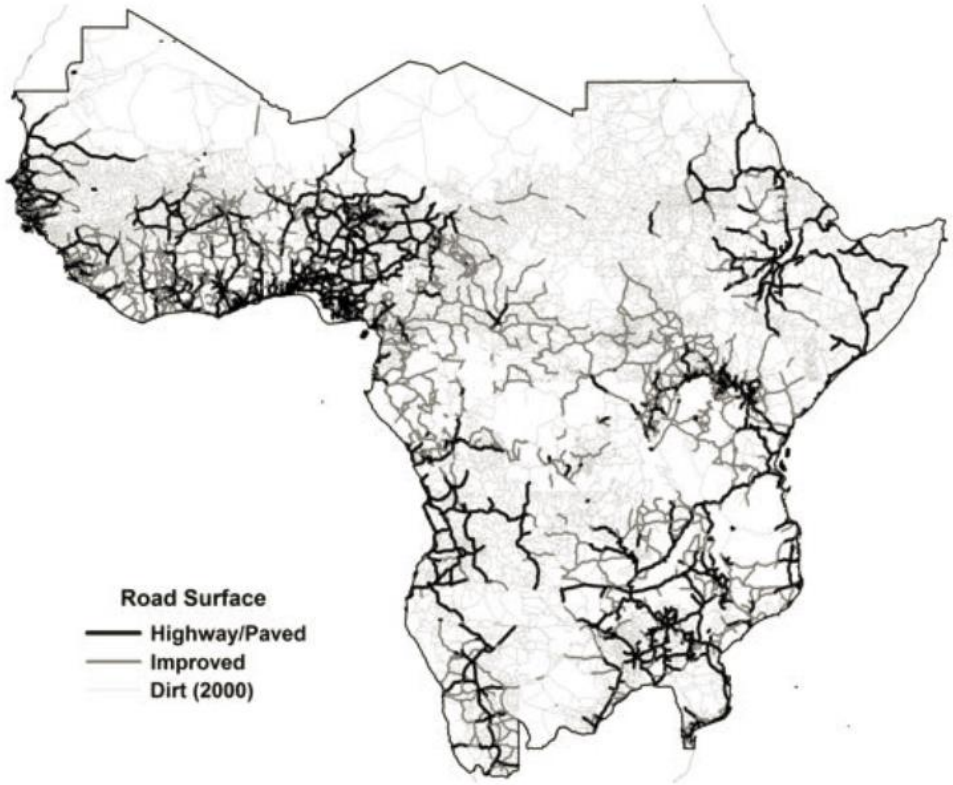
Examples: polyline data

Rivers

E.g. used in Lipscomb and Mobarak (2016)

Spatial data

Examples: polyline data

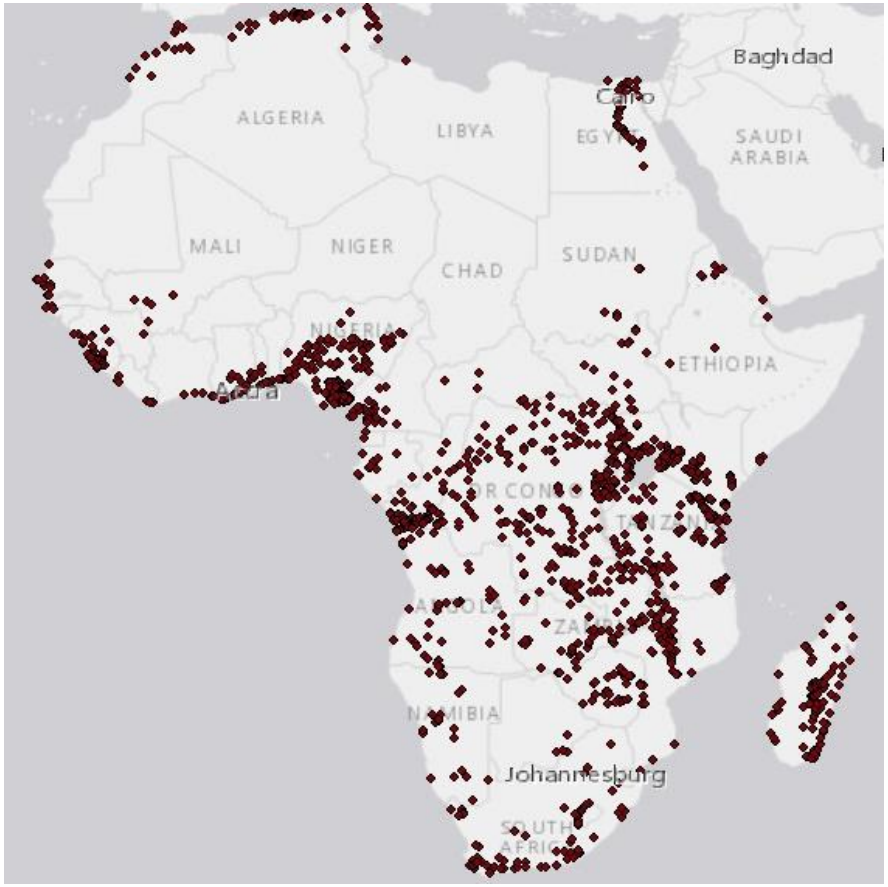


Roads

Source: Michelin maps digitized by Jedwab and Storeygard (2015)

Spatial data

Examples: point data



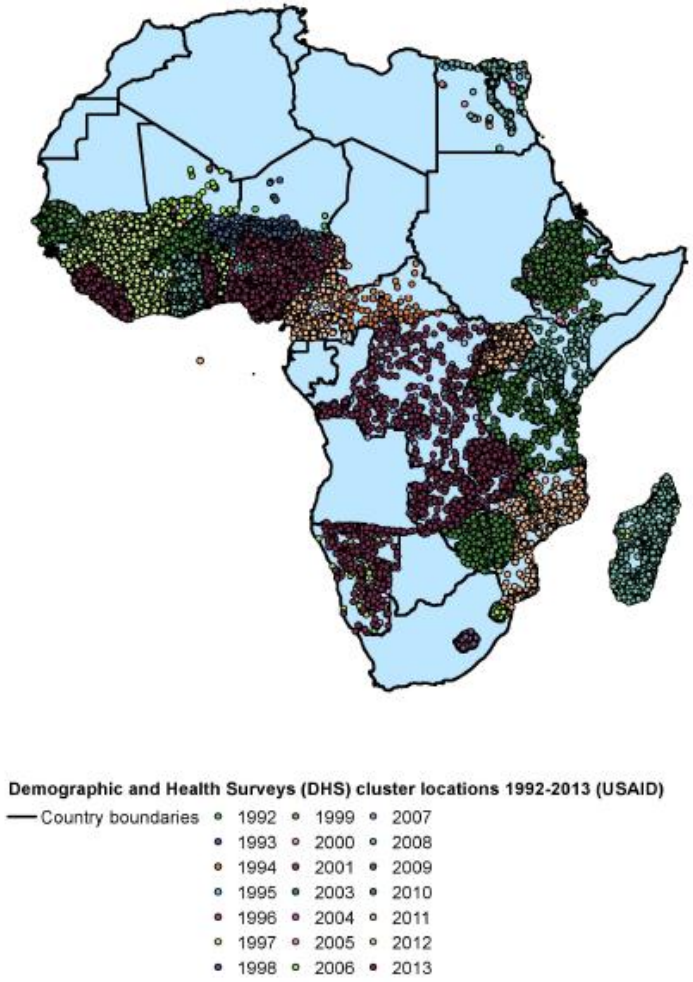
Mission stations in Africa in 1924

Each point = location of a mission

Source: Roome (1924) maps, digitized by Nunn (2010)

Spatial data

Examples: point data



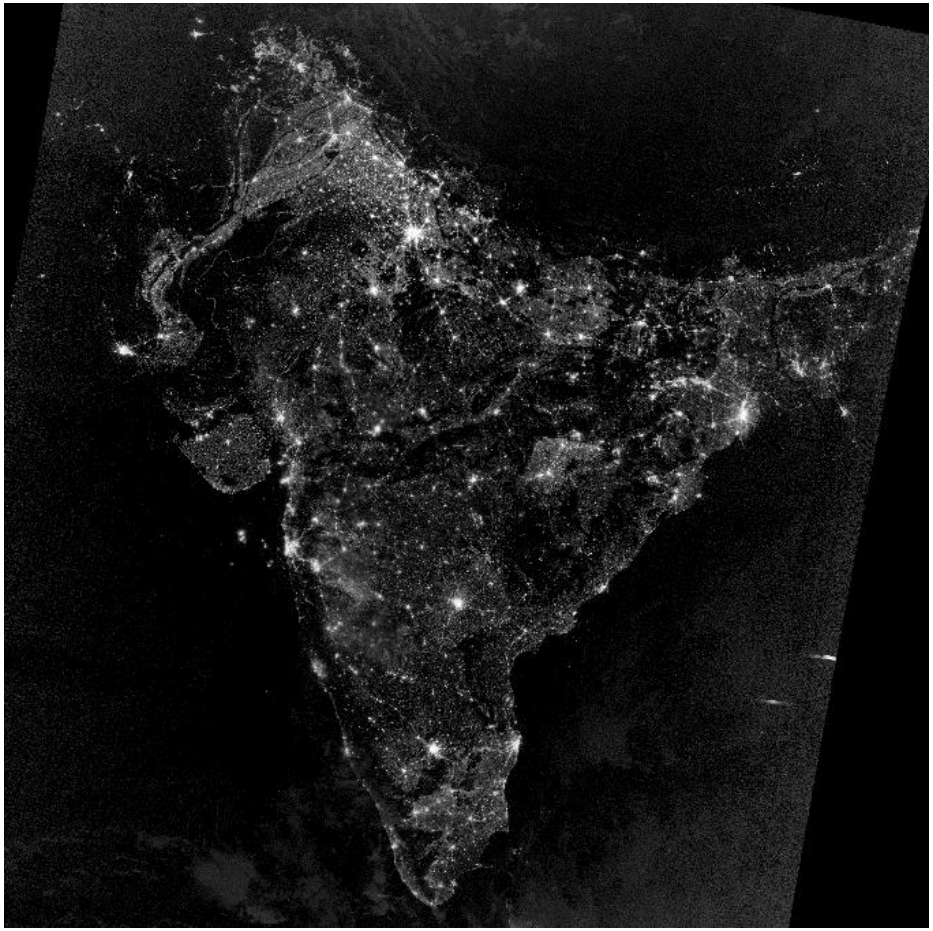
Demographic and Health Survey locations

Each point = location of a DHS cluster

Source: USAID, from Hodler (2016); used in Kudamatsu, Persson and Stromberg (2012)

Spatial data

Examples: raster data



DMSP-OLS Night-time lights dataset

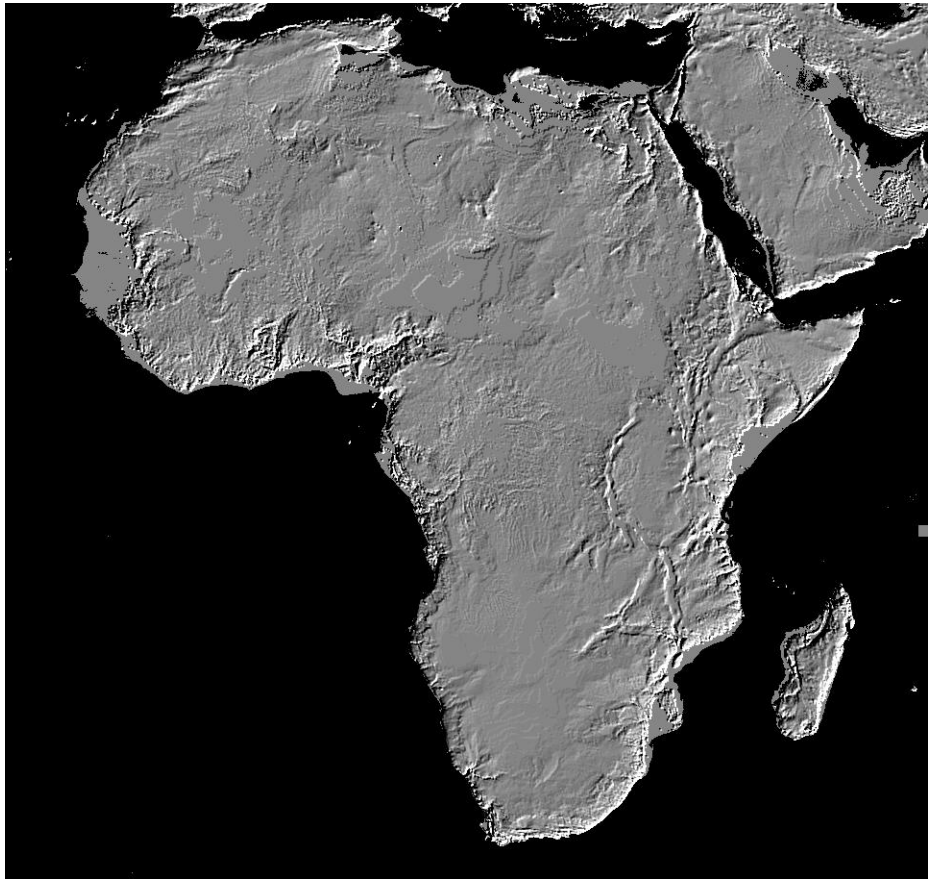
Resolution: 1km

Intensity of earth-based lights in each pixel

Source: National Oceanic and Atmospheric Administration

Spatial data

Examples: raster data



ASTER Digital Elevation Model

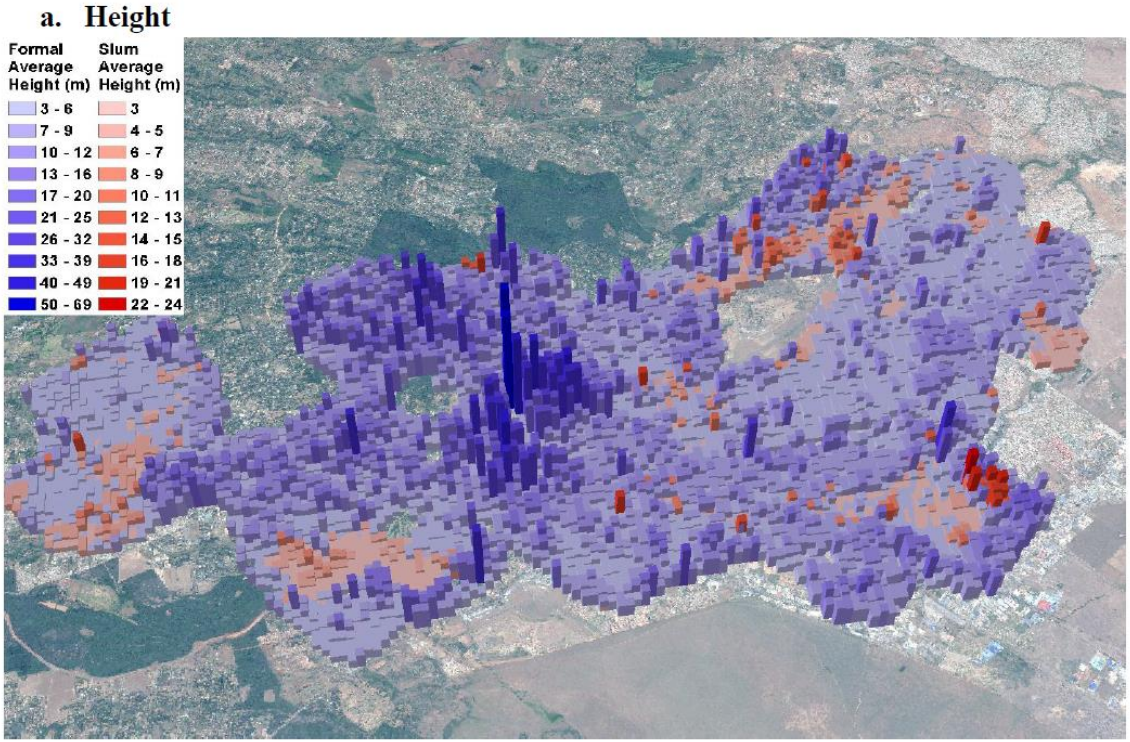
Resolution: 30m

Terrain elevation in each pixel

Source: NASA

Spatial data

Examples: raster data



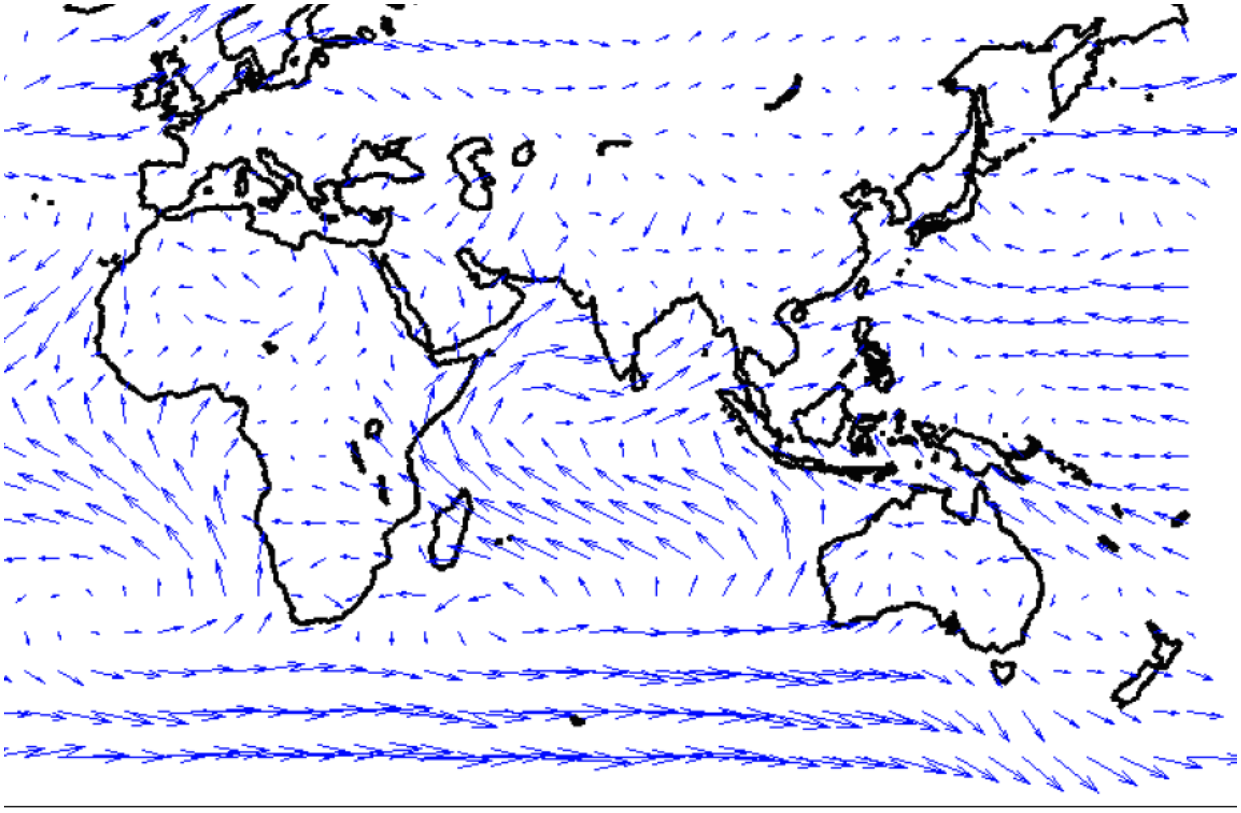
Digital Surface Model of building heights in Nairobi

Resolution: 30m

Source: Henderson et al (2016)

Spatial data

Examples: multidimensional raster data



Gridded climate datasets, e.g. wind patterns

Wind speed and direction in each pixel, over time

Source: CERSAT data, used in Pascali (2015)

Spatial data

Handling spatial data

- Geographical Information Systems (GIS) = tools used for inputting, storing, managing, analyzing and mapping data that have a spatial component

- Not all spatial data comes in GIS-friendly formats
  - Digital formats: e.g. shapefiles (.shp) for vector data, grid files (.tiff) for raster data, netCDF for multi-dimensional raster data...

- Inputting spatial data into GIS
  - can be relatively simple or very involved depending on the original source

Spatial data

Where do spatial data come from?

- **Ground survey**
  - Example: DHS survey clusters, censuses, directories...
- Remote sensing sources
  - Aerial imagery: photographs, LIDAR data (3D)
  - Satellite imagery: e.g. DMSP-OLS (nighttime), LandSat, MODIS (daytime)...

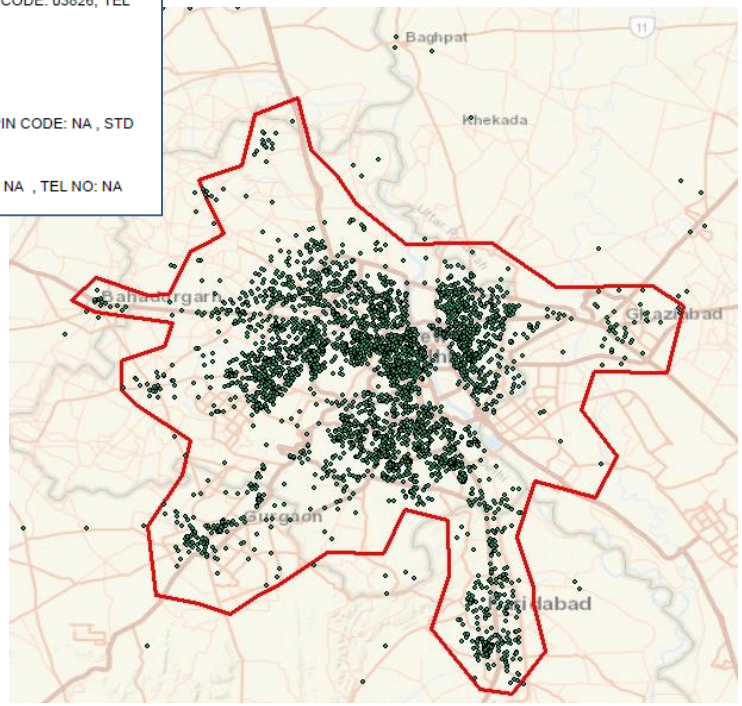
Spatial data

Example: from Indian firms' addresses to point data (Harari, 2016)

DIRECTORY ESTABLISHMENT

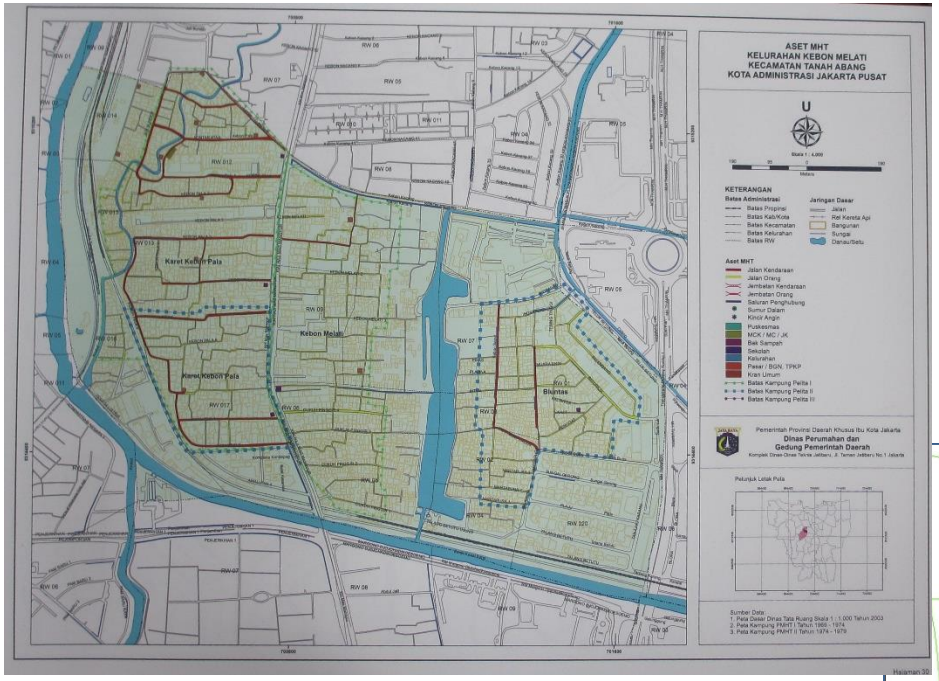
SECTOR :URBAN **STATE : TRIPURA** **DISTRICT : Dhalai**

Sl No	Name of Establishment	Address / Telephone / Fax / E-mail
(1)	(2)	(3)
NIC 2004 : 1513-Processing and preserving of fruit, vegetables and edible nuts		
1	DEYS SPICES AND FOOD PRODUCTS	KAMALPUR DHALAI TRIPURA DISTT. DHALAI TRIPURA PIN CODE: 799285, STD CODE: 03826, TEL NO: 262242, FAX NO: NA, E-MAIL : N.A.
NIC 2004 : 6511-Central banking_relates to the functions and working of the Reserve Bank of India		
2	THE ASSISTAN ENGINEER	PUBLIC HEALTH ENGINEER SUB - DIVISION, 6 KAMALPUR DHALAI TRIPURA PIN CODE: NA , STD CODE: NA , TEL NO: NA , FAX NO: NA, E-MAIL : N.A.
3	STATE BANK OF NDIA	KAMALPUR DHALAI, TRIPURA DISTT. TRIPURA PIN CODE: 799285, STD CODE: NA , TEL NO: NA , FAX NO: NA, E-MAIL : N.A.



