A Spatial Database for South Asia

Yue Li, Martin Rama, Virgilio Galdo and Maria Florencia Pinto

Office of the Chief Economist for South Asia, World Bank

October 15, 2015

About

Collaboration

- A product of the Office of the Chief Economist, South Asia Region.
- Benefited from close collaboration with:
  - Poverty GP: India Poverty and Shared Prosperity Cluster program
  - GP SURR: South Asia Regional Urbanization Flagship Report
  - GP SURR: Global Research Program on Spatial Development of Cities

Financing

- World Bank budget
- The Department for International Development of the United Kingdom as part of the Sustainable Urban Development Multi-Donor Trust Fund.
- The Partnership for South Asia Trust Fund provided by Australian Aid.
- South Asia window of the Trade Multi-Donor Trust Fund.

Why a spatial database?

- Most of the world’s urbanization in the coming two decades will take place in South Asia.
- Rural-urban transformation will be one of the main drivers of development; the opportunities it offers need to be fully tapped.
- Remote sensing, geographic information systems (GIS) and crowd-sourcing are resulting in a proliferation of geo-referenced data.

Why a new database?

- Remarkable geospatial portals exist already, e.g. Climate Change Knowledge Portal, Environmental Data Explorer, and WFPGeoNode.
- But these portals tend to have a strong sectoral focus: data on other themes is less granular.
- They often operate as a data catalog, offering links to other databases, but not integrating variables across levels of disaggregation.

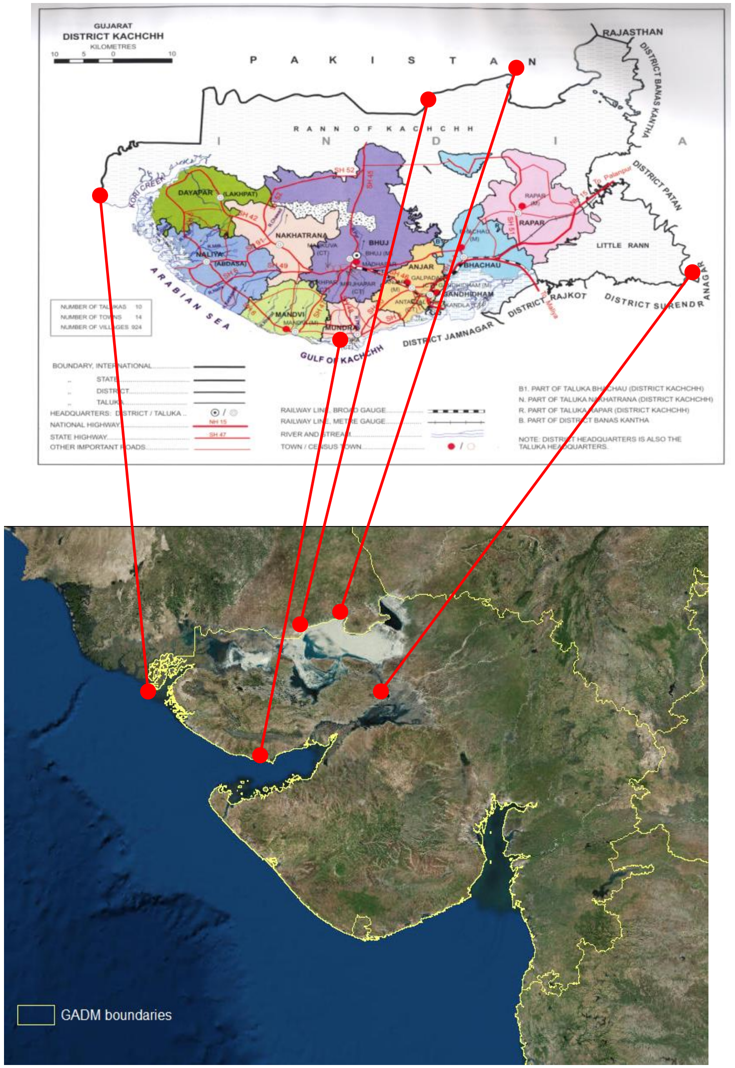
This spatial database

- Aims to integrate diverse socioeconomic and geographic variables from scattered sources into one single **platform** for the entire region.
- Uses a comparable spatial **hierarchy** across countries, including four administrative levels (from state/province to town/village) plus gridded cells/tiles.
- Ensures consistency between spatial **boundaries** and spatial features from paper maps and from remote sensing.
- Combines data **sources**, including traditional (e.g. administrative records, censuses, and surveys) and modern (e.g. remote sensing and crowd-sourced data).
- Pays special attention to the organization of all data under a unified **framework**, across sectors and spatial levels of disaggregation.
- Whenever possible **curates** indicators out of primary data to ensure their consistency across countries, years and sources.
- Presents the data for every indicator at the **lowest** possible level of disaggregation in the spatial hierarchy.
- Allows for consistent **aggregation** of indicators from lower to higher levels in the spatial hierarchy.
- Caters to multiple **users**, from those interested in visualization and descriptive tables to those keen to conduct econometric work.