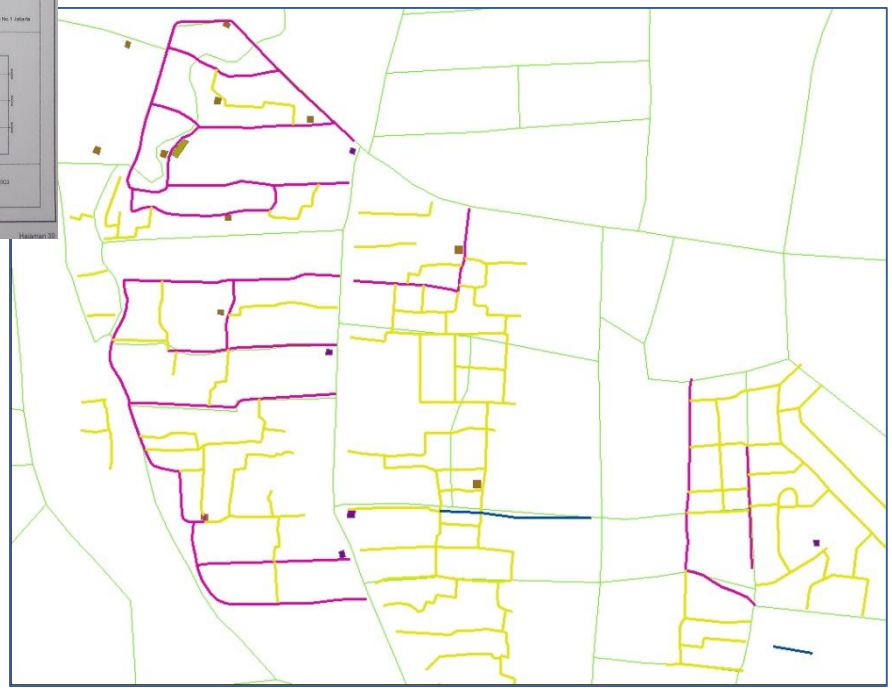
- Use Google maps to obtain coordinates from addresses
- Create point file from latitude and longitude

Spatial data



Example: digitizing analog maps

Slum upgrading in Jakarta, Harari and Wong (ongoing)



- Scan paper map
- Georeference image
- Manually trace roads and boundaries

Spatial data

Where do spatial data come from?

- Ground survey
  - Example: DHS survey clusters, censuses, directories...
- **Remote sensing sources**
  - Aerial imagery: photographs, LIDAR data (3D)
  - Satellite imagery: e.g. DMSP-OLS (nighttime), LandSat, MODIS (daytime)...

Spatial data

Advantages of remotely sensed data

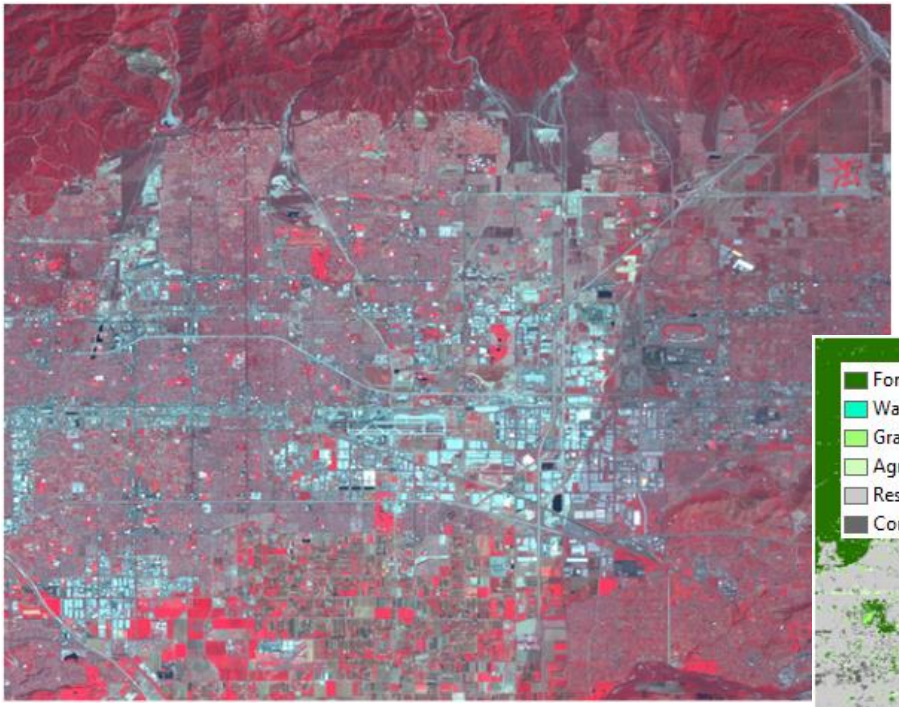
- Low-cost way of collecting panel data at a scale
- Uniform sampling and collection method across locations
- Can record hard-to-measure characteristics
- Free from
  - reporting biases / manipulation
    - E.g. pollution, deforestation, illegal crops...
  - arbitrariness of administrative units
- All of the above are especially important in a developing country setting!

Spatial data

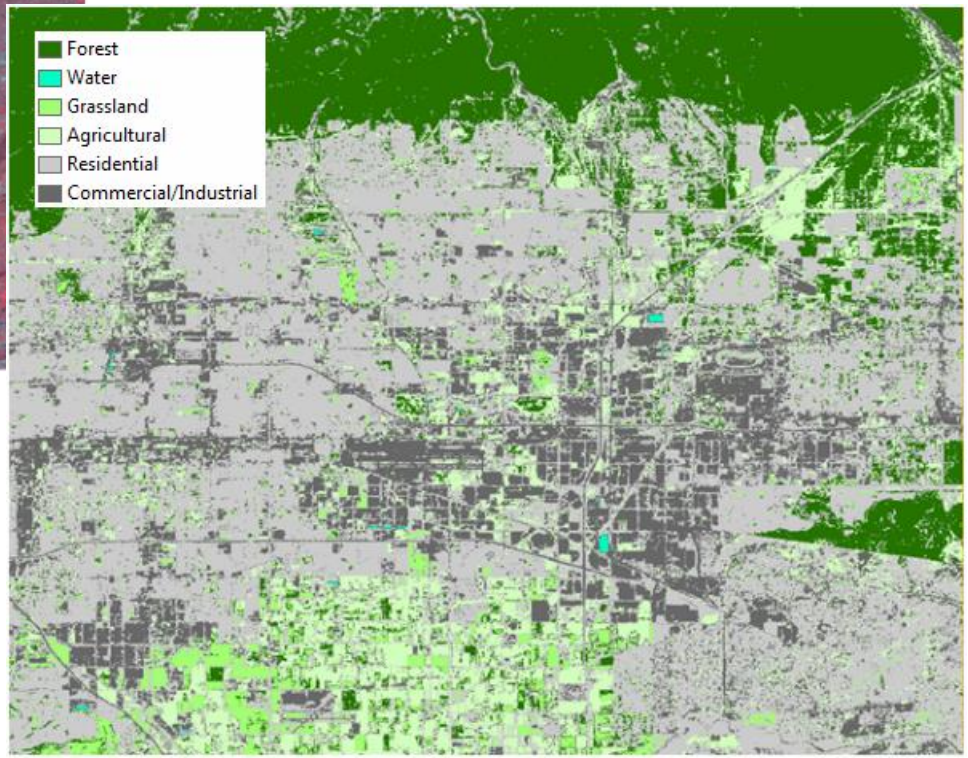
Issues with remotely sensed data

- Extensive processing may be needed to make data comparable and interpretable
- Researchers often need to make tradeoffs between coverage, accuracy, resolution...
- Remote sensing scientists and social scientists care about different things
  - E.g.: measurement error - does it matter?
  - Economists usually
    - want to avoid *non-classical* measurement error
    - can live with low resolution
    - can deal with a lot of sources of error through fixed effects

Spatial data



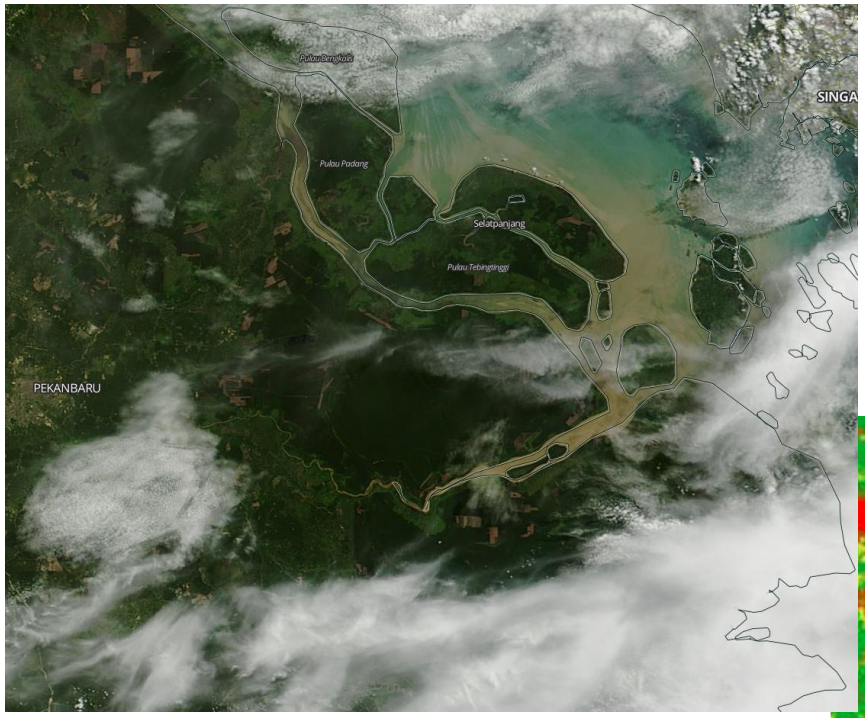
Example: classifying satellite imagery



Landsat imagery classified by land use

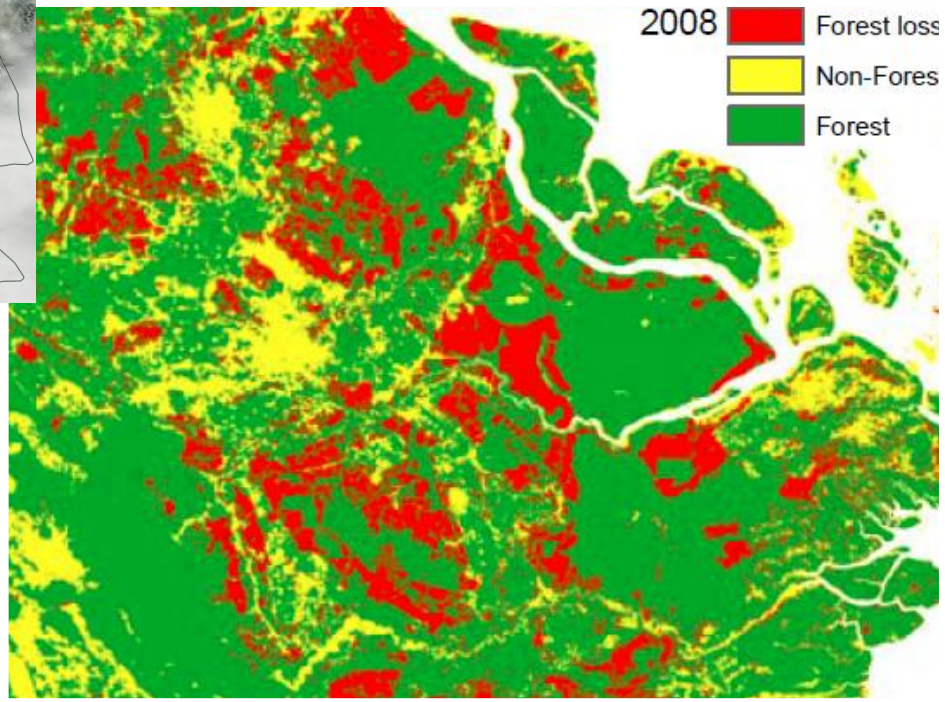
(Source: ESRI tutorial)

Spatial data



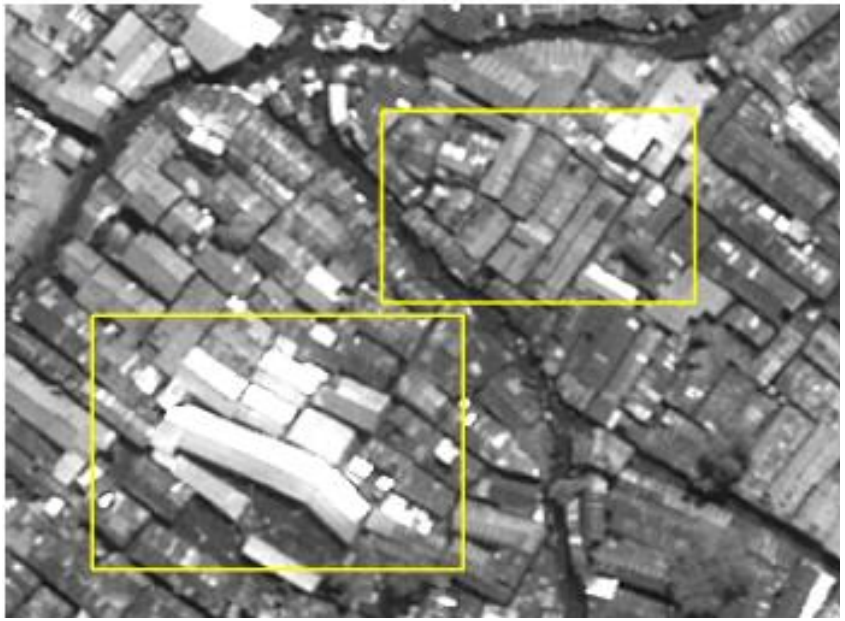
Example: classifying satellite imagery

Burgess et al. (2012)



From the raw MODIS satellite imagery (radiation patterns) to a forest cover change raster dataset

Spatial data



Example: object recognition

Identify newly renovated roofs in Nairobi slums based on reflection

Marx, Stoker and Suri (2015)

From high resolution daytime imagery to individual roofs

Spatial data

Issues with remotely sensed data

- Extensive processing may be needed to make data comparable and interpretable
- Researchers often need to make tradeoffs between coverage, accuracy, resolution...
- Remote sensing scientists and social scientists care about different things
  - E.g.: measurement error - does it matter?
  - Economists usually
    - want to avoid *non-classical* measurement error
    - can live with low resolution
    - can deal with a lot of sources of error through fixed effects

Roadmap

Outline

- Intro: spatial data
- What to do with spatial data
  - **Spatial correlation / dependence**
    - **Econometric implications**
    - Example from Harari and La Ferrara (2015)
  - Spatial data & identification
  - Spatial data as outcomes

Spatial correlation

Why should we care?

- Standard OLS assumptions require that units i are independent from one another

$$y_i = \alpha + \beta X_i + \varepsilon_i$$

- But this assumption may fail due to various forms of correlation:
  - Temporal correlation
  - Spatial correlation
- Depending on the nature of the cross-sectional dependence, OLS could be
  - Inefficient, with wrong standard errors
  - Biased and inconsistent

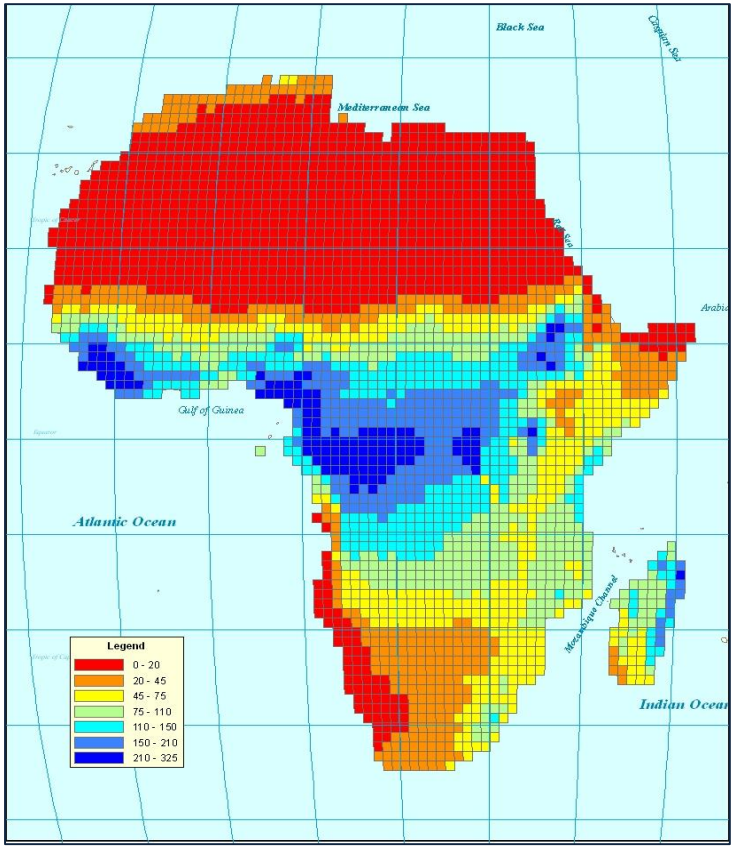
Spatial correlation

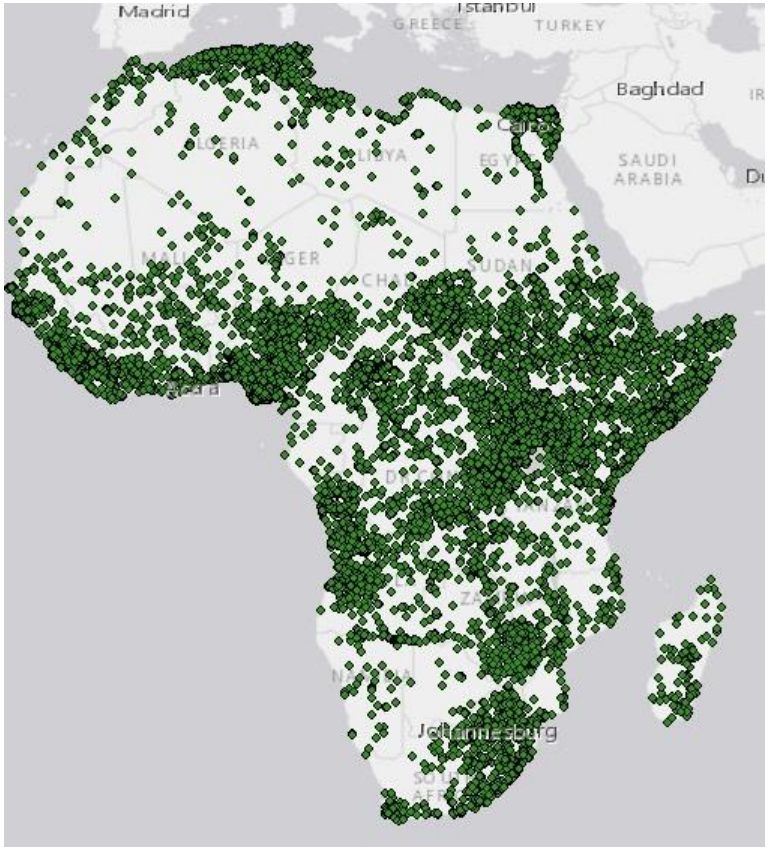
Spatial Effects

- Tobler's 'First Law of Geography': "Everything is related to everything else, but near things are more related than distant things." (Tobler, 1979)

Spatial correlation

Is there spatial correlation in the data?





Precipitation and conflict episodes 1997-2012 (Harari and La Ferrara, 2015)

Spatial correlation

Is there spatial correlation in the data?

- A number of diagnostic statistics to detect spatial correlation
- Moran's I: distance-weighted correlation coefficient used to detect departures from spatial randomness

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \cdot \frac{\sum_i \sum_j w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_i (y_i - \bar{y})^2}$$

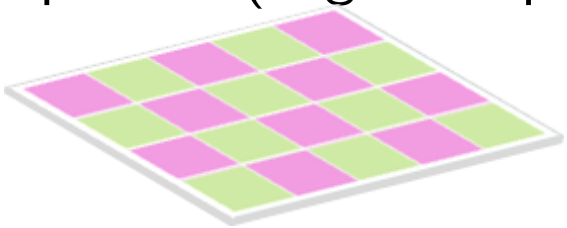
  - w_{ij} = distance between observations i and j

Spatial correlation

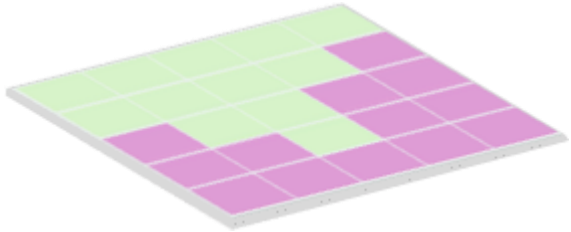
Is there spatial correlation in the data?

- A number of diagnostic statistics to detect spatial correlation
- Moran's I:

- -1: perfect dispersion (negative spatial autocorrelation)



- +1: perfect clustering (positive spatial autocorrelation)

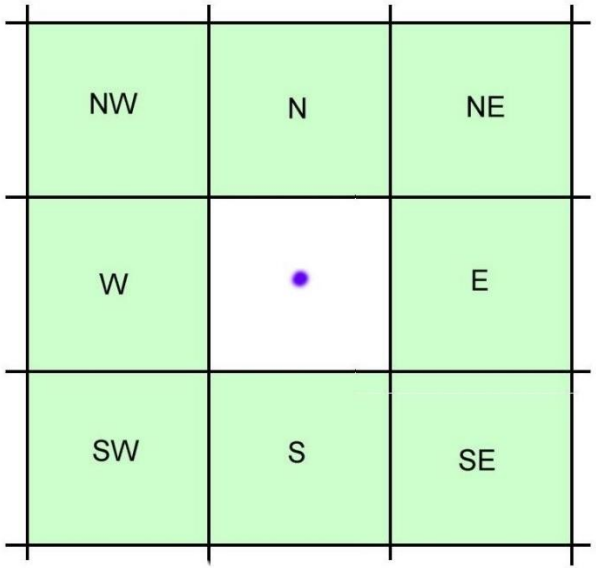


- 0: random spatial pattern

How to model spatial correlation?

A parallel between spatial and temporal correlation:

- Time: one-directional between two observations
  - This year is correlated with last year
- Space: two-directional (lat-lon) among several observations
  - Unit 1 is correlated with neighbors 2,3,4,5,6,7...



Spatial econometrics

How to model spatial correlation?

- Relative spatial positions between observations i and j are represented by a symmetric **spatial weights matrix W**
- Example: **binary contiguity** matrix

  Defines which pairs of observations i and j are neighbors ($w_{i,j} = 1$) and which are not ($w_{i,j} = 0$)

$$W = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

- Can define units as "neighbors" if
  - they share a common border
  - they are situated within a given distance band

How to model spatial correlation?

- Example: **inverse distance** matrix

  Observations i and j are more or less related depending on the inverse of the distance between them

$$\mathrm{W}=\begin{bmatrix} 0 & \frac{1}{(d_{1,2})^2} & \frac{1}{(d_{1,3})^2} \\ \frac{1}{(d_{2,1})^2} & 0 & \frac{1}{(d_{2,3})^2} \\ \frac{1}{(d_{3,1})^2} & \frac{1}{(d_{3,2})^2} & 0 \end{bmatrix}$$

- Often row standardized for ease of interpretation

$$w'_{ij}=\frac{w_{ij}}{\sum_j w_{ij}}$$

- Distance d can be Euclidean or not – even non-geographical

Spatial econometrics

How to model spatial correlation?

- Structure of spatial dependence is assumed, not estimated
- Some degree of arbitrariness in the specification of the weighting matrix
- Assumptions embedded in W might be considered strong... but it's even stronger to assume that all observations are spatially independent (w_{ij}=0)

Spatial econometrics

The spatial lag operator W

- Just as in time series you have lags in time, in spatial econometrics you have lags in space
- $W \cdot y$ = spatial lag of variable y
- Interpretation (in case of a row-standardized W): average value of variable y
  - in the neighborhood (contiguity weights)
  - in the whole sample with the weight decreasing with increasing distance (inverse distance weights)

Spatial econometrics

Spatial effects	Solutions
▶ Spatial correlation	
• in the error term ε	• *Spatial error model*, Conley standard errors
• in the covariates x	• *Spatial Durbin model* (include spatial lags of x)
▶ Spatial auto-correlation in the outcome variable y	• *Spatial autoregressive model* (include spatial lags of y), 2SLS

Note: for now, consider cross-sectional data only

Spatial econometrics

Spatial correlation in the error term

- OLS still unbiased, but standard errors wrong
- In most applied economics, this is dealt with by using **Conley (1999) standard errors**:
  - Allow for spatial dependence of an unknown form (non parametric)
  - Covariance matrix = weighted average of spatial autocovariances
    - Weights = product of Bartlett kernels in two dimensions (North/South and East/West)
    - Start at 1, decline linearly to 0 when a pre-specified cutoff distance is reached

Spatial econometrics

Spatial correlation in the error term

- **Spatial error model** (Anselin, 1988): parametrize spatial dependence in error term with matrix W

$$y = x\beta + \varepsilon$$

$$\varepsilon = \lambda W \varepsilon + \zeta$$

- Estimated by MLE (more efficient than OLS)
  - Likelihood function explicitly depends on W

Spatial econometrics

Spatial correlation in the covariates

- Just as in time series, we can add lags of the covariates
- Sometimes knows as **spatial Durbin model** (Anselin, 1988)

$$y = x\beta + \lambda W x + \varepsilon$$

- No particular econometric problem other than multicollinearity: OLS is unbiased and consistent

Spatial econometrics

Spatial dependence in the dependent variable

- Spatial autoregressive model, sometimes knows as spatial lag model (Anselin, 1988):

$$y = \rho W y + x\beta + \varepsilon$$

- Autoregressive term in y is simultaneous (reflection problem)
- OLS is biased and inconsistent
  - OLS will also be biased (due to omitted variables) if the term ρWy is ignored and dropped from the regression

Spatial econometrics

Spatial dependence in the dependent variable

- Spatial autoregressive model, sometimes knows as spatial lag model (Anselin, 1988):

$$y = \rho W y + x\beta + \varepsilon$$

- Autoregressive term in y is simultaneous (reflection problem)
- **2SLS**: must find suitable instruments for Wy
- **MLE**:
  - multiply right and left hand side by $(\rho W)^{-1}$, then write down likelihood
  - intuitively, like 2SLS using Wx as instruments for Wy

Spatial econometrics

New developments in spatial econometrics

- Spatial panel models
- Spatial Probit, Logit, Tobit
- Estimation: GMM

Roadmap

Outline

- Intro: spatial data
- What to do with spatial data
  - **Spatial correlation / dependence**
    - Econometric implications
    - **Example from Harari and La Ferrara (2015) "Conflict, Climate and Cells a Disaggregated Analysis"**
  - Spatial data & identification
  - Spatial data as outcomes

Harari and La Ferrara (2015)

- Empirical literature has stressed the link between weather shocks and civil conflict, esp. in Africa
  - Cross-country evidence
  - Variation in annual average precipitation

- **This paper**: a step further in understanding this relationship by taking the analysis to a different scale
  - Geographic disaggregation: units of observation = 110 x 110 km subnational "cells"
  - Temporal disaggregation: within-year climate variation
  - Rich, detailed georeferenced dataset, 36 African countries 1997-2011

Theory: relationship between weather shocks & civil conflict

- Effects mediated by agriculture:

  Weather shock = negative income shock

  - Opportunity cost channel: ↓opportunity cost of fighting => ↑conflict
  - Greed channel: smaller "pie" to be appropriated => ↓ conflict
  - State capacity channel: lower tax revenue => ↑conflict

- Direct effects: logistics, transit, psychology...

This paper: shed light on the channels by isolating variation in weather that is relevant for local agriculture

Methodology: disaggregated "grid" approach

General advantages of gridded (raster) data:

- Can be merged together, aggregated, interacted with vector data seamlessly
- Uniform sampling, as opposed to arbitrary (and potentially endogenous) boundaries
- Mitigate risk of "ecological fallacy" = drawing individual level inferences based on area level analyses
- Mitigates measurement error from measuring spatially continuous phenomena at the level of arbitrary administrative regions

Methodology: disaggregated "grid" approach

What does this buy us?

- Avoids endogeneity of administrative boundaries to conflict
- Rich set of geographic covariates whose effect on conflict might have been observed at the wrong scale (ruggedness, forest cover...)
- Allows to evaluate effects of covariates such as borders, distances from capital city...

Methodology: spatial econometrics techniques

- Explicit econometric modelling of spatial and temporal dependence: *dynamic spatial lag model*
  - Spatial and temporal dependence in covariates (e.g. weather)
  - Spatial and temporal dependence in outcomes (conflict)
  - Disentangle direct conflict spillovers in time and space from correlation in the covariates
- What does this buy us?
  - Investigate propagation of conflict over time and across space

Methodology: focus on climate within the year

We isolate the component of climate shocks that is relevant for local agriculture by constructing crop/cell-specific weather shock measures

- Identify cell specific main crop and growing season
- Exploit extra source of variation across cells: even if spatially clustered weather, possibly different shocks if cultivate crops w/ different growing seasons
- Climate indicator: **SPEI** Standardized Precipitation Evapotranspiration Index
  - precipitation + potential evaporation + temperature

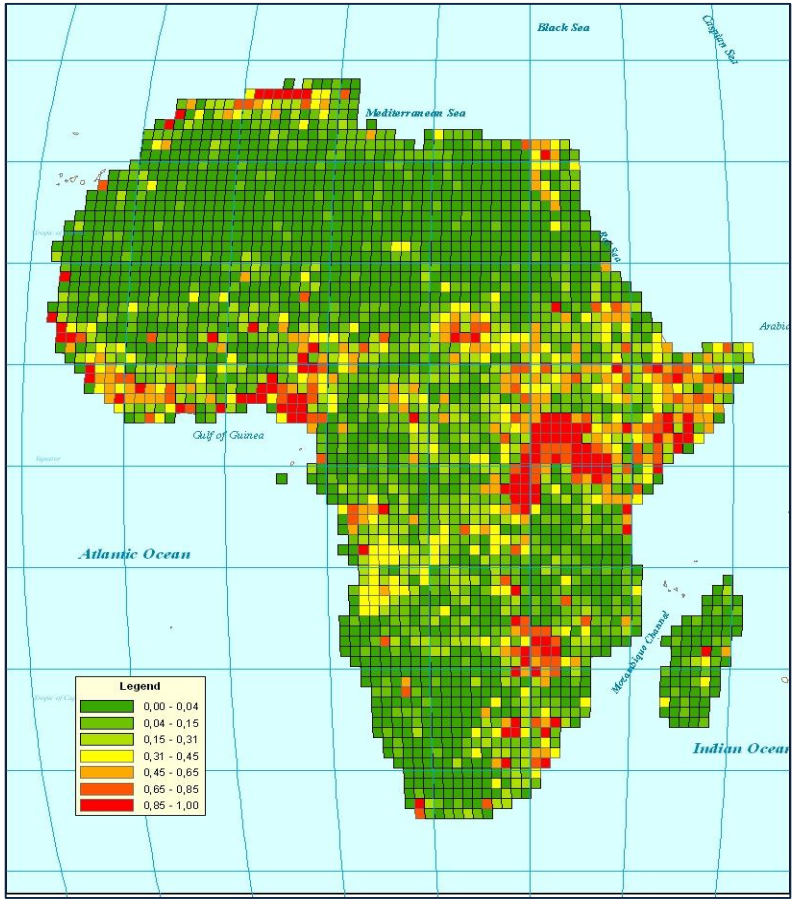
Data on conflict

- PRIO/Uppsala Armed Conflict Location and Event (ACLED)

• 46 African countries, 1997-2011
• codes exact locations and dates of civil conflict episodes

- Benchmark indicator:

• ANY EVENT it = whether cell i experienced a conflict event in year t



Fraction of sample years with at least one conflict event, 1997-2011

Data on climate – a premise

Many different datasets with different resolution, temporal coverage, spatial coverage.

Two formats:

- **Station data:**
  - Point data – each point is a weather station
  - Reports measurements of precipitation, temperature etc. recorded in the weather station
- **Gridded data:**
  - Raster datasets with various resolutions
  - Report a precipitation/temperature/... figure for each pixel

Data on climate – a premise

Gridded format is convenient and flexible, but coverage may be illusive due to interpolation (see e.g. Dell et al, 2014):

- Many gridded climate datasets are derived from interpolated station data
  - E.g.: CRU-TS
- Climatological stations can be extremely sparse in low income regions...
- ...and their ability to record data is endogenous to conflict itself (Fetzer, 2014)
- Problematic for disaggregated studies: when stations are sparse, most of the local variation is artificial

Data on climate – a premise

The alternative to (interpolated) station data : **reanalysis data**

- Combine ground-level measures (e.g. rainfall from weather stations, vessels, etc.)..
- ...with remotely sensed recordings (e.g. cloud cover from satellites, aircrafts etc.)
- Essentially, "intelligent interpolation"
- Available at a variety of resolutions – now the best practice for disaggregated studies
- Examples: ERA (European Centre for Medium-Range Weather Forecasts) dataset

Data on climate

Our benchmark indicator:

Standardized Precipitation-Evapotranspiration Index (SPEI)

- Combines information on: precipitation + temperature + Potential EvapoTranspiration (a function of sunshine exposure, latitude, wind patterns...)
- Extent to which the soil retains moisture received through precipitation
- Expressed in units of standard deviation from cell historical average (1901-2006)

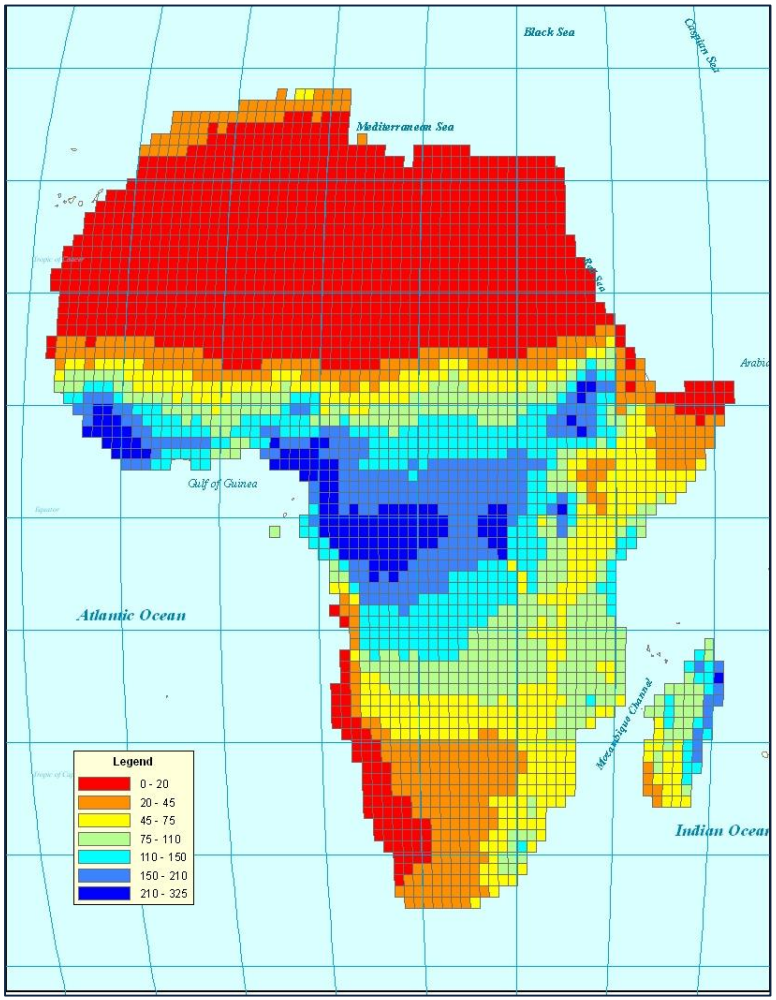
Data on climate

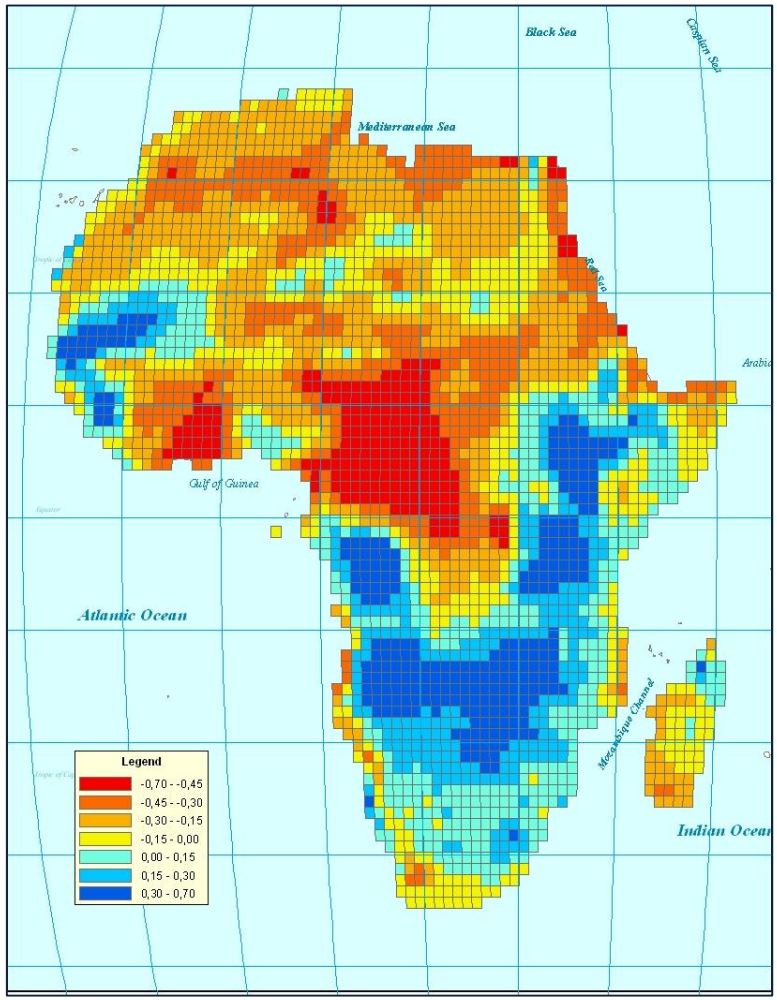
Standardized Precipitation-Evapotranspiration Index (SPEI)

- Original index: monthly data, 0.5x0.5 resolution, 1901-2006 , from Vicente-Serrano et al (2010)
  - Inputs: interpolated station data from CRU
- We re-calculate SPEI using reanalysis input data from ERA-Interim:
  - Observations are fed into ECMWF atmospheric circulation model to produce grid-specific forecasts at 6-hour frequency

Harari and La Ferrara (2015)
Data sources

Precipitation (left) vs. SPEI (right), average 1997 - 2011

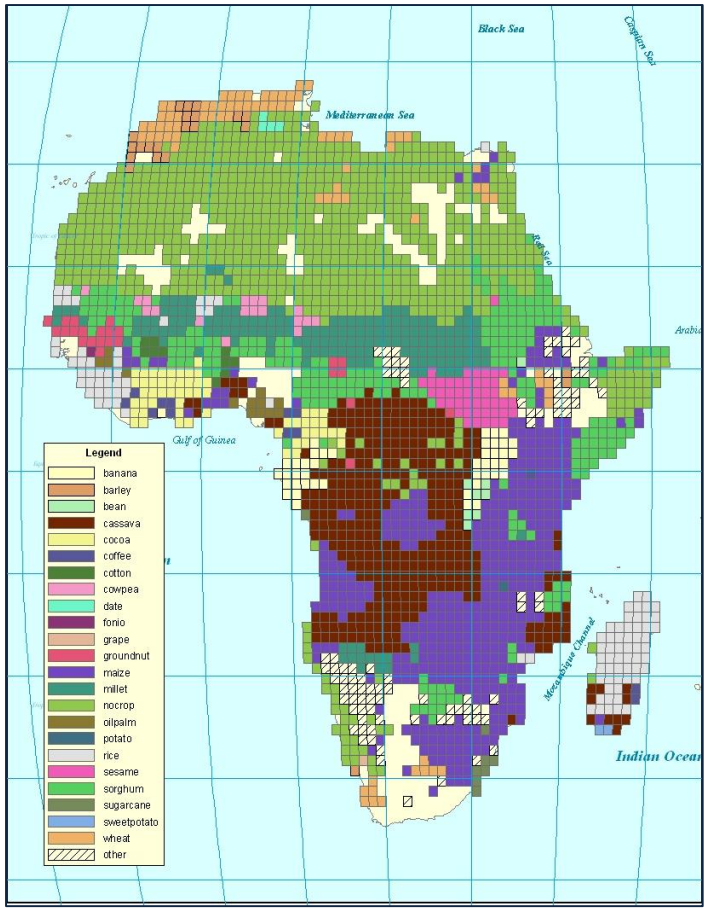




Data on crop cover

M3-Crops Data by Monfreda et al. (2008)

- raster dataset 5''x5'' resolution;
- reports harvested area as proportion of the grid cell area for 137 crops, as in year 2000;
- used to identify main crop for each cell

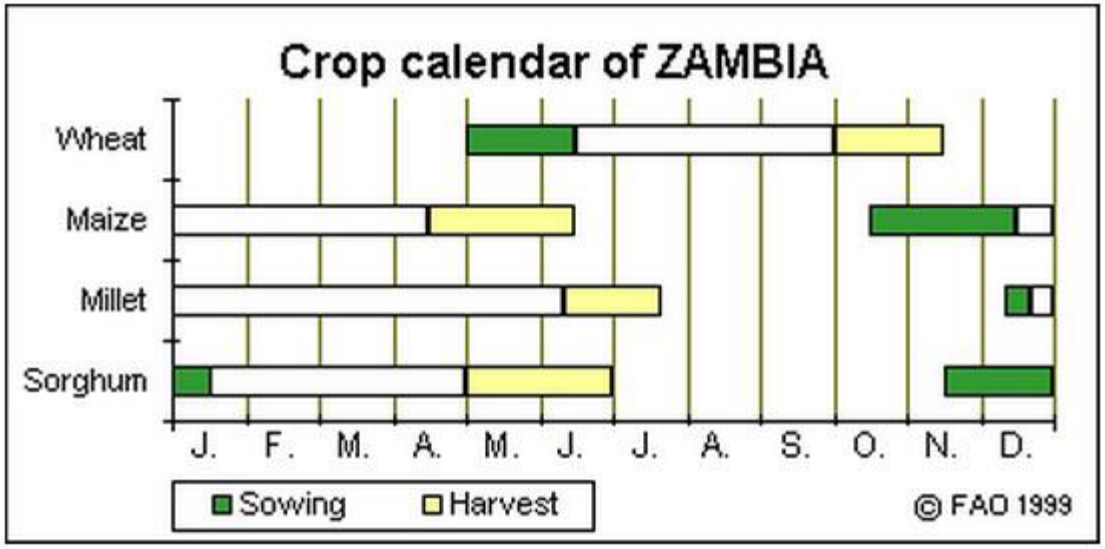


Main crop by cell, year 2000

Data on crop calendars

- Global Monthly Irrigated and Rainfed Crop Areas around the year 2000 (MIRCA 2000), Goethe Universität Frankfurt am Main
  - monthly growing seasons of 26 irrigated and rainfed crops at different latitudes and longitudes; resolution 5"x5"
- FAO Food security and Early warning Network for Information eXchange Workstation (FENIX) - Crop Calendar tool
- FAO Seeds and Plant Genetic Resources Crop Calendars
  - monthly planting and harvesting seasons for different countries

Harari and La Ferrara (2015)
Data sources



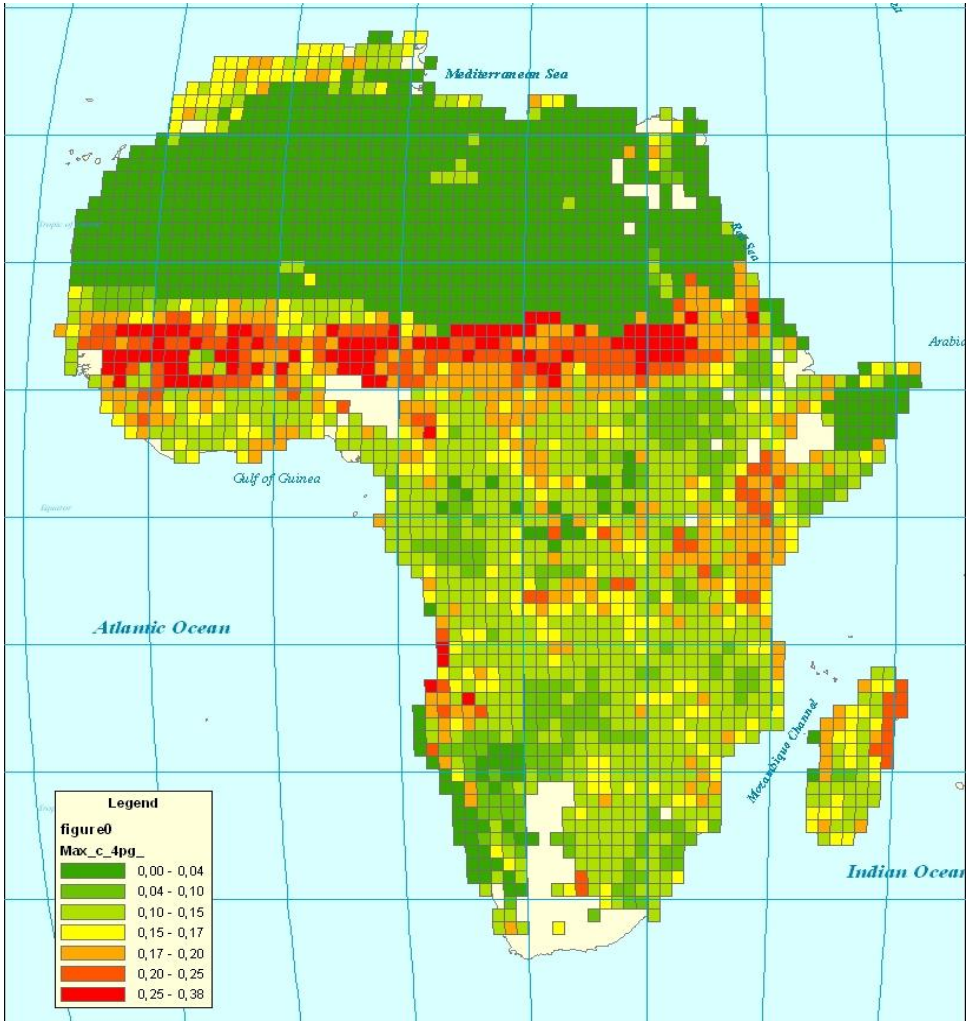
Our benchmark measure:

SPEI Shock Growing Season = fraction of the main crop's growing season in which SPEI was abnormally low

- consider the growing season of the main crop in the cell;
- take the number of consecutive growing season months in which SPEI was 1std dev below mean;
- express this measure as a fraction of the number of growing season months

Takes values from 0 (= "good year") to 1 (="bad year")

SPEI Shock Growing Season, year 2000



Other controls

- Ethno-Linguistic Fractionalization
  - Cell-level ELF computed based on relative land shares of different ethnic groups
  - Source for ethnic groups: GREG dataset
- Mineral deposit locations (USGS, Petrodata, Gemdata, Diadata)
- Presence of roads (Digital Chart of the World), rivers, elevation, ruggedness, distance from capital city (G-Econ dataset)

Spatial dependence in conflict/climate data

- Both conflict and weather appear strongly clustered, both in space and in time
- Part of the correlation in conflict is due to the correlation in the underlying determinants (e.g. weather)
- Part of it is due to direct spillovers in space / persistence in time
- Disentangling them is a version of the reflection problem
- Not just spatial correlation in the errors – most likely, spatial dependence in the conflict process itself

We address this by **modelling the spatial/temporal process explicitly** and fitting spatial econometric models

Model I – OLS

- Accounts for spatial correlation in the error term only

$$Conflict_{i,t} = \sum_{k=0}^{2} \gamma_k \, Climate_{i,t-k} + \rho X_{it} + \varepsilon_{c,t}$$

- $Climate_{i,t} = (Climate\_AllYear\ , Climate\_GrowingSeason)_{i,t}$
- $X_{it} = (Controls_i\ , yearFE, countryFE)$
- Standard errors corrected for spatial and serial correlation
  - Spatio-temporal version of Conley (1999): Hsiang (2010)

Model II – OLS

- Spatial Durbin model: accounts for spatial and temporal correlation in the covariates and in the error term

$$Conflict_{i,t} = \sum_{k=0}^{2} \gamma_k \, Climate_{i,t-k} + \rho X_{it} + \\ + \sum_{k=0}^{2} \delta_k \, WClimate_{t-k} + \theta W X_{it} + \varepsilon_{c,t}$$

- Benchmark weighting matrix W: binary contiguity matrix, distance 180 km, non standardized
  - "Neighbors" = 8 adjacent cells

Model III – MLE

- Dynamic autoregressive spatial Durbin model: accounts for spatial and temporal correlation in the covariates and in the error term

$$Conflict_{i,t} = \alpha Conflict_{i,t-1} + \beta WConflict_t + $$

$$+ \sum_{k=0}^{2} \gamma_k \, Climate_{i,t-k} + \rho X_{it}$$

$$+ \sum_{k=0}^{2} \delta_k \, WClimate_{t-k} + \theta WX_{it} + \varepsilon_{c,t}$$

Model III – MLE

- Where does the identification come from?
- Intuitively, use excluded lags (in time and space) of the covariates as instruments for the autoregressive terms in Conflict
    - Functional form assumptions on W
- By definition these cannot be tested, but can do sensitivity analyses with different W's

Preliminary cross sectional analysis:

- Collapse our dataset (1 obs per cell) to highlight geographic correlates of conflict
- Dependent variable: average conflict incidence over time in the cell

A: Cross sectional sample	No. Obs.	Mean	Std Dev
Fraction of years with conflict	2669	0.171	0.252
Shared	2669	0.252	0.434
Border	2669	0.050	0.218
Area, in km^2	2669	10926.7	2577.3
Elevation, in m	2669	314.9	269.6
Rough	2669	0.093	0.102
Distance to river, in km	2669	628.0	476.1
Road	2669	0.241	0.428
Minerals	2669	0.210	0.408
ELF	2669	0.203	0.240

Harari and La Ferrara (2015)
Results

Table 2: Conflict incidence, cross section

Dependent variable: fraction of years with at least one conflict event

	(1) Model I
Shared	0.0379***
	(0.0128)
Border	0.00212
	(0.0181)
Area(a)	0.0000
	(0.00250)
Elevation(a)	-0.0705*
	(0.0397)
Rough	0.377***
	(0.106)
Distance to river(b)	-0.00584**
	(0.00272)
Road	0.0932***
	(0.0164)
ELF	0.0514*
	(0.0265)
Minerals	0.0616***
	(0.0138)
Observations	2,669
R-squared	0.587

(a) Coefficient and std error multiplied by 10^3 (b) Coefficient and std error multiplied by 10^2
Regressions include country FE * p<0.01, ** p<0.05, * p<0.1
Standard errors in parenthesis corrected for spatial dependence, following Conley (1999).

Harari and La Ferrara (2015)
Results

Table 2: Conflict incidence, cross section

Dependent variable: fraction of years with at least one conflict event

	(2) Model II	
	X	W · X
Shared	0.0273**	-0.000858
	(0.0118)	(0.00330)
Border	-0.0173	0.00991
	(0.0160)	(0.00724)
Area[a]	0.00281	-0.000812
	(0.00493)	(0.00110)
Elevation[a]	-0.337***	0.0401**
	(0.128)	(0.0183)
Rough	0.279***	0.0145
	(0.0833)	(0.0205)
Distance to river[b]	-0.00173	-0.000618
	(0.00433)	(0.000706)
Road	0.106***	-0.00646
	(0.0152)	(0.00406)
ELF	0.0393*	0.00467
	(0.0225)	(0.00720)
Minerals	0.0486***	0.0111**
	(0.0117)	(0.00450)
Observations	2,669	
R-squared	0.630	

(a) Coefficient and std error multiplied by 10^3 (b) Coefficient and std error multiplied by 10^2
Standard errors in parenthesis corrected for spatial dependence, following Conley (1999).
* p<0.01, ** p<0.05, * p<0.1 W = binary contiguity matrix, cutoff 180 km. Regressions include country FE

Harari and La Ferrara (2015)
Results

Dependent variable: fraction of years with at least one conflict event		
	(2)	
	Model III	
W · Y	0.0237***	
	(0.00453)	
	X	W · X
Shared	0.0340***	-0.000460
	(0.00988)	(0.00334)
Border	-0.00704	0.0122**
	(0.0162)	(0.00547)
Area[a]	-0.000821	0.000133
	(0.00232)	(0.000525)
Elevation[a]	-0.0759***	0.00102
	(0.0241)	(0.00742)
Rough	0.430***	-0.0236
	(0.0652)	(0.0165)
Distance to river[b]	-0.00496**	-8.19e-05
	(0.00197)	(0.000539)
Road	0.0949***	0.00294
	(0.0127)	(0.00393)
ELF	0.0422**	0.00222
	(0.0201)	(0.00595)
Minerals	0.0560***	0.00564
	(0.0113)	(0.00383)
Observations	2,669	
R-squared	0.421	

(a) Coefficient and std error multiplied by 10^3 (b) Coefficient and std error multiplied by 10^2
Standard errors in parenthesis corrected for spatial dependence, following Conley (1999).
* p<0.01, ** p<0.05, * p<0.1 W = binary contiguity matrix, cutoff 180 km. Regressions include country FE

B: Panel sample	No. Obs.	Mean	Std Dev
ANY EVENT	37425	0.170	0.376
BATTLE	37425	0.098	0.297
CIVILIAN	37425	0.098	0.297
RIOT	37425	0.056	0.231
REBEL	37425	0.029	0.168
SPEI	37425	-0.089	0.575
SPEI Growing Season, Main Crop	37425	-0.022	0.362
SPEI Shock, Growing Season, Main Crop	37425	0.105	0.188
Rain	37425	65.108	69.129
Rain Growing Season, Main Crop	37425	51.582	63.859
Temperature, abs dev	37425	0.782	0.203
Temperature abs dev, Growing Season, Main Crop	37425	0.332	0.320

Table 3: Conflict incidence and climate, panel

Dependent variable (Y) = 1 if conflict event in year t (ANY EVENT)

	(1)
	Model I
SPEI	0.00771
	(0.00539)
SPEI, t-1	-0.0102*
	(0.00568)
SPEI, t-2	0.00322
	(0.00561)
SPEI Shock Growing Season	0.0469***
	(0.0179)
SPEI Shock Growing Season, t-1	0.0550***
	(0.0180)
SPEI Shock Growing Season, t-2	0.0594***
	(0.0180)
Observations	37,425
R-squared	0.327

Notes: Each observation is a cell/year. All regressions include controls listed in table 2, country and year fixed effects. W = binary contiguity matrix, cutoff 180 km. *** p<0.01, ** p<0.05, * p<0.1 Standard errors in parenthesis. Corrected for spatial and serial correlation.

Harari and La Ferrara (2015)
Results

Dependent variable (Y) = 1 if conflict event in year t (ANY EVENT)

	(2) Model II	
	X	W · X
SPEI	0.0263*	-0.00241
	(0.0136)	(0.00208)
SPEI, t-1	0.0124	-0.00341
	(0.0138)	(0.00213)
SPEI, t-2	0.00559	-0.000113
	(0.0140)	(0.00216)
SPEI Shock Growing Season	0.0272	0.00604
	(0.0195)	(0.00405)
SPEI Shock Growing Season, t-1	0.0499**	0.00218
	(0.0212)	(0.00422)
SPEI Shock Growing Season, t-2	0.0458**	0.00470
	(0.0213)	(0.00425)
Observations	37,425	
R-squared	0.353	

Notes: Each observation is a cell/year. All regressions include controls listed in table 2, country and year fixed effects. W = binary contiguity matrix, cutoff 180 km. Standard errors in parenthesis. Corrected for spatial and serial correlation.
*** p<0.01, ** p<0.05, * p<0.1

Harari and La Ferrara (2015)
Results

Dependent variable (Y) = 1 if conflict event in year t (ANY EVENT)

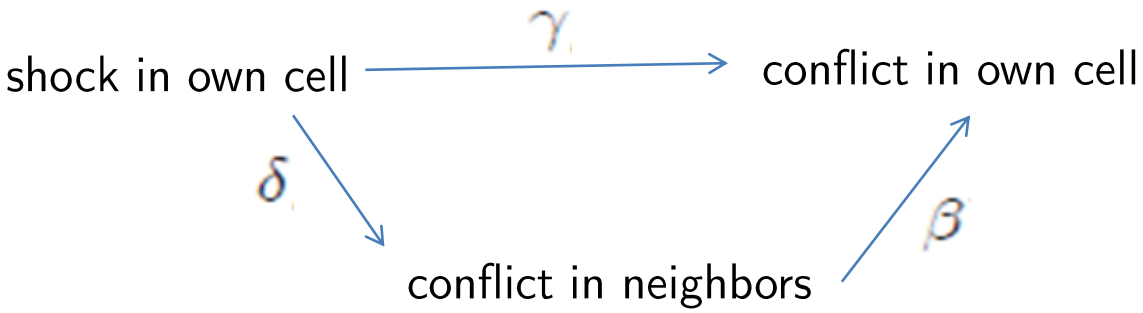
	(3)	
	Model III	
Y, t-1	0.332***	
	(0.00493)	
W · Y	0.0451***	
	(0.00103)	
	X	W · X
SPEI	0.0103	0.0000
	(0.0123)	(0.00182)
SPEI, t-1	-0.000535	-0.00101
	(0.0122)	(0.00184)
SPEI, t-2	0.00432	-0.000474
	(0.0117)	(0.00175)
SPEI Shock Growing Season	0.0103	0.00245
	(0.0190)	(0.00333)
SPEI Shock Growing Season, t-1	0.0401**	-0.00290
	(0.0196)	(0.00337)
SPEI Shock Growing Season, t-2	0.0407**	-0.00307
	(0.0194)	(0.00338)
Observations	37,425	
R-squared	0.347	

Notes: Each observation is a cell/year. All regressions include controls listed in table 2, country and year fixed effects. W = binary contiguity matrix, cutoff 180 km.
Standard errors in parenthesis. Corrected for clustering at the cell level.

Feedback mechanism

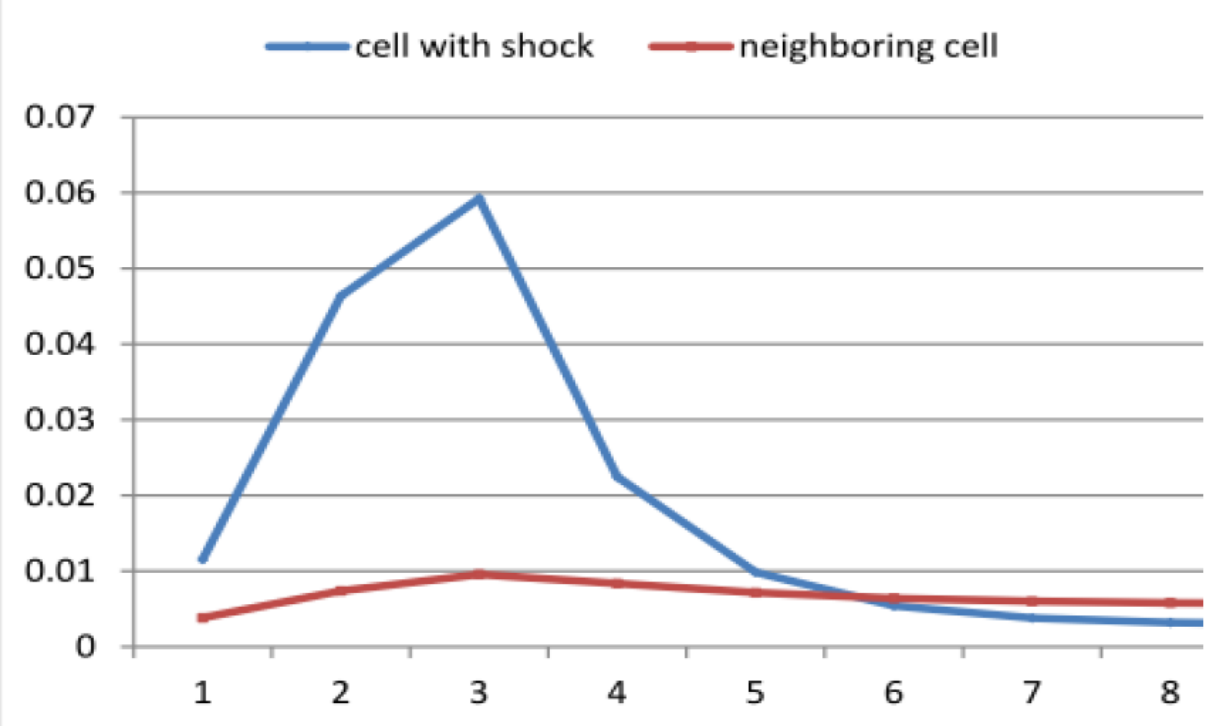
$$Conflict_{it} = \alpha Conflict_{i,t-1} + \beta WConflict_{t} +$$
$$+ \sum_{k=0}^{2} \gamma_k Shock_{i,t-k} + \sum_{k=0}^{2} \delta_k WShock_{t-k}$$
$$+ \rho X_i + \theta WX + \varepsilon_{it}$$

A one-time SPEI Growing Season Shock propagates in time and space feeding back into the process through autoregressive terms



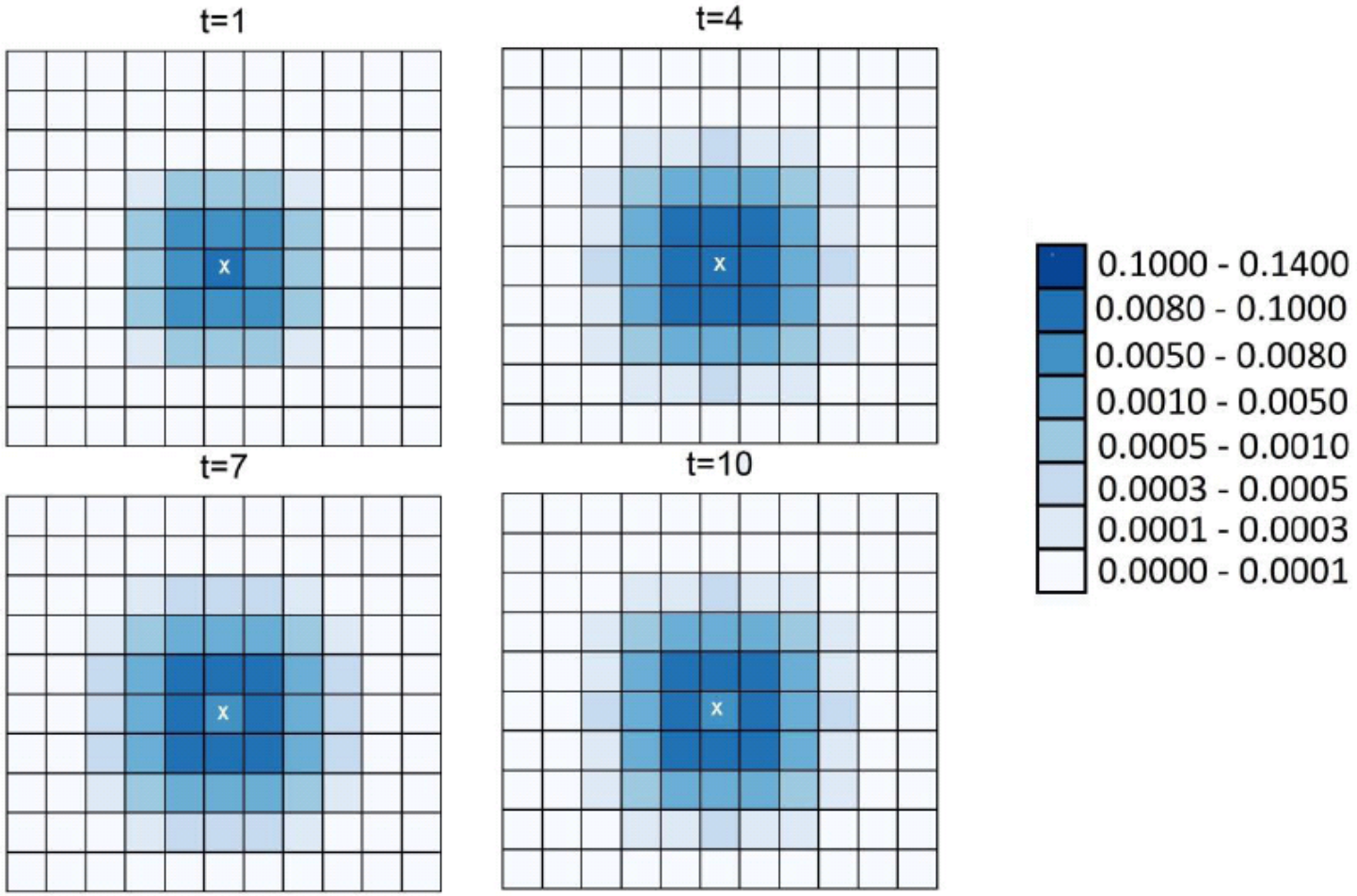
"Impulse response":

- set all covariates and prior conflict to 0;
- give to a hypothetical cell a one-time *SPEI Shock Growing Season =1*
- use Model III estimated coefficients to track the estimated marginal impact



Harari and La Ferrara (2015)
Results

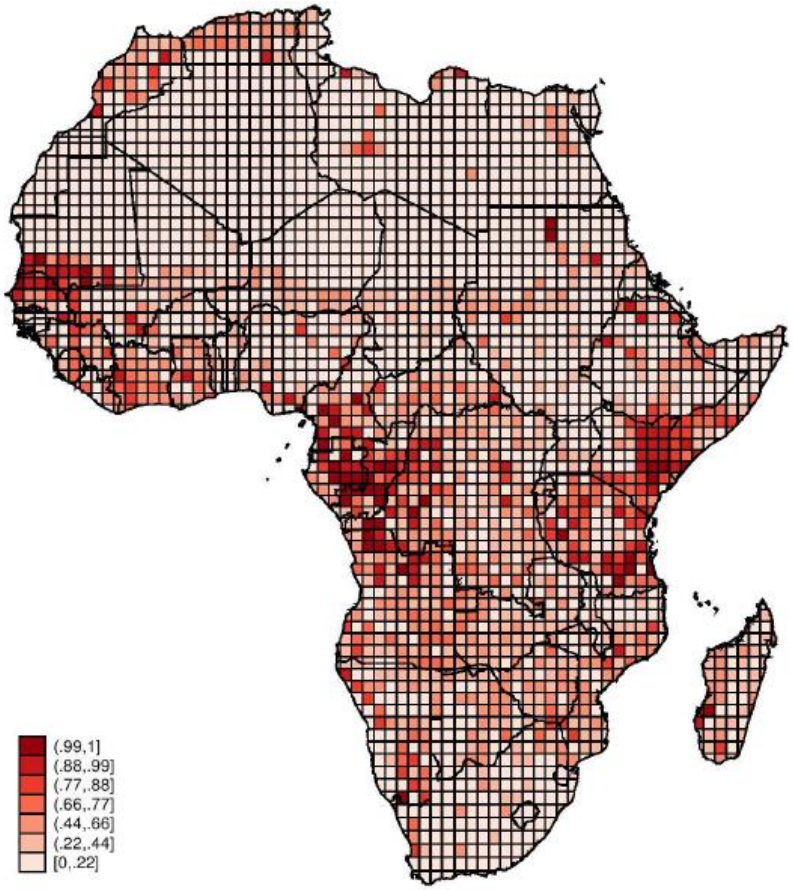
"Impulse response":



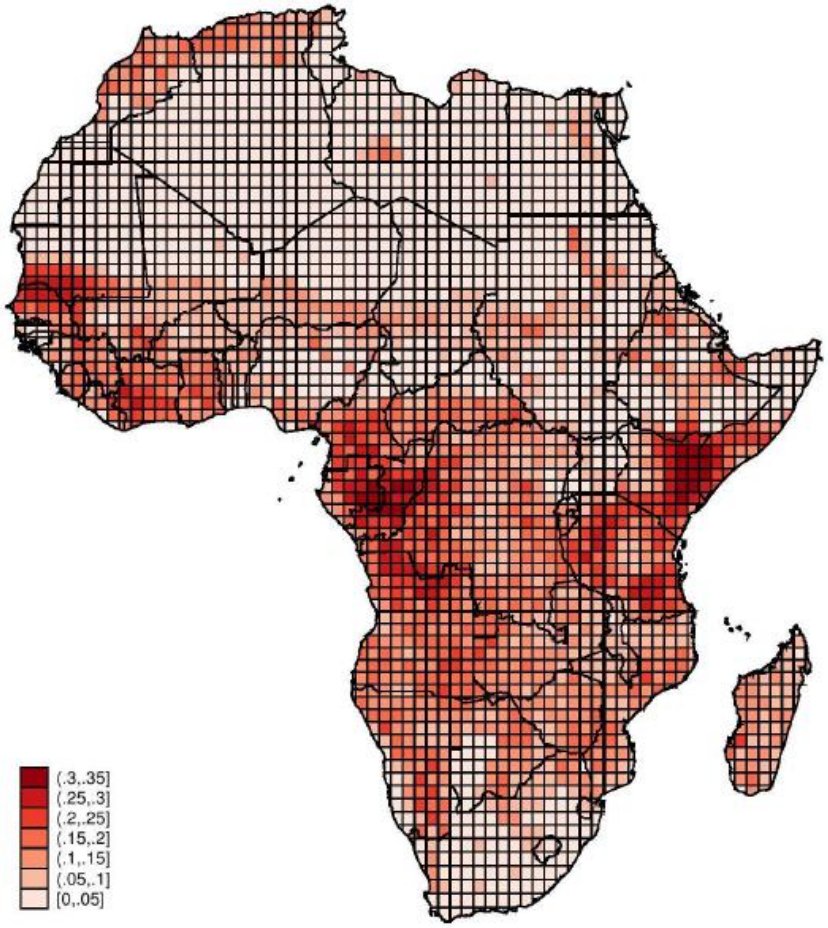
Instead of arbitrary shock=1, feed forecatsed shock over 2012-2030

- Cell-level climate projections from CORDEX Archive under CAN-ESM2 model, for a midrange-mitigation emissions scenario (RCP4.5)
- Avg. SPEI Shock Growing Season (which is 0.10 in 1997-2011) becomes 0.25 in 2012-2030

Projected SPEI shocks



Projected pc point increases in conflict

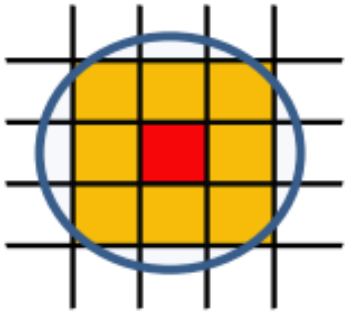


Robustness

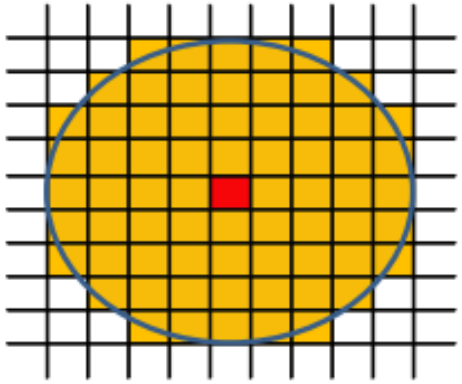
- Choice of weighting matrix
  - binary vs. inverse distance
  - distance cutoff

- Modifiable Areal Unit Problem =
imposition of artificial units of spatial reporting on continuous geographical phenomenon resulting in the generation of artificial spatial patterns
  - scale
  - zonation

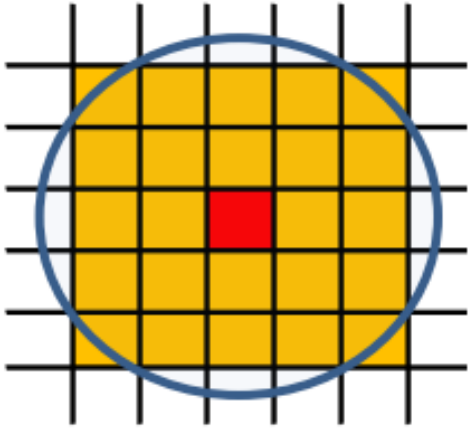
Different definitions of neighborhood



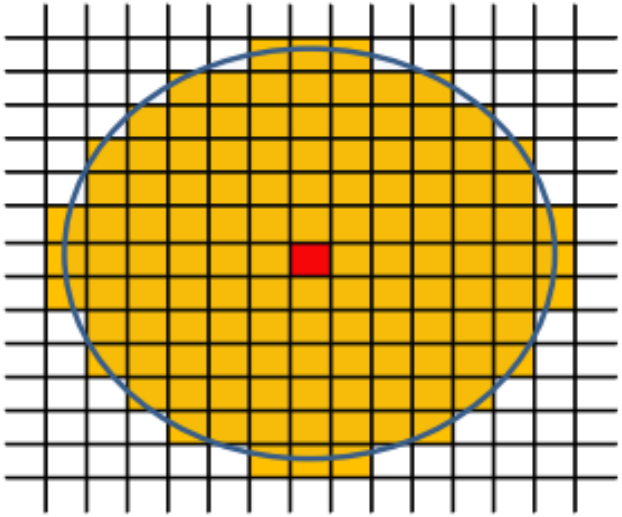
190 km



450 km



290 km



600 km

Harari and La Ferrara (2015)
Results

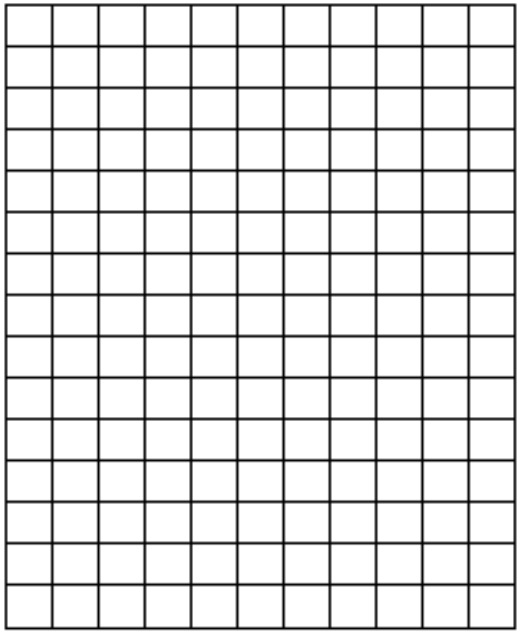
Dependent variable (Y) = 1 if conflict event in year t (ANY EVENT)

	(1)	(2)	(3)	(4)
	Binary contiguity matrix			
	180 km	290 km	450 km	600 km
Y_{t-1}	**0.352*** **	**0.351*** **	**0.361*** **	**0.371*** **
	(0.00498)	(0.00500)	(0.00502)	(0.00502)
$W \cdot Y$	**0.0360*** **	**0.0187*** **	**0.00809*** **	**0.00319*** **
	(0.00109)	(0.000620)	(0.000387)	(0.000314)
SPEI Shock Growing Season	0.0123	0.0167	0.0187	0.0206
	(0.0169)	(0.0155)	(0.0143)	(0.0135)
SPEI Shock Growing Season, t-1	0.0429**	0.0259	0.00359	-0.00716
	(0.0174)	(0.0160)	(0.0147)	(0.0140)
SPEI Shock Growing Season, t-2	0.0403**	0.0301*	0.0389***	0.0353**
	(0.0176)	(0.0162)	(0.0148)	(0.0141)
Observations	34,930	34,930	34,930	34,930
R-squared	0.388	0.391	0.391	0.390
Controls	X	X	X	X
Country x year fixed effects	X	X	X	X

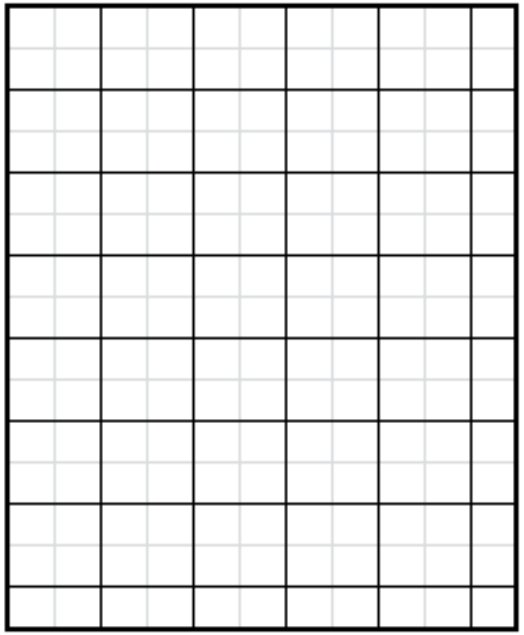
Notes: Each observation is a cell/year. Estimation by MLE. Standard errors in parenthesis, corrected for clustering at the cell level. *** p<0.01, ** p<0.05, * p<0.1

MAUP: scale

Which resolution?



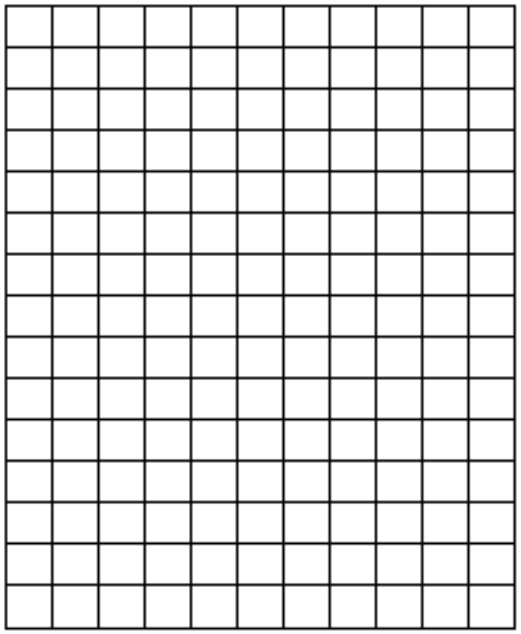
original grid



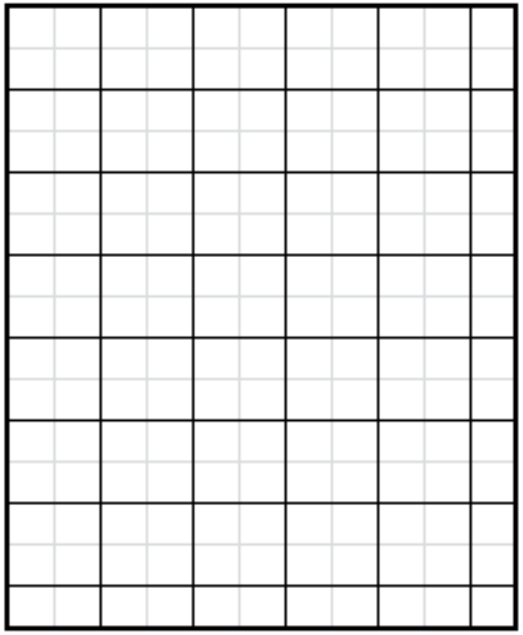
2x2 grid, example (1)

MAUP: zonation

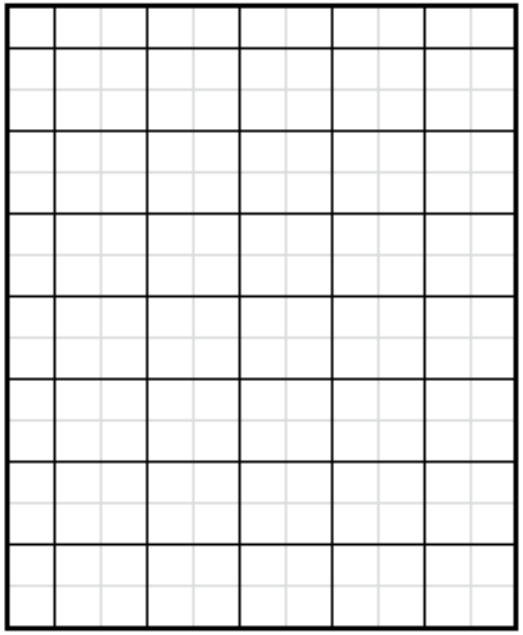
2x2 grid can be centered in 4 possible ways:



original grid



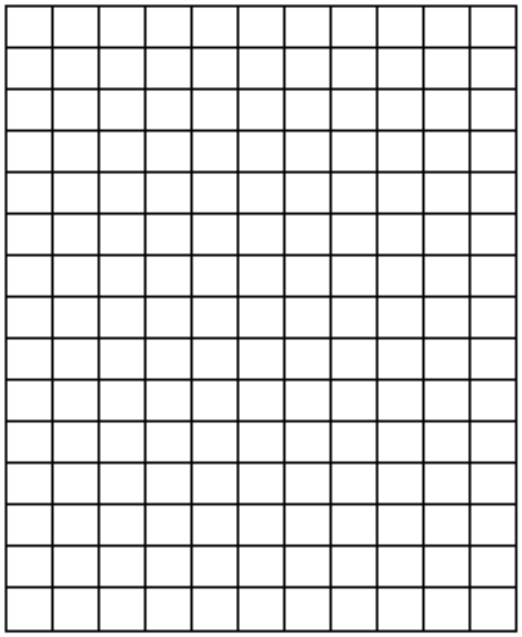
2x2 grid, example (1)



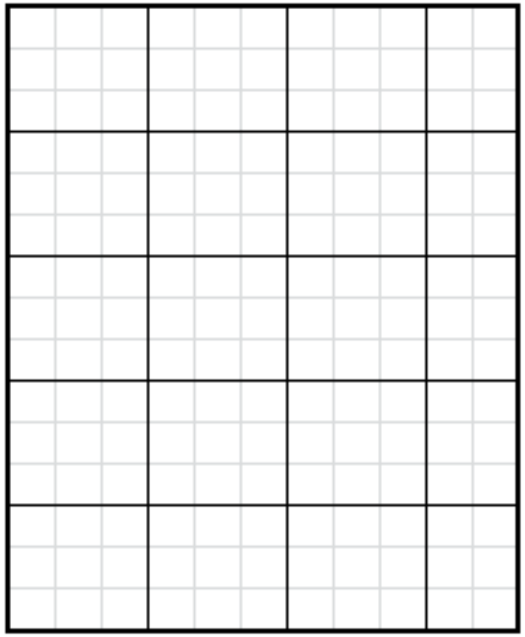
2x2 grid, example (2)

MAUP: zonation

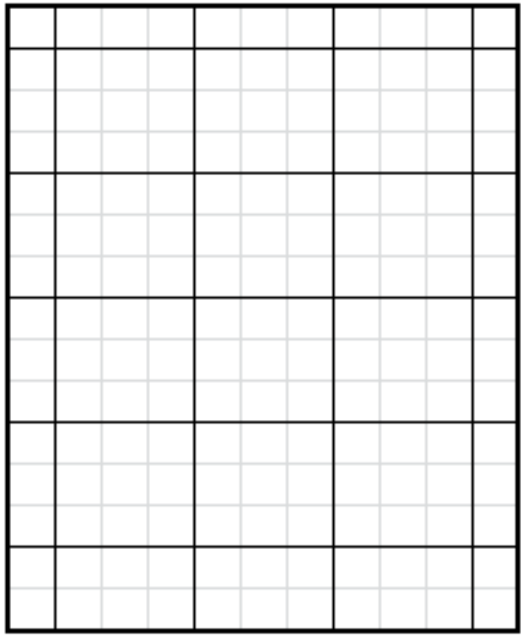
3x3 grid can be centered in 9 possible ways:



original grid



3x3 grid, example (1)



3x3 grid, example (2)

Harari and La Ferrara (2015)
Results

Panel A: 2x2 cells

Dependent variable (Y) = 1 if conflict event in year t (ANY EVENT).

	Model III - MLE			
	avg. coefficient	coefficient std. dev.	avg. std. error	no. panels 10% significant
Y_{t-1}	0.487	0.005	0.005	4
W · Y	0.020	0.001	0.001	4
SPEI Shock Growing Season	0.030	0.011	0.011	2
SPEI Shock Growing Season, t-1	0.004	0.011	0.011	0
SPEI Shock Growing Season, t-2	0.052	0.011	0.011	4
W · SPEI Shock Growing Season	0.002	0.003	0.003	0
W · SPEI Shock Growing Season, t-1	0.003	0.003	0.003	0
W · SPEI Shock Growing Season, t-2	0.002	0.003	0.003	0
Average nr of obs	8733			
Average R squared	0.672			

Harari and La Ferrara (2015)
Results

Panel B: 3x3 cells

Dependent variable (Y) = 1 if conflict event in year t (ANY EVENT).

	Model III - MLE			
	avg. coefficient	coefficient std. dev.	avg. std. error	no. panels 10% significant
Y, t-1	0.572	0.025	0.013	9
W · Y	0.005	0.004	0.004	4
SPEI Shock Growing Season	0.042	0.018	0.025	4
SPEI Shock Growing Season, t-1	-0.027	0.019	0.026	2
SPEI Shock Growing Season, t-2	0.062	0.023	0.026	7
W · SPEI Shock Growing Season	-0.004	0.006	0.008	0
W · SPEI Shock Growing Season, t-1	0.019	0.005	0.009	7
W · SPEI Shock Growing Season, t-2	-0.002	0.007	0.009	0
Average nr of obs	3881			
Average R squared	0.807			

Notes: Regressions include controls listed in table 2 and country-year fixed effects. Standard errors corrected for clustering at the cell level. Panel A: Results of the estimation of model III in 4 possible panels of 2x2 cells. W = binary contiguity matrix, cutoff 390 km. Panel B: Results of the estimation of model III in 9 possible panels of 3x3 cells. W = binary contiguity matrix, cutoff 490 km.

Other results from heterogeneous effects analysis

- Effects of SPEI shocks most pronounced for events such as violence against civilians and riots
- Some evidence that effect of shock mitigated in more democratic settings ("grievances" channel)
- Some evidence that effect of shock mitigated if higher tax revenue for state government ("state capacity" channel)
- No differential impact by road intensity (no support for the "logistics" channel)

Key takeaways:

geographic and temporal disaggregation highlights interesting patterns and allows to shed light on mechanisms:

- Weather affects conflict likelihood locally mostly through the channel of agriculture
- Sizeable spillovers over time and across space → caution w/ studies that do not incorporate spatial dynamics
- Our estimates allow to quantify predicted effects of global warming & identify high-risk areas
- Methodology applicable beyond conflict analysis, given rich geo-referenced datasets

Roadmap

Outline

- Intro: spatial data
- What to do with spatial data
  - Spatial correlation / dependence
  - **Spatial data & identification**
    - Spatial data / distances as instruments
    - Examples from Harari (2016)
  - Spatial data as outcomes

Spatial instruments

Space as an instrument

- How spatial data can help with the identification:
  "Suitability" types of instruments
  - Evaluating the impact of infrastructure
  - Crop suitability and long-run development

- How spatial patterns can help with the identification:
  Distances as instruments
  - Instrumenting routes
  - Spatial RD
  - Propagation patters as sources of variation

Spatial instruments

Suitability instruments

- Faber and Gaubert (2015): impact of tourism on local development in Mexico
  - Problem: endogenous location of tourist destinations
  - Compare satellite imagery with beach rankings to identify spectral characteristics associated with beach quality
  - Develop measure of beach quality for every segment of coastline, based on satellite imagery
  - Use natural beach quality to predict tourism prevalence
  - Exclusion restriction: must show that this "exogenous beach quality" did not affect location decisions of locals

Spatial instruments

Evaluating the impact of infrastructure

- Duflo and Pande (2007): impact of dams in Indian rural districts
  - Dams require a certain range of river gradient – neither too steep not too shallow
  - Time variation: interact district-level dam suitability with state-level trends in dam construction
  - Spatially heterogeneous effects: dam districts lose, downstream districts gain
- Dinkelman (2011): rural electrification in South Africa
- Lipscomb, Mobarak and Barham (2013): hydropower in Brazil

Spatial instruments

Crop suitability and long-run development

- FAO Global Agro-Ecological Zones (GAEZ)
  - Potential yields by crop based on climate and soil
  - 10 x 10 km raster
- Nunn and Qian (2011): long-run impact of nutritional improvements
  - Use potato suitability to predict potato adoption across regions of Europe

Spatial instruments

Crop suitability and long-run development

- Alesina, Giuliano and Nunn (2013): do modern-day gender roles stem from traditional agricultural practices?
  - Use suitability for crops requiring ploughing to predict plough adoption
- Mayshar et al (2015): early state formation on present outcomes
  - Role of cereals (storable, incentive to confiscate) vs. tubers (perishable, no incentive to confiscate)
  - Use suitability for cereals to predict early state formation

Spatial instruments

Distances as instruments

- Nunn (2008): long-term impacts of slave trade in African countries

- Problem: selection into slave trade
- IV strategy: instrument country-level slave exports with distance to the location of the demand for slaves (ports)

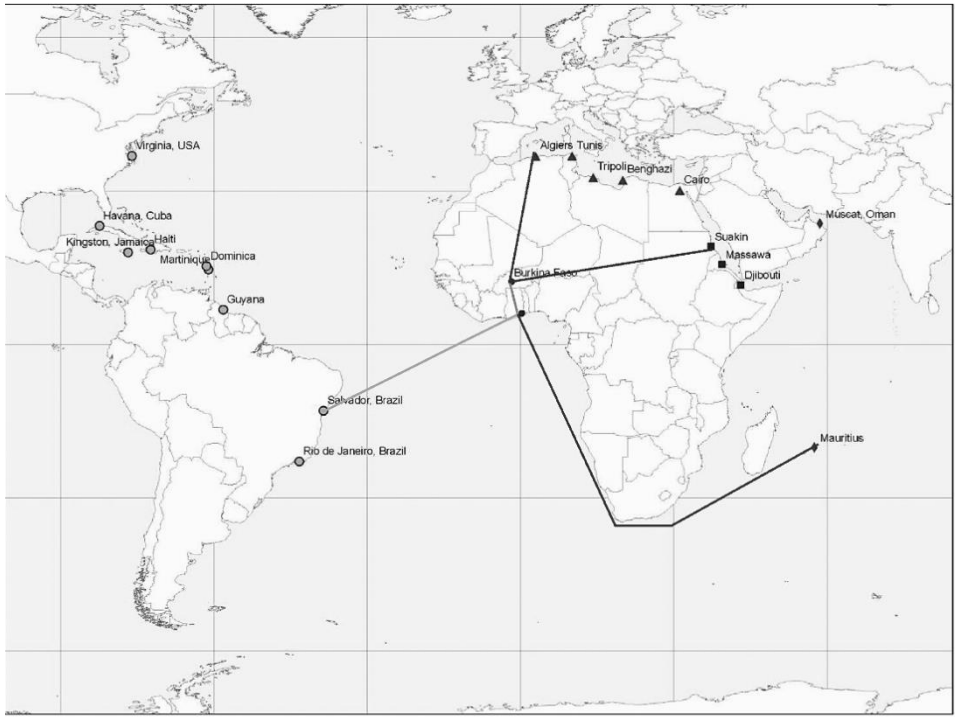


FIGURE V
Example Showing the Distance Instruments for Burkina Faso

Distances as instruments

Issues with this type of strategy:

- Distance from X could be correlated with market access trade routes, other economic opportunities...
- Selection into space
  - In this case: location of ports explained by location of mines and crop suitability
- Is the LATE interesting?
  - Compliers = selected just because they were close by

Spatial instruments

Instrumenting routes / road networks

Problem: road placement is endogenous

Strategies:

- Exploit geographical variation in transit cost due to climate/topography
- Instrument actual network with least cost network
  - Minimum spanning tree (graph theory): shortest path that connects all targeted nodes
- Instrument actual network with historical network
  - Roads planned, but not built as "control group"

Spatial instruments

Instrumenting routes / road networks

- Pascali (2015), Feyrer and Sacerdote (2009):

  predict patterns of maritime trade with wind patterns

- Morten and Oliveira (2014): long-run impacts of road networks on spatial allocation of economic activity in Brazil
  - Instrument road network with minimum spanning tree to connecting state capitals to Brasilia
  - In the simplest case: instrument "having a road" with "being on a straight-line path between a state capital and Brasilia".

Spatial instruments

Instrumenting routes / road networks

- Faber (2014): roads and trade integration in China
  - Uses data on topography to calculate least cost routes between major Chinese cities
  - Uses spanning tree algorithm to construct the least cost network
  - Instrument Chinese highway network using this network

Spatial instruments

Instrumenting routes / road networks

- Historic/planned network as an instrument for current network:
  - Donaldson (2015): "Railroads of the Raj" and economic development in India
  - Baum-Snow et al. (2012): roads and decentralization in Chinese cities

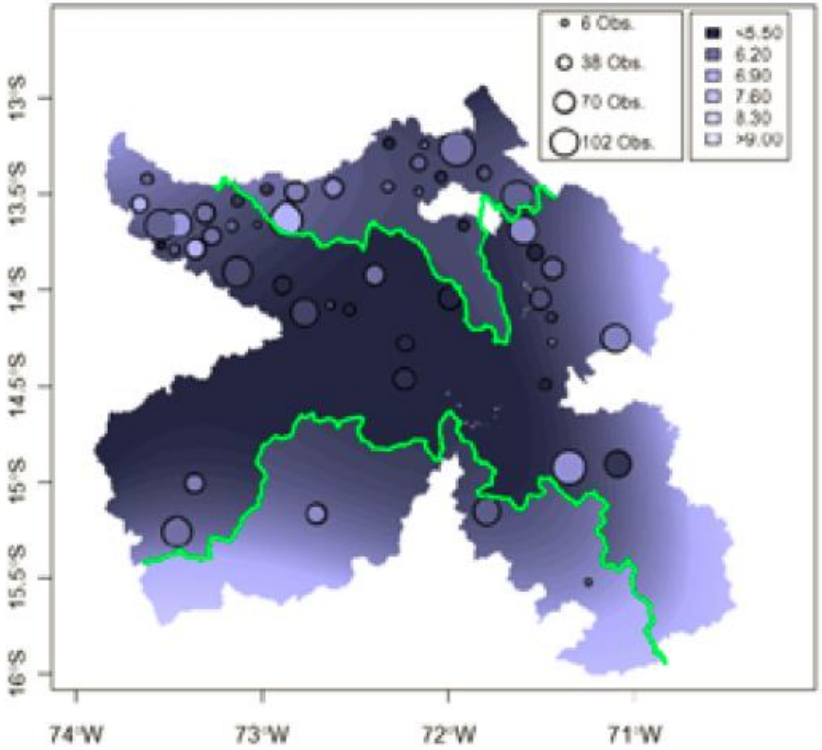
Spatial RD

- A regression discontinuity set up where
  - the discontinuity occurs on a boundary in 2-dimensional space
  - forcing variable is a two-dimensional vector: latitude and longitude

Spatial instruments

Spatial RD

- Dell (2010): long-run impacts of "mita" (forced labor system in place during Spanish colonial rule in Peru)
  - Strategy: discontinuities at the boundary of "mita" regions



(a) Consumption (2001)

Spatial instruments

Spatial RD

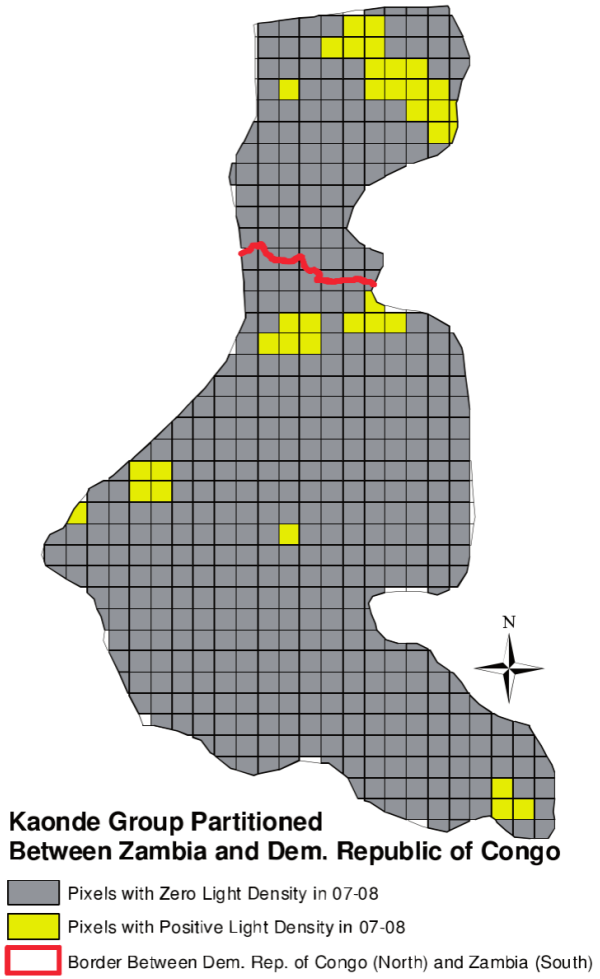
Estimation strategies:

- Turn it into a scalar RD:
  - Project coordinates into distance from the boundary
  - RD polynomials in distance
  - Control for boundary segment fixed effect
  - Disadvantage: can't estimate heterogeneous effects at different boundary points
- Boundary RD:
  - RD polynomials in latitude and longitude
  - Disadvantage: requires a lot of observations near the boundary

Spatial instruments

Spatial RD

- Michalopoulos & Papaioannou (2014)
- Look at ethnic homelands split by national borders in Africa to identify effect of national institutions
- Outcome: luminosity
- Find heterogeneous effects by ethnicity



Spatial instruments

Propagation patterns

- Hanna and Oliva (2015): impact of pollution on labor supply of Mexico City households
  - Consider closure of a refinery (time variation)
  - Main strategy: use distance from the refinery as cross-sectional source of variation
  - But there could be confounding factors we highlighted when discussing distance as instruments
  - Complementary strategy: use wind patterns as an additional source of variation (orthogonal to distance from refinery)

Spatial instruments

Propagation patterns

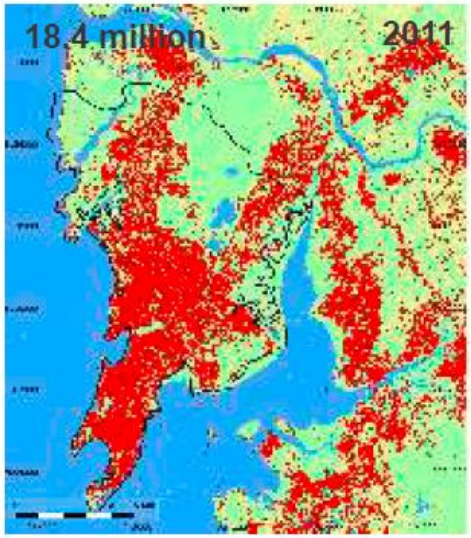
- Olken (2009): do TV and radio destroy social capital?
  - Instrument TV/radio penetration using model of electromagnetic signal propagation + topography

Roadmap

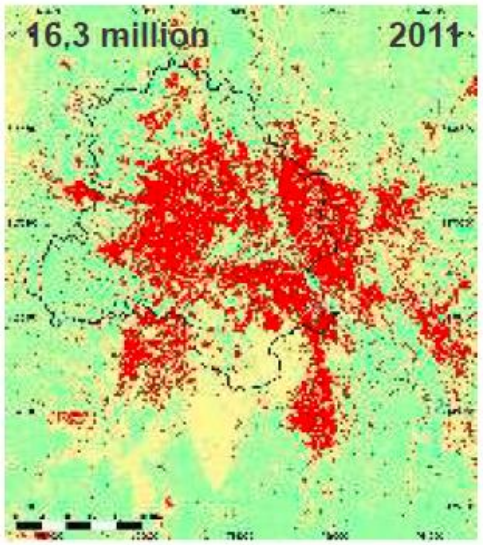
Outline

- Intro: spatial data
- What to do with spatial data
  - Spatial correlation / dependence
  - **Spatial data & identification**
    - Spatial data / distances as instruments
    - **Example from Harari (2016) "Cities in Bad Shape: Urban Geometry in India"**
  - **Spatial data as outcomes**

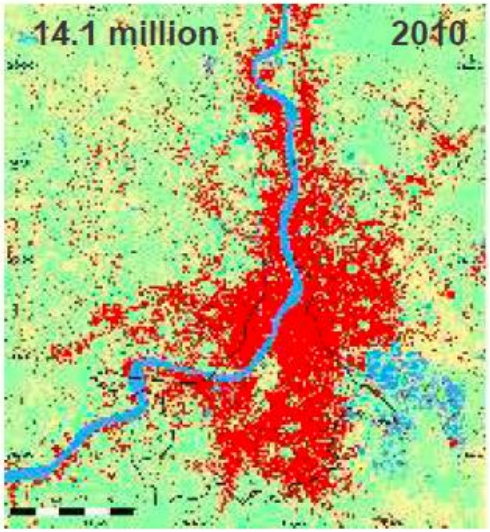
- This study focuses on one particular feature of urban form: the **spatial layout of urban footprints**



Mumbai



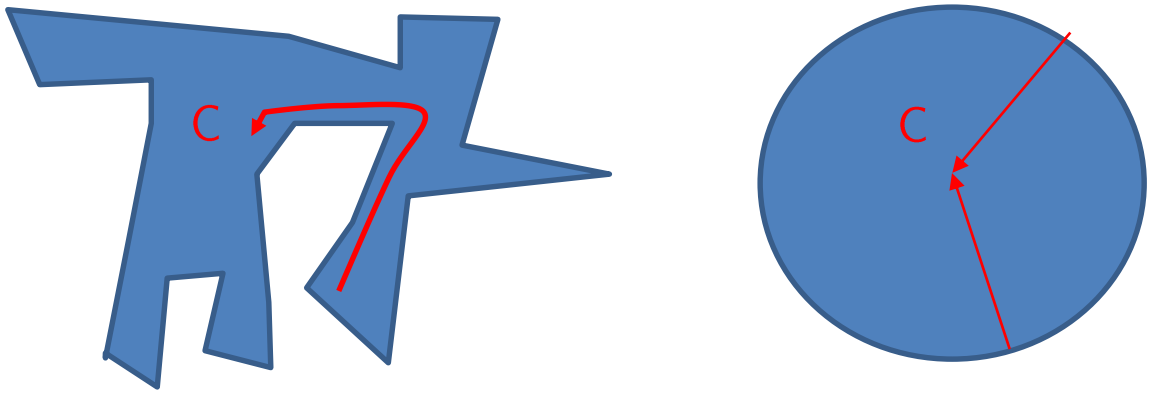
Delhi



Kolkata

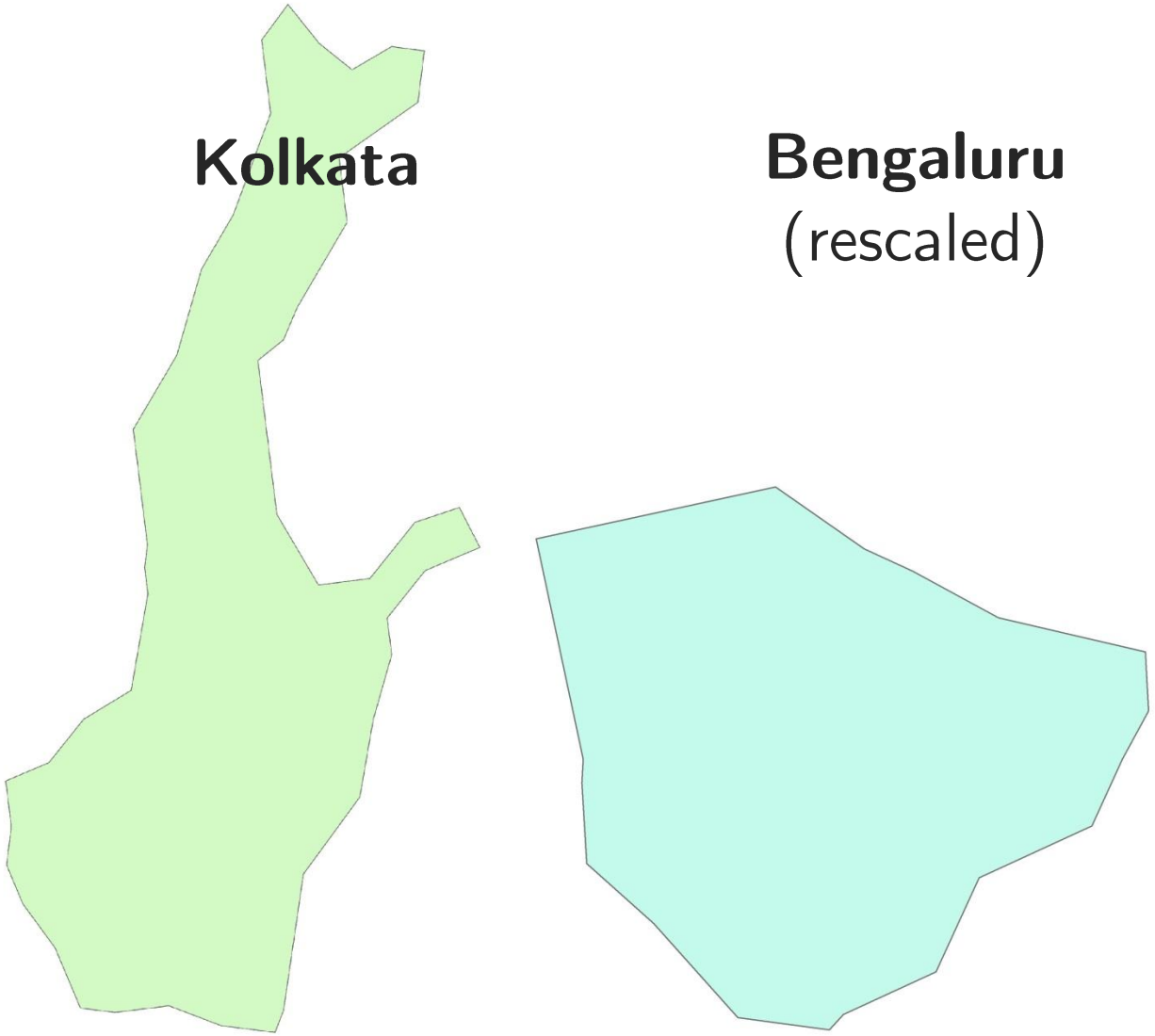
- This study focuses on one particular feature of urban form: the **spatial layout of urban footprints**

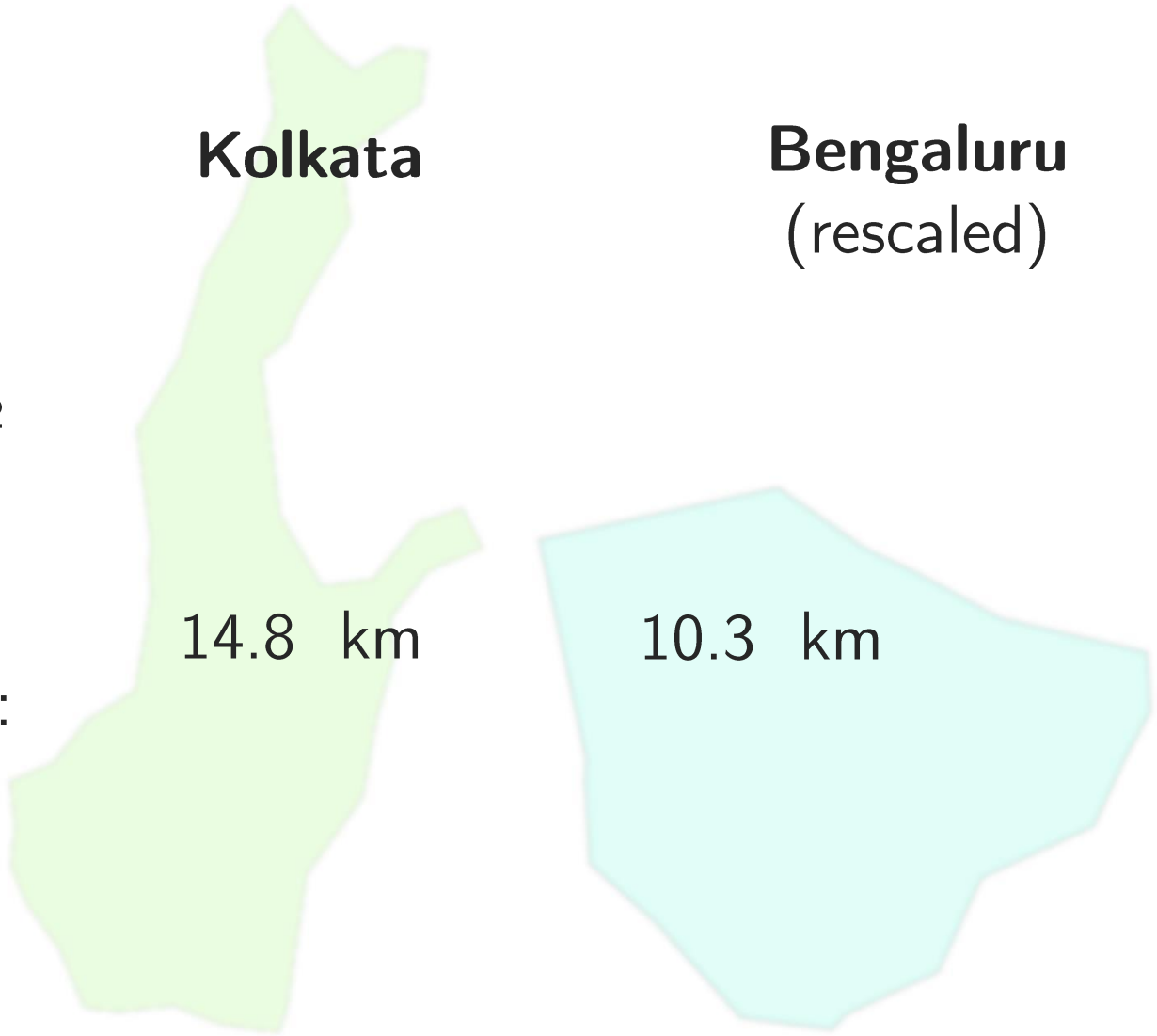
- City shape affects commuting efficiency: all else being equal, **more compact geometries = shorter within-city distances**

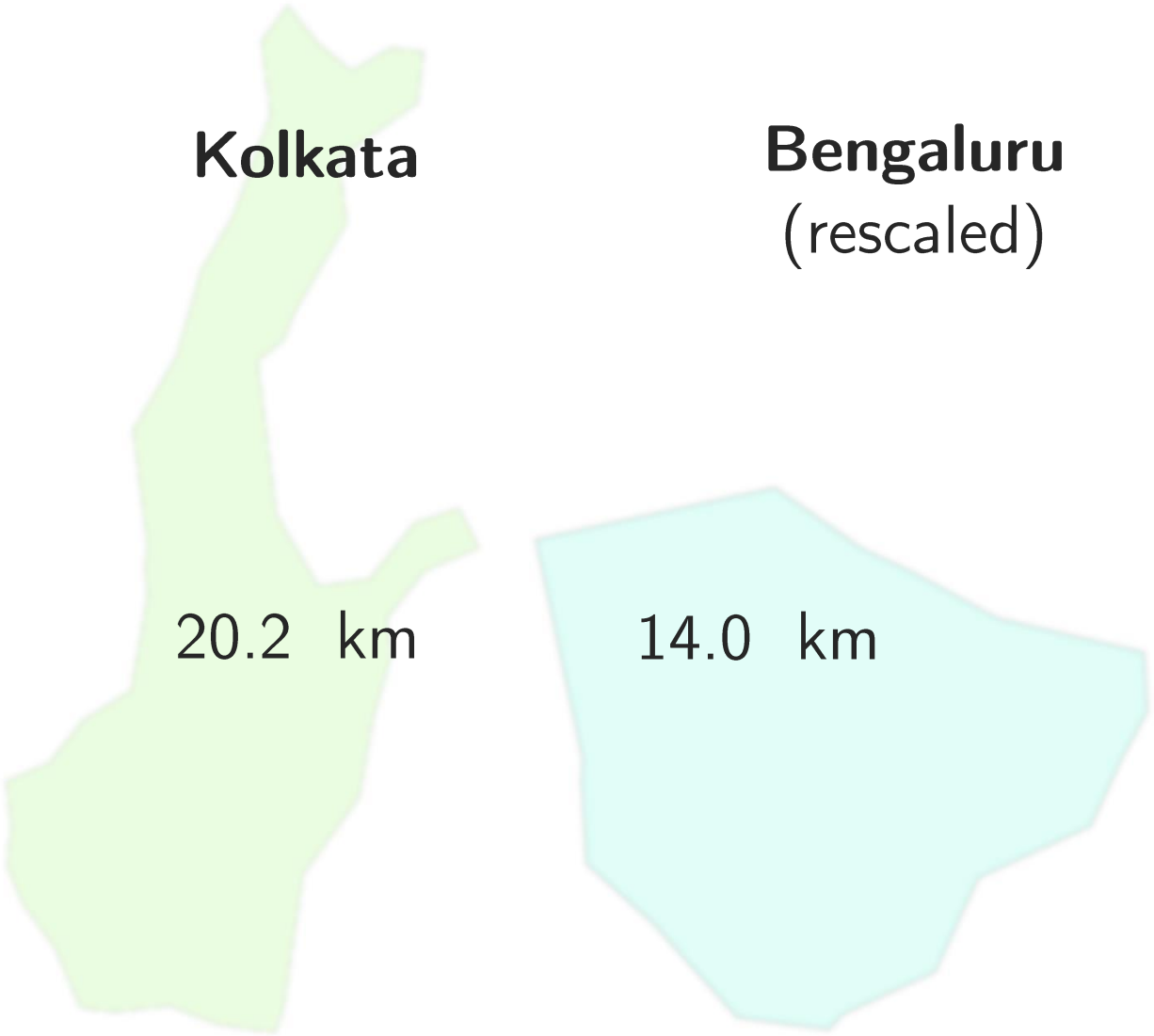


- This study focuses on one particular feature of urban form: the **spatial layout of urban footprints**
- City shape affects commuting efficiency: all else being equal, **more compact geometries = shorter within-city distances**
- Does city shape matter from an economics standpoint?
  - What influence does a city's shape have on the location decisions of consumers and firms?
  - Do households and firms benefit from locating in cities with particular shapes?

Area: 702 km^2




Kolkata
Bengaluru
(rescaled)
Area: 702 km^2
Avg. distance
to the centroid:
14.8 km
10.3 km


Kolkata
Bengaluru
(rescaled)
Area: 702 km²
Avg. within-city distance:
20.2 km
14.0 km

- **Are households / firms hurt by locating in a city like Kolkata?**
- **If so, by how much?**
- **If so, how should policy respond?**

This paper:

Empirical investigation of the economic implications of urban geometry in India

- Shape as a shifter of *potential* commuting distances
- Do consumers / firms value compact shape?
- Revealed preference Rosen-Roback framework
- Exogenous variation in shape due to geography

A unique setting:

- Rapid urbanization: observe city shape as it evolves
  - We need enough variation in city shape!
- Diffused urbanization: large sample
- Policy relevance:
  - 40% of Indian population will be urban by 2030
  - Highly debated regulatory tools:
    e.g. vertical limits

Methodology

- Panel of 460 cities over 1951-2011 (with gaps)
  - Outlines of urban footprints
  - Micro-geography
  - Outcome variables from Census / surveys
    - Main outcomes: population, wages, rents

Methodology

- Identification challenge: city shape is endogenous
  - Determined by city growth, infrastructure, land use regulations, urban planning...
  - Compact cities could be systematically different in dimensions other than geometry
- Identification strategy: exploit exogenous variation in city shape due to topography
  - Use geography to construct a time-varying instrument
  - What happens when a city becomes less compact, as a result of hitting a geographic obstacle?

Model

Spatial equilibrium across cities - Rosen (1979), Roback (1982)

- Consumers and firms optimally chose in which city to locate
- In equilibrium they must be indifferent across cities
  - Indirect utility and profits equalized
- Cities have different levels of consumption and production amenities
  - Consumption amenities make consumers better off
  - Production amenities make firms more productive

▶ Factor prices – wages, rents – allocate people and firms across cities

Revealed preferences approach

- Prediction of the Rosen-Roback model:
  - Differences in rents net of wages across cities are informative of differences in consumption amenities
  - Differences in population and wages across cities are informative of differences in production amenities
- Can compact city shape be viewed as a consumption / production amenity?
  - Consider the response of population, wages, rents to changes in city shape
  - Elicit how much households and firms value compact city shape

(A) Consumers:

$$max_{C,H}\ U(C, H, \theta)\ \ s.t.\ C = W - p_H H$$

- Consumption C, housing H, Wages W, rents p_H
- Consumption amenities θ

(B) Firms :

$$max_{N,K}\ \ Y(N, K, \bar{Z}, A) - WN - K$$

- Workers N, traded capital K, non-traded capital $\bar{Z}$
- Production amenities A

(C) Developers:

$$max_{H}\ p_H H - C(H)$$

- $H = Lh =$ land * height

Equilibrium:

(1) consumers' optimal location choice
(2) firms' labor demand
(3) housing market equilibrium

$\rightarrow$ population N, wages W, rents p_H as functions of amenities A, θ

If compact shape is a pure consumption amenity, compact cities should have:

- larger population
- lower wages
- higher rents

Equilibrium:

(1) consumers' optimal location choice
(2) firms' labor demand
(3) housing market equilibrium

$\rightarrow$ population N, wages W, rents p_H as functions of amenities A, θ

If compact shape is a consumption and production amenity, compact cities should have:

- larger population
- ? wages
- higher rents

Tracing urban footprints over time

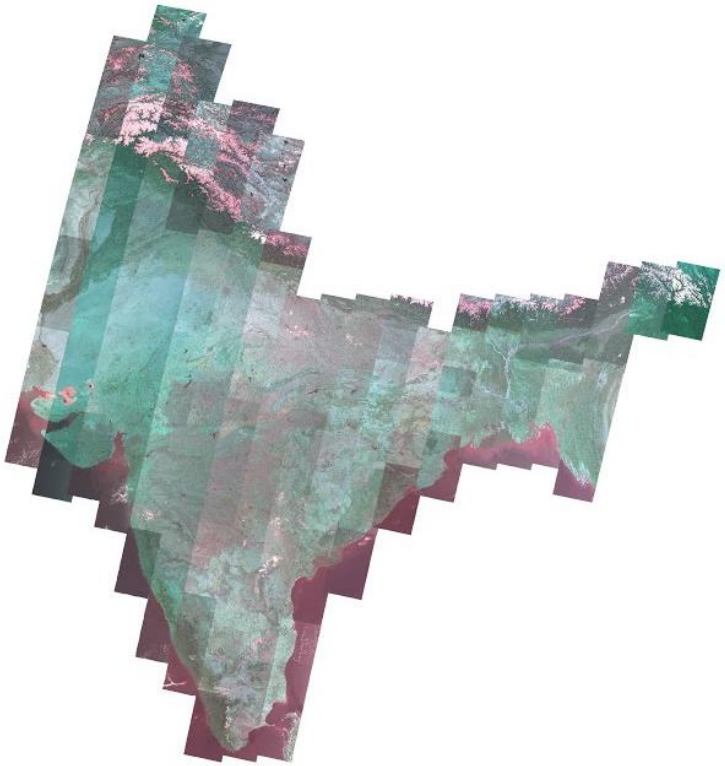
Tradeoffs between possible data sources

- Administrative city boundaries: rarely updated, not reflective of actual built-up area, arbitrary, endogenous to urban development (through regulation)
- Maps: not clear what original source is, probably not uniform across cities
  - If available, remote sensing imagery may be preferable

Tracing urban footprints over time

Tradeoffs between possible data sources

- **Day-time** imagery (e.g. Landsat)



Tracing urban footprints over time

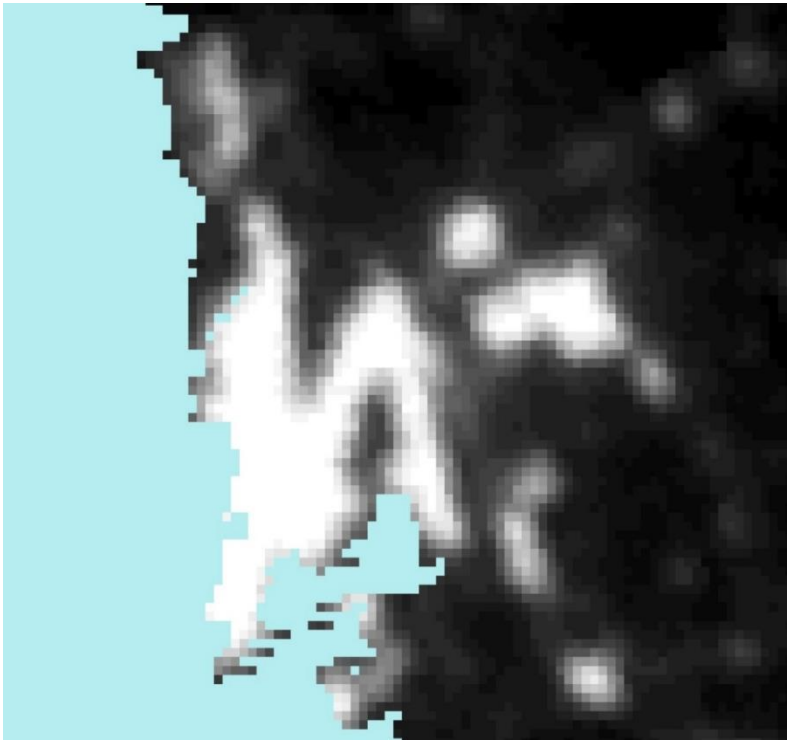
Tradeoffs between possible data sources

- **Day-time** imagery (e.g. Landsat, Modis, Quickbird satellites / Google Earth etc.)
  - Low temporal frequency: cross-sections / very short time series
  - Raw imagery requires extensive processing to classify pixels as built up and non-built up
    - Can only be automated to a limited extent
    - Accuracy depends on the availability of other sources for cross-checking (e.g. aerial photos)
  - Few pre-processed products (land use rasters) have global coverage (e.g. Global Human Settlements Layer)

Tracing urban footprints over time

Tradeoffs between possible data sources

- **Night-time** imagery: DMSP/OLS night-time lights

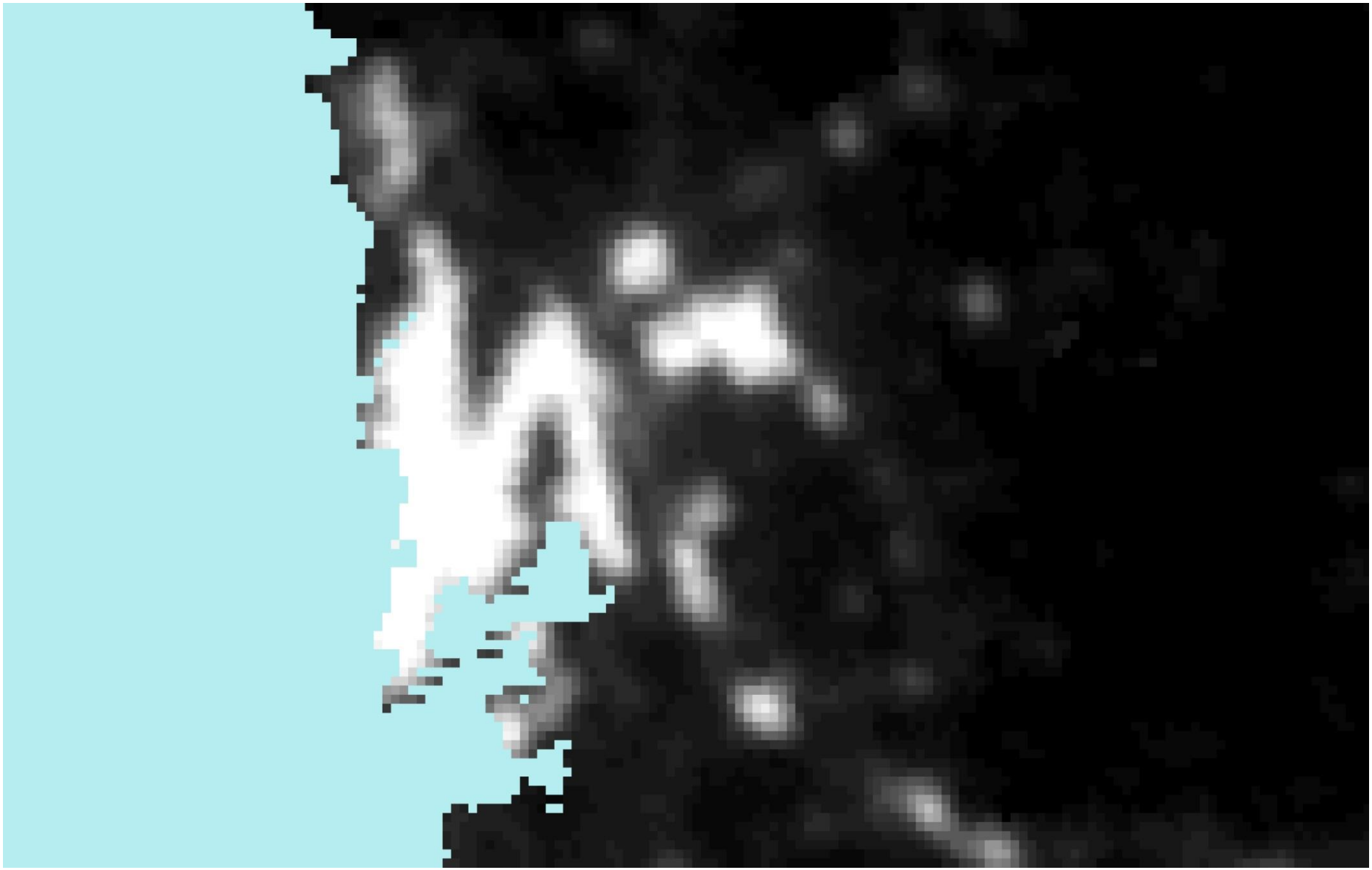


Tracing urban footprints over time

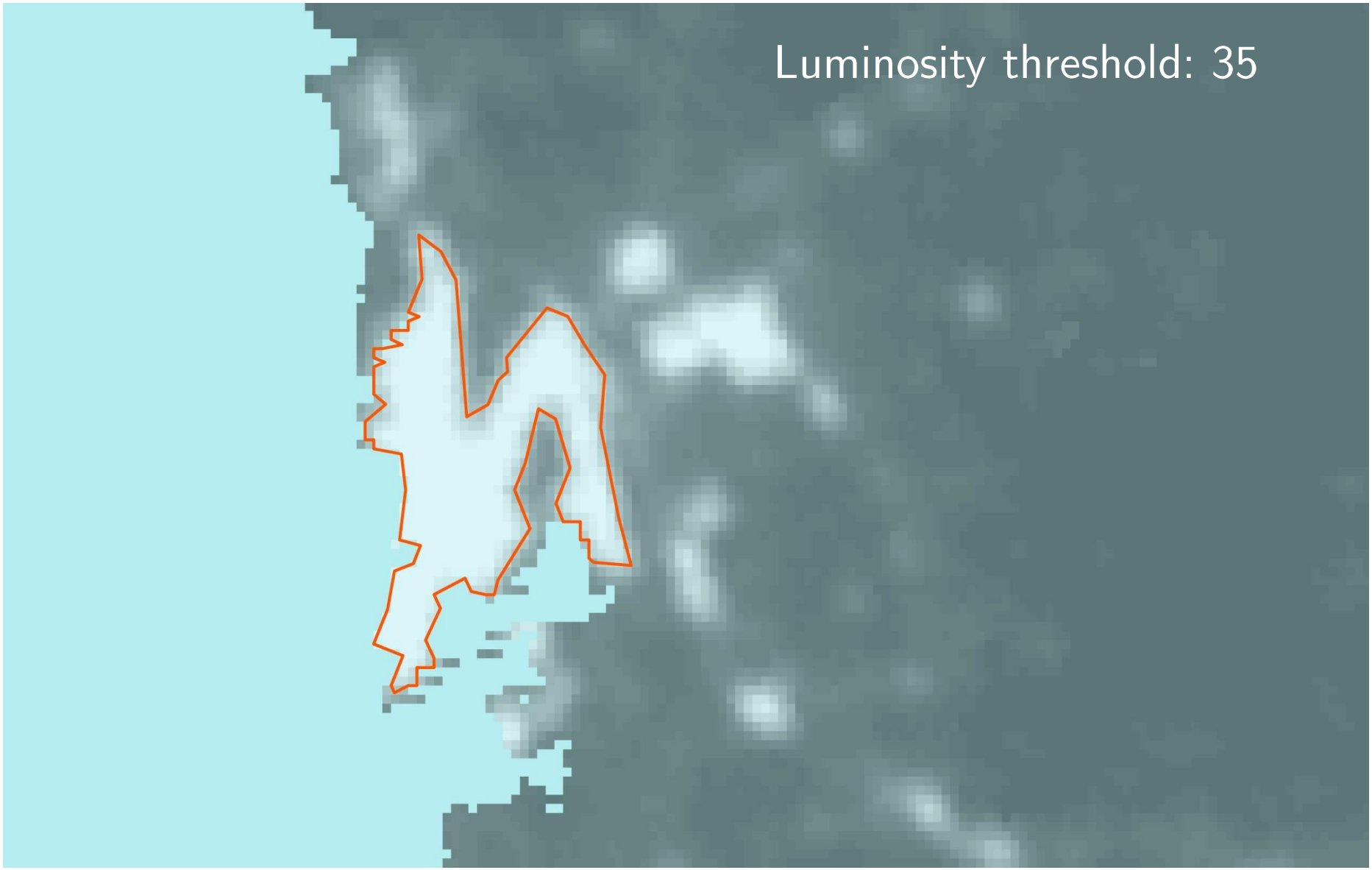
Tradeoffs between possible data sources

- **Night-time** imagery: DMSP/OLS night-time lights
  - Yearly frequency, time series 1992-2010
  - Pre-processed: available as a 1km by 1 km raster coding luminosity on a scale 0 to 63
  - Global coverage
  - Allows for immediate, replicable, objective way of tracing urban areas – set luminosity threshold (that can be varied for robustness)

Harari (2016)
Data



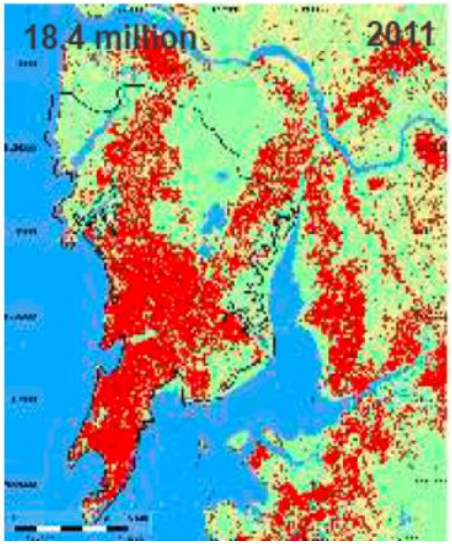
Harari (2016)
Data



Tracing urban footprints over time

Tradeoffs between possible data sources

- Urban mapping through night-time lights: **measurement error**
  - Low resolution
  - "Blooming" effects
  - Satellites with different calibration across years
  - Luminosity correlated with income
- **Solutions:**
  - City and year FE
  - IV strategy



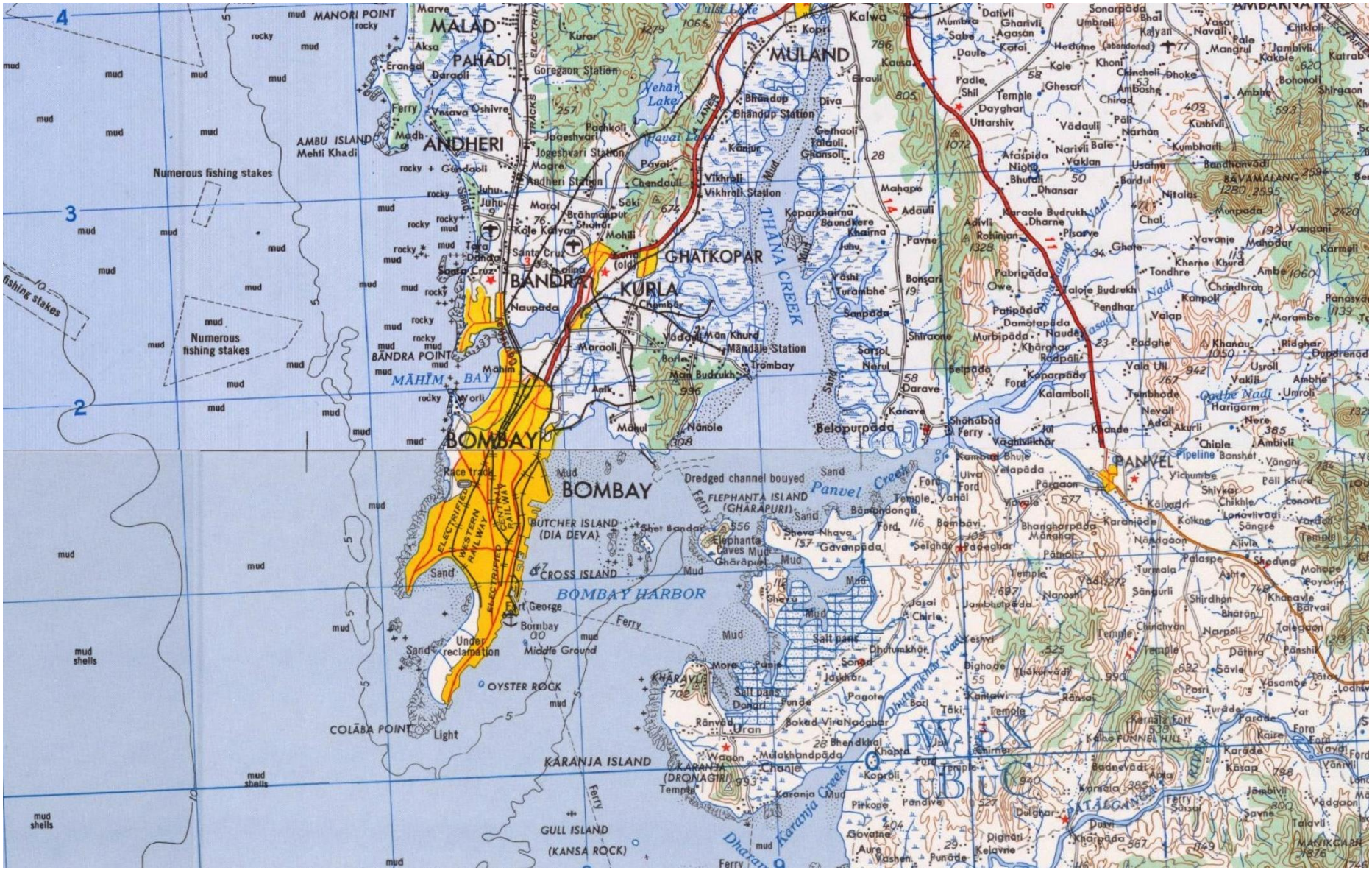
Tracing urban footprints over time

- 1992-2010: DMSP/OLS night-time lights (res. 1 km)

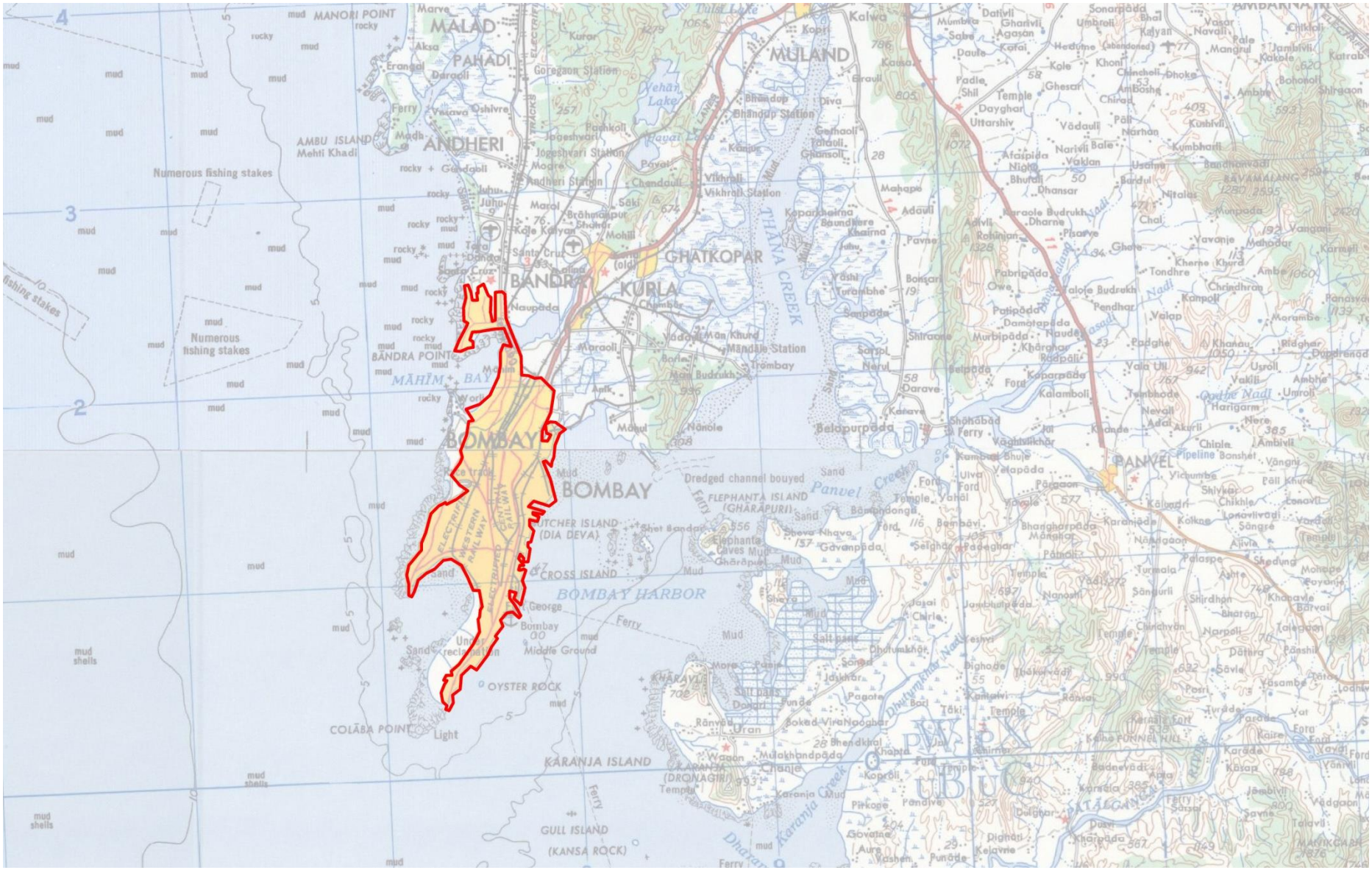
  +

- 1950 : India and Pakistan Topographic Maps, Series U502, 1:250,000, U.S. Army Map Service

Harari (2016)
Data

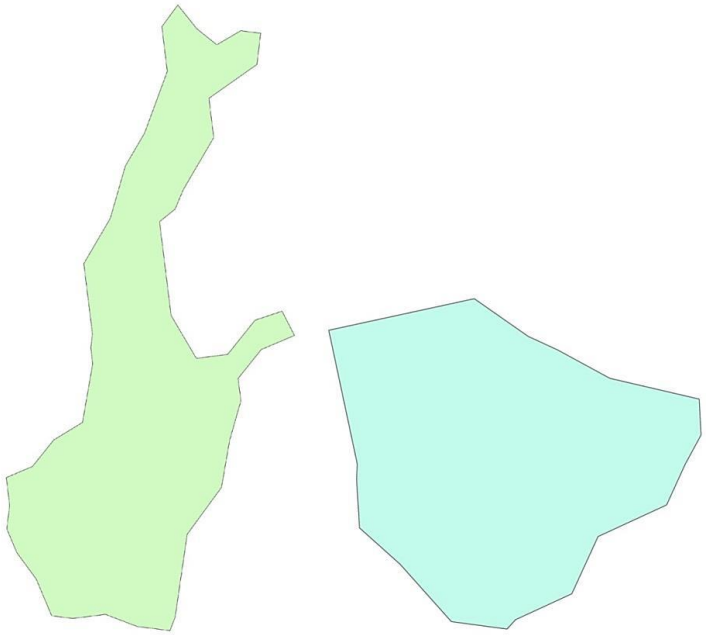


Harari (2016)
Data



Shape metrics (Angel, Civco and Parent, 2009)

- **Disconnection**
  Avg. distance between all pairs of interior points
- **Remoteness**
  Avg. distance between interior points and centroid
- **Range**
  Max. distance between two points on the shape perimeter
- All shape metrics correlated with footprint size: can be normalized by radius of equivalent-area circle



	Kolkata		Bengaluru	
Shape metric		Normalized		Normalized
remoteness, km	14.8	0.99	10.3	0.69
disconnection, km	20.2	1.35	14	0.94
range, km	62.5	4.18	36.6	2.45

Geographic constraints

- ASTER Digital Elevation Model (res. 30m)
  - compute slope raster from elevation raster
- Global MODIS raster water mask (res. 500 m)
  - binary, water or not
- "Constrained" = (water=1 or slope $>$ 15%)

- Population: Census of India, 1871-2011
- Wages : Annual Survey of Industries
  (1990, 1994, 1995, 1997, 2009, 2010)
- Housing rents:
  NSS Household Consumer Expenditure Survey
  (2005, 2006, 2007)

} District urban averages

- Other data
  - Floor Area Ratios (Sridhar, 2010)
  - Directory of Establishments (2005 Economic Census)

Note: no data at the sub-city level, no commuting data

Descriptive Statistics

	Obs	Mean	St.Dev.	Min	Max
Area, km^2	6276	62.63	173.45	0.26	3986.02
Remoteness, km	6276	2.42	2.22	0.20	27.43
Disconnection, km	6276	3.30	3.05	0.27	38.21
Range, km	6276	9.38	9.11	0.86	121.12
Norm. remoteness	6276	0.71	0.06	0.67	2.10
Norm. disconnection	6276	0.97	0.08	0.91	2.42
Norm. range	6276	2.74	0.35	2.16	7.17
City population	1440	422869	1434022	5822	22085130
City population density (per km^2)	1440	15011	19124	432	239179
Avg. yearly wage, thousand 2014 Rs.	2009	93.95	66.44	13.04	838.55
Avg. yearly rent per m^2, 2014 Rs.	896	603.27	324.81	104.52	3821.59
Avg. yearly rent, thousand 2014 Rs.	1574	24.77	10.01	8.06	147.68

$$Outcome_{c,t} = a \cdot shape_{c,t} + controls_{c,t} + u_{c,t}$$

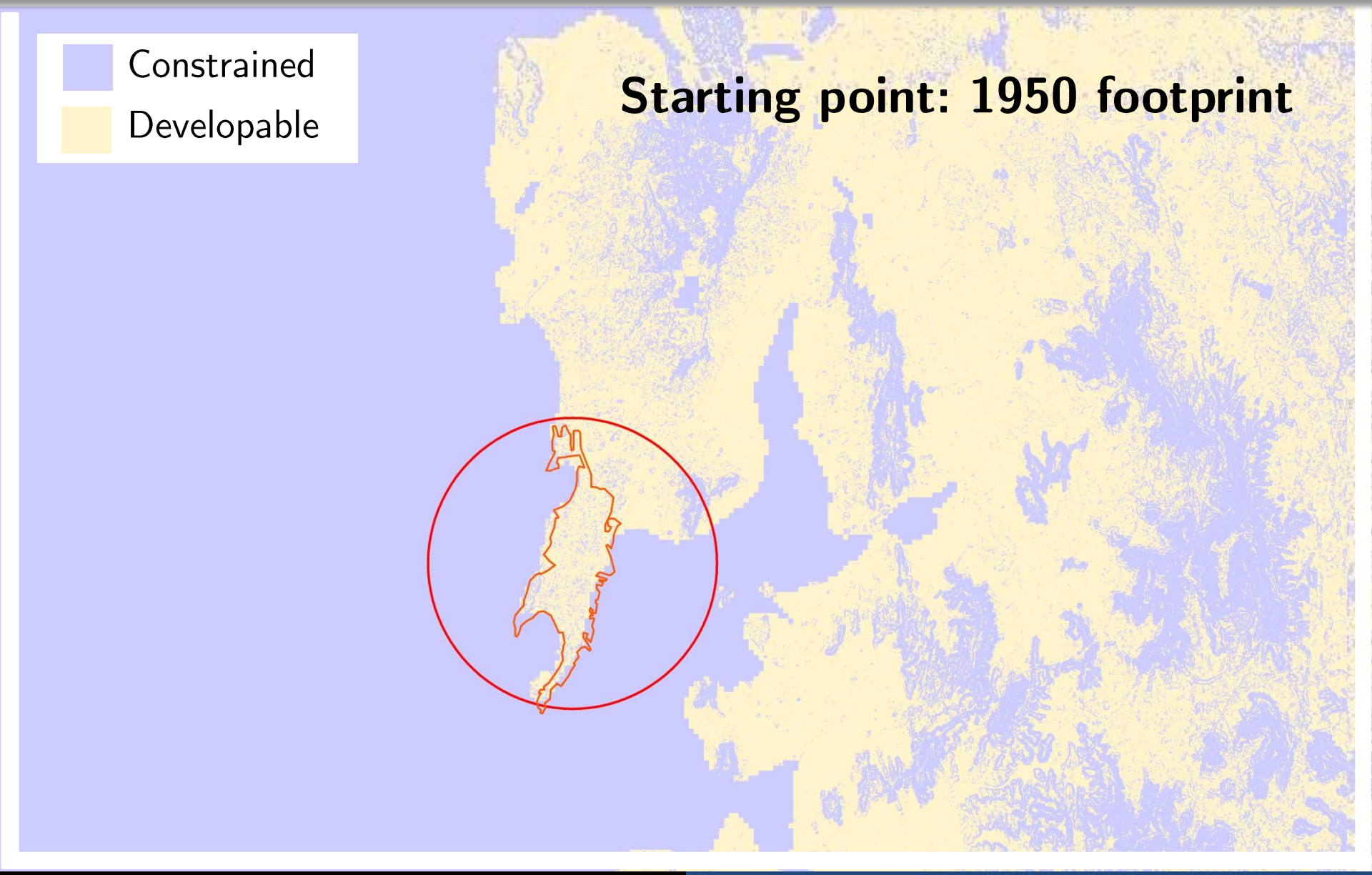
$$Outcome \in \{N, W, p_H\}$$

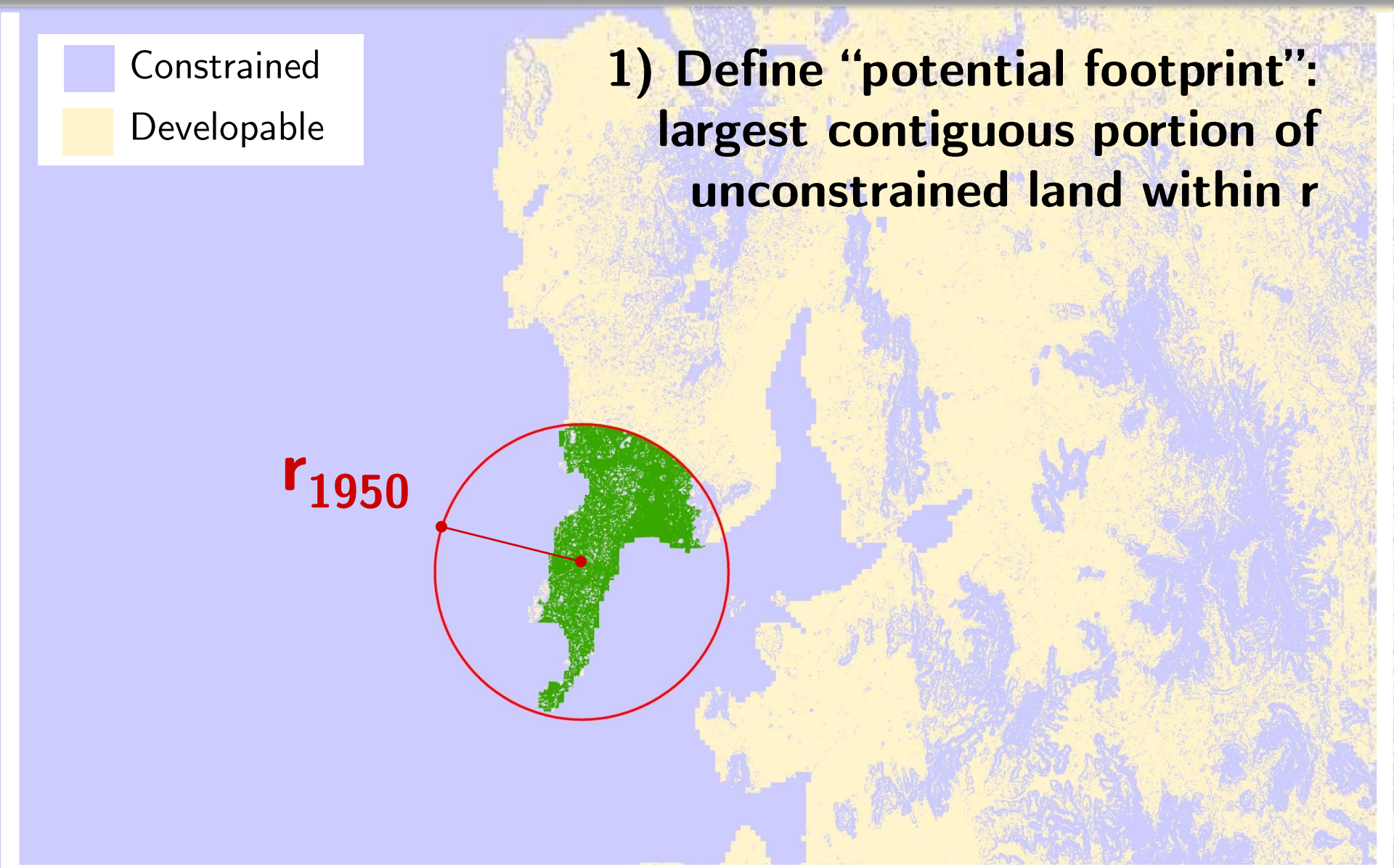
- City shape is endogenously determined by land use, planning, infrastructure, growth...

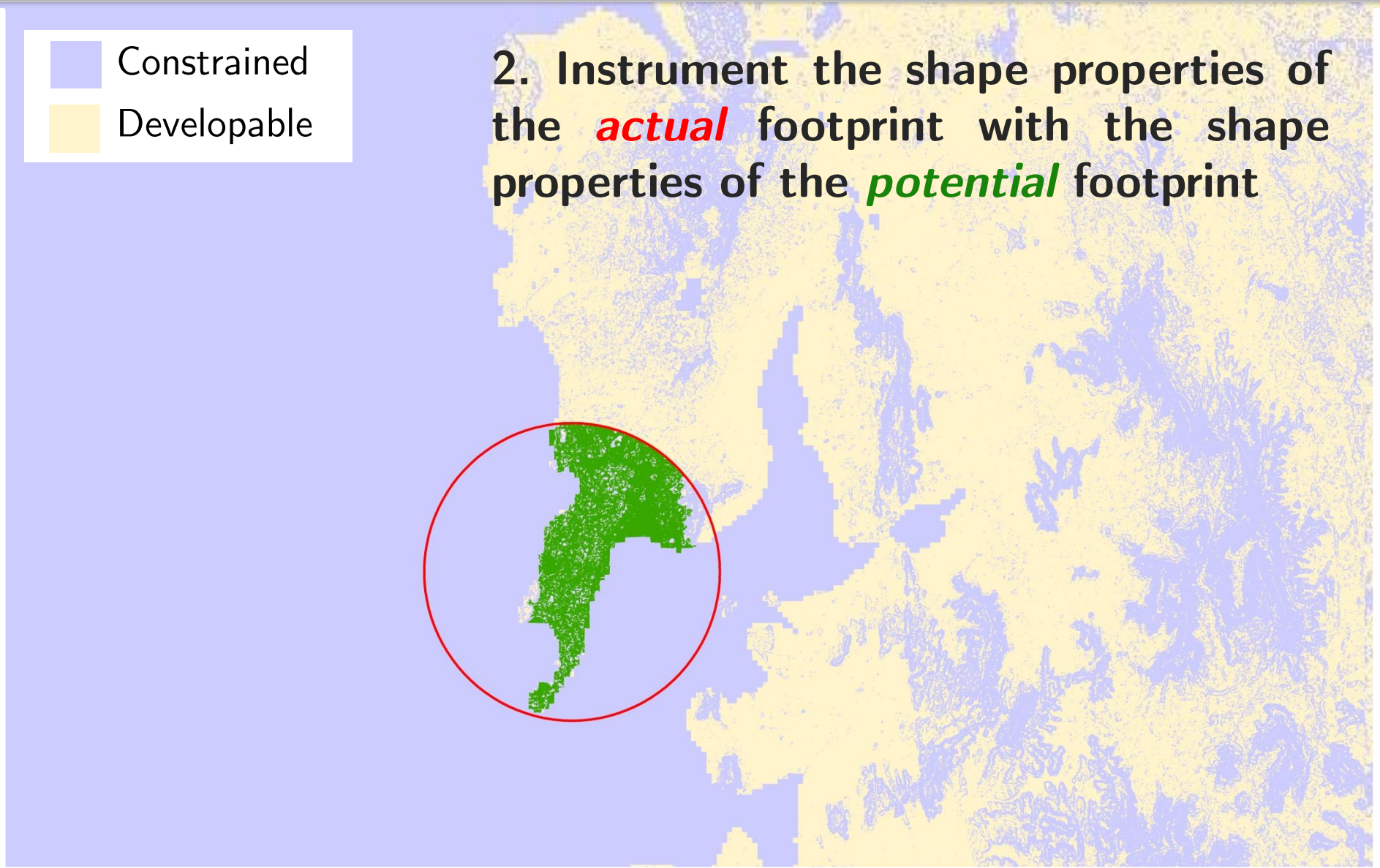
$$Outcome_{c,t} = a \cdot shape_{c,t} + controls_{c,t} + u_{c,t}$$

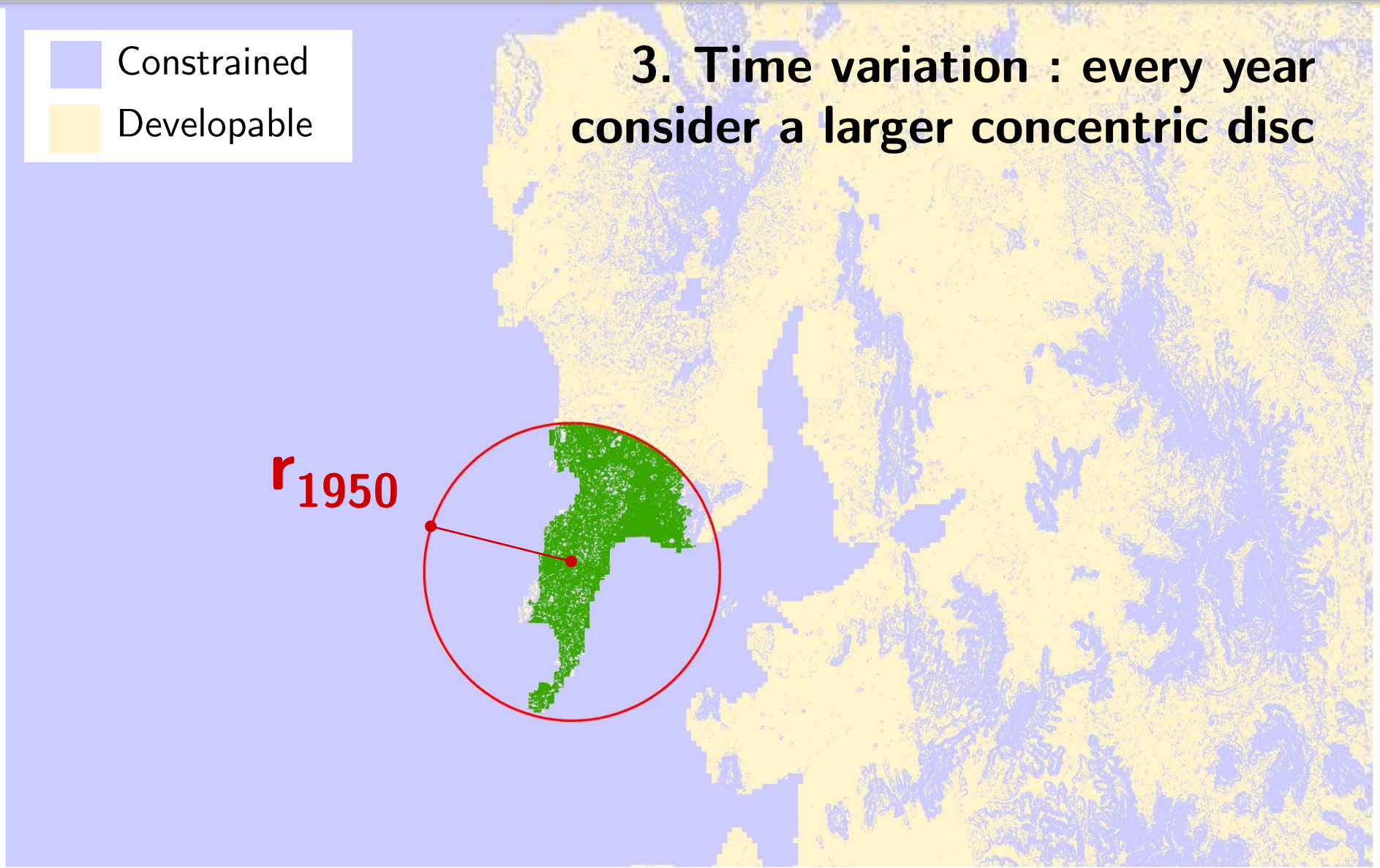
$$Outcome \in \{N, W, p_H\}$$

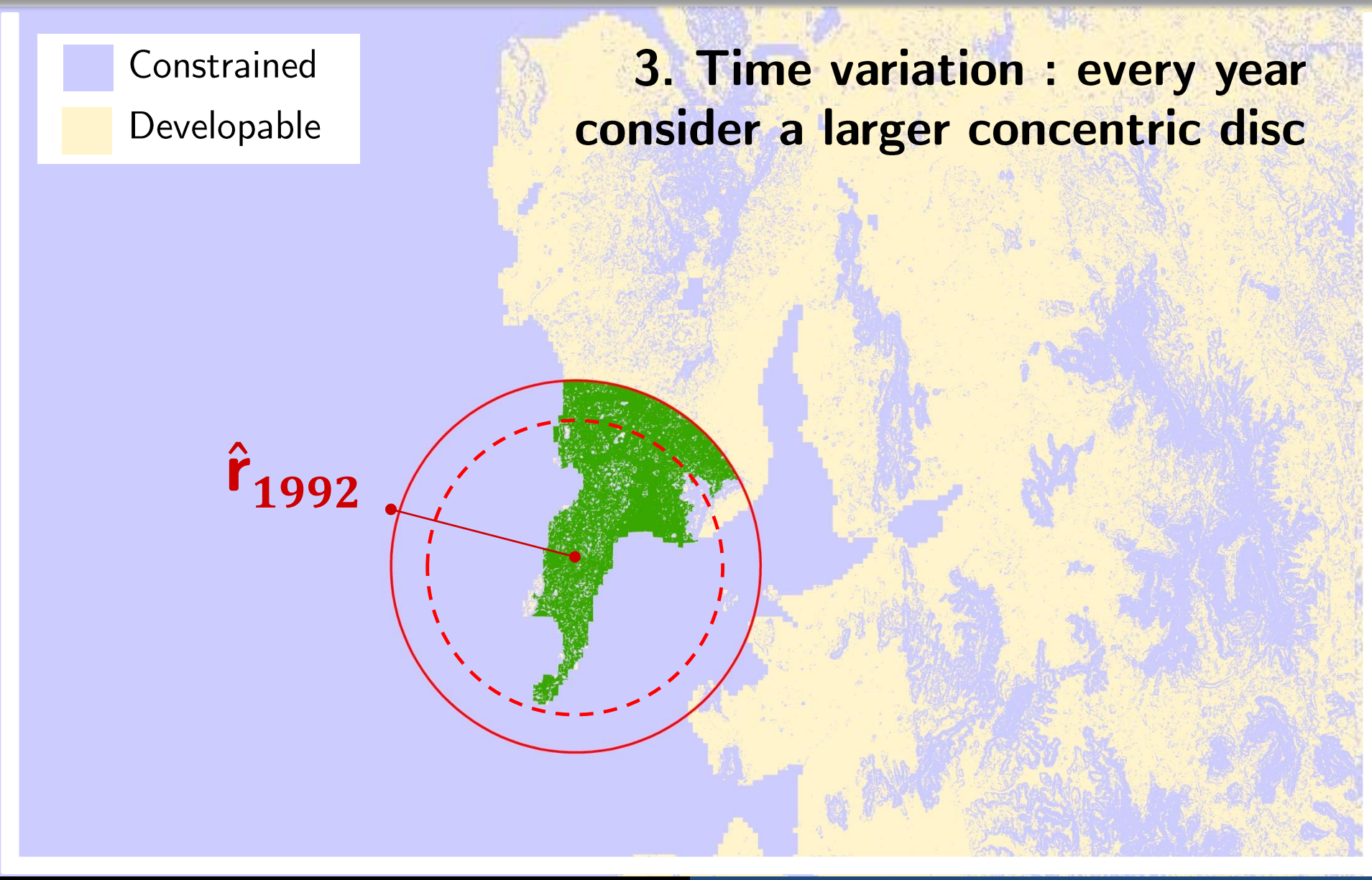
- City shape is endogenously determined by land use, planning, infrastructure, growth...
- Instrument *actual* city shape with the *potential* shape that a city can have based on the geographic constraints that surround it
  - Constraints = slopes, water
- Time variation: as cities expand, they face different topographic constraints
  - City FE, year FE

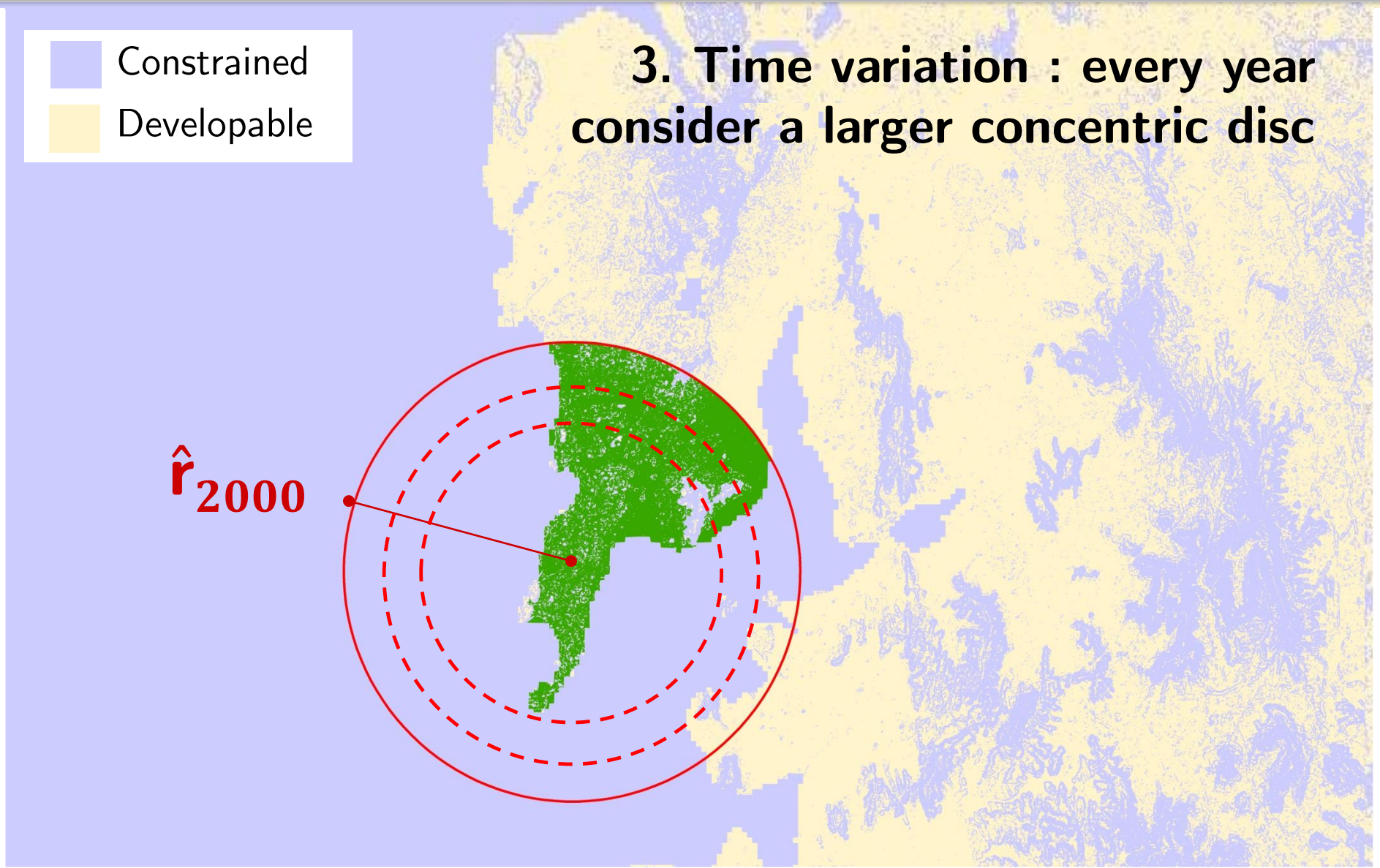

Constrained
Developable
Starting point: 1950 footprint


Constrained
Developable
1) Define "potential footprint": largest contiguous portion of unconstrained land within r
r_{1950}


Constrained
Developable
2. Instrument the shape properties of the *actual* footprint with the shape properties of the *potential* footprint


Constrained
Developable
3. Time variation : every year consider a larger concentric disc
r_{1950}


Constrained
Developable
3. Time variation : every year consider a larger concentric disc
$\hat{r}_{1992}$


Constrained
Developable
3. Time variation : every year consider a larger concentric disc
$\hat{r}_{2000}$

How is $\widehat{r_{ct}}$ determined?

- City-specific model

  rate of expansion of the radii varies across cities based on historic population growth rates

  $\widehat{pop_{c,t}}$: log-linear projection of city 1871-1951 population

- Common rate model

  rate of expansion of the radii is equal to the average expansion rate for all cities

Harari (2016)
Empirical strategy

- Variation comes from the relative position of topographic obstacles encountered over time
  - City FEs: changes in shape that a city undergoes over time
  - Explanatory power not limited to coastal/mountainous cities

- Variation comes from the relative position of topographic obstacles encountered over time
  - City FEs: changes in shape that a city undergoes over time
  - Explanatory power not limited to coastal/mountainous cities

- Exclusion restriction:

conditional on city and year FEs, this function of topography affects outcomes Y only by constraining urban shape

- Variation comes from the relative position of topographic obstacles encountered over time
  - City FEs: changes in shape that a city undergoes over time
  - Explanatory power not limited to coastal/mountainous cities

- Exclusion restriction:

conditional on city and year FEs, this function of topography affects outcomes Y only by constraining urban shape

- Threats to identification (discussed in paper):
  - Inherent amenity value of topography
  - Geographic constraints affect housing supply
  - Differential trends in cities with different initial shapes

Double instrument specification

$$\log(Y_{c,t}) = a \cdot S_{c,t} + b \cdot \log(area_{c,t}) + \mu_c + \rho_t + \eta_{c,t}$$

- 2 endogenous regressors: city shape $S_{c,t}$, $log(area_{c,t})$
- 2 instruments: shape of potential footprint $\widetilde{S_{c,t}}$, $\log(\widehat{pop_{c,t}})$

Double instrument specification

$$\log(Y_{c,t}) = a \cdot S_{c,t} + b \cdot \log(area_{c,t}) + \mu_c + \rho_t + \eta_{c,t}$$

- 2 endogenous regressors: city shape $S_{c,t}$, $log(area_{c,t})$
- 2 instruments: shape of potential footprint $\widetilde{S_{c,t}}$, $\log(\widehat{pop_{c,t}})$
- 1st stage (1):

$$S_{c,t} = \sigma \cdot \widetilde{S_{c,t}} + \delta \cdot \log(\widehat{pop_{c,t}}) + \omega_c + \varphi_t + \theta_{c,t}$$

- 1st stage (2):

$$\log(area_{c,t}) = \alpha \cdot \widetilde{S_{c,t}} + \beta \cdot \log(\widehat{pop_{c,t}}) + \lambda_c + \gamma_t + \varepsilon_{c,t}$$

Single instrument specification

Outcome $Y_{c,t}$= population:

$$\frac{pop_{c,t}}{area_{c,t}} = a \cdot nS_{c,t} + \mu_c + \rho_t + \eta_{c,t}$$

- 1 endogenous regressor: normalized city shape $nS_{c,t}$
- 1 instrument: normalized shape of potential footprint $\widetilde{nS_{c,t}}$

- 1st stage: $nS_{c,t} = \beta \cdot \widetilde{nS_{c,t}} + \lambda_c + \gamma_t + \varepsilon_{c,t}$

Potential shape (driven by geography) predicts actual shape

First Stage

	(1)	(2)	(3)
	Norm. shape of actual footprint	Shape of actual footprint, km	Log area of actual footprint, km^2
		Shape Metric: Disconnection	
Norm. shape of potential footprint	**0.0663*** **		
	(0.0241)		
Shape of potential footprint, km		**1.392*** **	0.152***
		(0.229)	(0.0457)
Log projected historic population		-1.180***	**0.307*** **
		(0.271)	(0.117)
Observations	6,276	6,276	6,276
Model for r	common rate	city-specific	city-specific
City FE	YES	YES	YES
Year FE	YES	YES	YES

Notes: each observation is a city-year. Disconnection is the average length of within-city trips. Standard errors clustered at the city level.*** p<0.01,** p<0.05,* p<0.1.

IV: bad shapes → slower population growth

Impact of City Shape on Population

	(1) IV	(2) IV	(3) OLS
	Population density	Log population	Log population
	Shape Metric: Disconnection		
Norm. shape of actual footprint	**-254.6*** **		
	(80.01)		
Shape of actual footprint, km		**-0.0991** **	0.0249***
		(0.0386)	(0.00817)
Log area of actual footprint, km^2		0.782***	0.167***
		(0.176)	(0.0318)
Observations	1,440	1,440	1,440
Model for r	common rate	city-specific	
City FE	YES	YES	YES
Year FE	YES	YES	YES

Notes: each observation is a city-year. Disconnection is the average length of within-city trips. Population density is measured in thousand inhabitants per km^2. Standard errors clustered at the city level.*** p<0.01,** p<0.05,* p<0.1.

OLS: cities tend to deteriorate in shape as they grow

Impact of City Shape on Population

	(1) IV	(2) IV	(3) OLS
	Population density	Log population	Log population
	Shape Metric: Disconnection		
Norm. shape of actual footprint	-254.6***		
	(80.01)		
Shape of actual footprint, km		-0.0991**	**0.0249*** **
		(0.0386)	(0.00817)
Log area of actual footprint, km^2		0.782***	0.167***
		(0.176)	(0.0318)
Observations	1,440	1,440	1,440
Model for r	common rate	city-specific	
City FE	YES	YES	YES
Year FE	YES	YES	YES

Notes: each observation is a city-year. Disconnection is the average length of within-city trips. Population density is measured in thousand inhabitants per km^2. Standard errors clustered at the city level.*** p<0.01,** p<0.05,* p<0.1.

Robustness:

- Different shape metrics
- Different luminosity thresholds – resulting in more or less restrictive definitions of urban areas
- Excluding cities with "extreme" topographies (coastal and mountainous)
- Initial shape x year FEs

IV: bad shapes → higher wages

Impact of City Shape on Wages

	(1) IV	(2) IV	(3) OLS
	Dependent variable: log wage		
Shape of actual footprint, km	**0.0996*** **	**0.0626**	0.0586***
	(0.0336)	(0.0536)	(0.0150)
Log area of actual footprint, km^2		-0.167	-0.00936
		(0.465)	(0.0516)
Obs.	1,075	1,075	1,075
Model for r	common rate	city-specific	
City FE	YES	YES	YES
Year FE	YES	YES	YES

Notes: each observation is a city-year. Dependent variable: log urban average of individual yearly wages in the city's district, in thousand 2014 Rupees. Sample includes only districts with one city. Shape is captured by the disconnection index, which measures the average length of trips within the city footprint, in km. Wages are from the Annual Survey of Industries, waves 1990, 1994, 1995, 1997, 1998, 2009, 2010. Standard errors are clustered at the city level. *** p<0.01,** p<0.05,* p<0.1.

IV: bad shapes → lower rents

Impact of City Shape on Rents

	(1) IV	(2) IV	(3) OLS
	Dependent variable: log yearly rent per square meter		
Shape of actual footprint, km	**-0.636**	**-0.518***	-0.00857
	(1.661)	(0.285)	(0.0736)
Log area of actual footprint, km^2		-0.919	-0.0632
		(0.870)	(0.108)
Obs.	476	476	476
Model for r	common rate	city-specific	
City FE	YES	YES	YES
Year FE	YES	YES	YES

Notes: each observation is a city-year. Dependent variable : log urban average of housing rent per m^2 in the city's district. Sample includes only districts with one city. Shape is captured by the disconnection index, which measures the average length of trips within the city footprint, in km. Housing rents are from the NSS Household Consumer Expenditure Survey, rounds 62 (2005-2006), 63 (2006-2007) and 64 (2007-2008). Standard errors are clustered at the city level. *** p<0.01,** p<0.05,* p<0.1.

Key takeaways:

- Consumers are better off in more compact cities:
  - As cities grow into more compact shapes
    - Population ↑
    - Wages ↓
    - Housing rents ↑
  - Spatial equilibrium: households pay for "good shapes" by foregoing wages and paying higher rents
- Calibrating the model:
  - implied welfare cost of "bad shape" for consumers: one std. dev. deterioration in shape = - 5% income
  - limited impact on firms' productivity: -0.1%

- Why aren't firms that affected in equilibrium?
  - Probably, because they cluster in relatively few central locations – and it is consumers facing longer commutes
- Question: are cities with worse shapes more polycentric?

- Why aren't firms that affected in equilibrium?
  - Probably, because they cluster in relatively few central locations – and it is consumers facing longer commutes
- Question: **are cities with worse shapes more polycentric**?

- Methodology: detect employment sub-centers
  - Geocode firms' addresses from Directory of Establishments in Economic Census
  - Use a non-parametric procedure (Mc Millen, 2001) to detect employment clusters within each city
  - Use number of employment sub-centers per city as outcome

Non-parametric detection of employment subcenters

Intuition:

- In a purely monocentric city, employment density declines from central business district (CBD) to periphery
- If there are sub-centers, distance to the CBD is not the only predictor of density

Detect sub-centers as locations that

- have larger employment density than nearby locations
- have a significant impact on the overall employment density gradient in a city
- Locations could be wards, zipcodes... in my case: 1km grid cells

Non-parametric detection of employment subcenters

Consider employment density y_i in grid cell i.

Step 1: find candidate sub-centers

- Fit a it a non-parametric model $y_i = f(x_i^N, x_i^E) + \varepsilon_i$
  - x_i^N is distance North from the CBD
  - x_i^E is distance East from the CBD
- Pick those location I's for which: $(y_i - \widehat{y_i}) / \widehat{\sigma_i} > 1.96.$

Non-parametric detection of employment subcenters

Step 1: find candidate sub-centers j

Step 2: select those locations, among candidate sub-centers, with significant explanatory power in the following:

$$y_i = g(DCBD_i) + \sum_{j=1}^{S} \delta_j^1 (D_{ji})^{-1} + \delta_j^2 (-D_{ji}) + u_i$$

- $DCBD_i$ = distance from the CBD
- D_{ij} = distance from candidate subcenter j

- Stepwise procedure: start with all candidate j's in the regression, then drop the one with the lowest t stat, then re-run....
- ...until all Dj's are significant.

Non-parametric detection of employment subcenters

Step 1: find candidate sub-centers j

Step 2: select those locations, among candidate sub-centers, with significant explanatory power in the following:

$$y_i = g(DCBD_i) + \sum_{j=1}^{S} \delta_j^1 (D_{ji})^{-1} + \delta_j^2 (-D_{ji}) + u_i$$

- $DCBD_i$ = distance from the CBD
- D_{ij} = distance from candidate subcenter j

- Stepwise procedure: start with all candidate j's in the regression, then drop the one with the lowest t stat, then re-run....
- ...until all Dj's are significant.

Cities with bad shapes are more monocentric

Impact of City Shape on the Number of Employment Subcenters, 2005

	(1)	(2)	(3)
	IV	IV	OLS
	Shape Metric: Disconnection		
	Subcenters/km^2	Log subcenters	Log subcenters
Norm. shape of actual footprint	**-0.371**		
	(0.507)		
Shape of actual footprint, km		**-0.0639***	-0.0579***
		(0.0379)	(0.0154)
Log area of actual footprint, km^2		0.611***	0.571***
		(0.125)	(0.0568)
Observations	188	188	188
Model for r	common rate	city-specific	

Notes: each observation is a city in year 2005. Number of employment subcenters computed following Mc Millen (2001). Data on firms' addresses and employment are from the Economic Census (2005). Disconnection is the average length of trips within the city footprint, in km. *** p<0.01,** p<0.05,* p<0.1

Other results:

- Channels:
  - Infrastructure mitigates the negative effects of poor geometry on population
- Interactions of policy and city shape: land use regulations
  - Restrictions on building height (FARs) result in larger and less compact footprints

Key takeaway:

- Spatial configuration of cities matters for quality of life
  - Example of spatial patterns as outcomes

Roadmap

Outline

- Intro: spatial data
- What to do with spatial data
  - Spatial correlation / dependence
  - Spatial data & identification
  - **Spatial data as outcomes**

Spatial data as outcomes

Spatial data as proxies for outcomes that we can't measure directly

...at the appropriate scale:

- Nighttime light intensity: proxy for economic activity / development
  - 4 DMSP-OLS Nighttime Lights Time Series
  - Idea: many forms of production, consumption, and public goods emit light
  - Henderson et al. (2012): changes in luminosity are correlated with changes in GDP
  - Pinkovskiy and Sala-i-Martin (2016): accuracy of GDP versus household surveys – crosswalk through "weights"

Spatial data as outcomes

Spatial data as proxies for outcomes that we can't measure directly

...at the appropriate scale:

- Advantages of Nighttime Lights data as proxies for income:
  - 1 km x 1 km raster format – flexible
  - Pixel values can be aggregated for any geographical unit - not only administrative units
  - E.g.: Alesina, Michalopoulos and Papaioannou (2014): calculate luminosity at the ethnic homeland level and compute measure of "ethnic inequality" within countries

Spatial data as outcomes

Spatial data as proxies for outcomes that we can't measure directly

...in an accurate way:

- Data that may be subject to manipulation:
  - (Illegal) deforestation: Burgess et al. (2102), Alesina et al. (2014), Burgess, Costa and Olken (2016), Jayachandran et al. (2016)
  - Illegal crops
  - Land values (for tax assessment purposes):
    - E.g.: use LIDAR data on building heights as a proxies for "true" realized land values in Jakarta (Harari and Wong, in progress)

Spatial data as outcomes

Spatial data: where I think the frontier is

- Very advanced image recognition techniques
  - Henderson et al (2016) and Suri (2015): recognize individual buildings
  - Glaeser et al. (2015) and Naik et al. (2015): machine learning algorithm predicts income using Google Streetview images
    - So far, in the US, but possible extension in slums in developing countries?

Spatial data as outcomes

Spatial data: where I think the frontier is

- Tracking people's movements with precision
  - Transit data from Google Maps traffic information
  - Cell phone location data (Kreindler and Miyauchi 2015)

Conclusion

Spatial methods and data can help answer interesting questions in creative ways

Further readings:

- Masayuki Kudamatsu's "GIS for economics research" course notes: https://sites.google.com/site/mkudamatsu/gis
- Solomon Hsiang's blog (with code): http://www.fight-entropy.com/
- Melissa Dell's "ArcGis for applied economists" notes
- Donaldson and Storeygard (forthcoming), Journal of Economic Perspectives article on remote sensing in economics