Next section ->

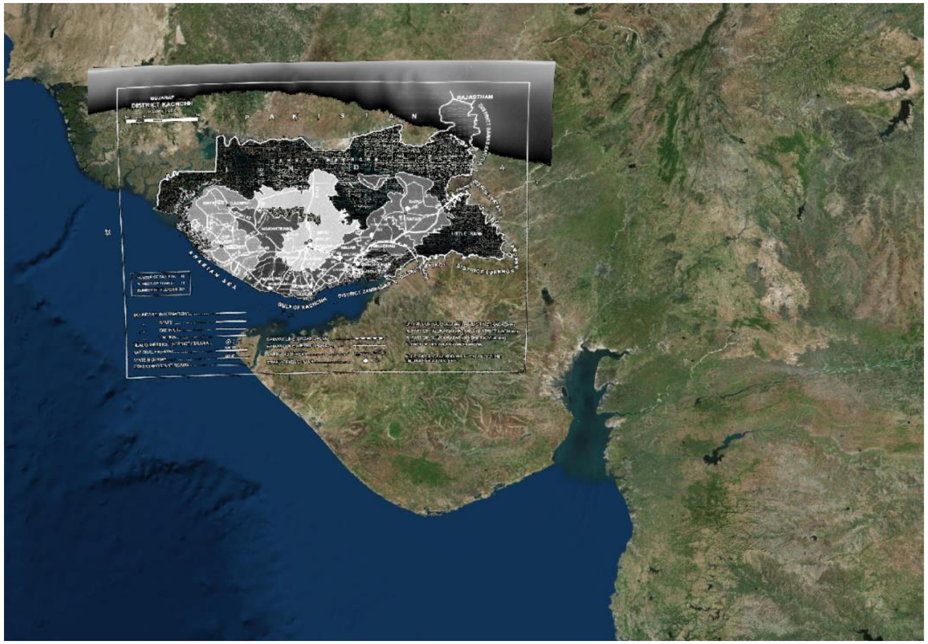
Boundaries: paper maps are digitized

Original paper map



Ground controls points

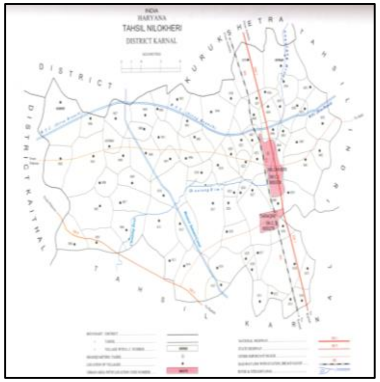
Geo-referenced map



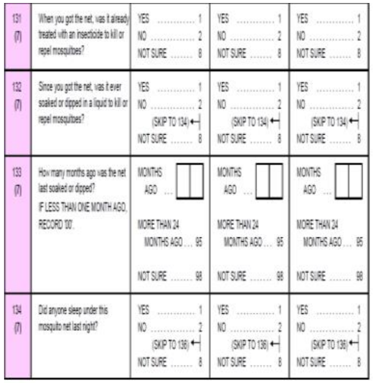
<-back

Sources: traditional and modern data are curated

Censuses



Surveys

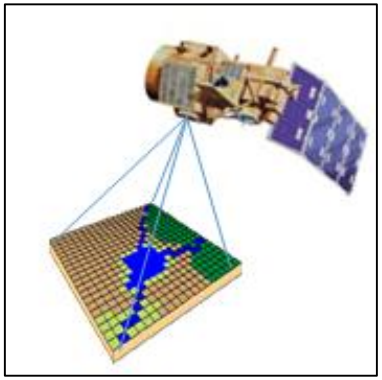


131 (7)	When you got the net, was it already treated with an insecticide to kill or repel mosquitoes?	YES .... 1 NO .... 2 NOT SURE 8	YES .... 1 NO .... 2 NOT SURE 8	YES .... 1 NO .... 2 NOT SURE 8
132 (7)	Since you got the net, was it ever soaked or dipped in a liquid to kill or repel mosquitoes?	YES .... 1 NO .... 2 (SKIP TO 134) ← NOT SURE 8	YES .... 1 NO .... 2 (SKIP TO 134) ← NOT SURE 8	YES .... 1 NO .... 2 (SKIP TO 134) ← NOT SURE 8
133 (7)	How many months ago was the net last soaked or dipped? IF LESS THAN ONE MONTH AGO, RECORD '00'.	MONTHS AGO ... □□ MORE THAN 24 MONTHS AGO ... 95 NOT SURE 98	MONTHS AGO ... □□ MORE THAN 24 MONTHS AGO ... 95 NOT SURE 98	MONTHS AGO ... □□ MORE THAN 24 MONTHS AGO ... 95 NOT SURE 98
134 (7)	Did anyone sleep under this mosquito net last night?	YES .... 1 NO .... 2 (SKIP TO 136) ← NOT SURE 8	YES .... 1 NO .... 2 (SKIP TO 136) ← NOT SURE 8	YES .... 1 NO .... 2 (SKIP TO 136) ← NOT SURE 8

Administrative records

	States/ provinces	Districts	Sub-districts	Villages/ towns	Tiles
Theme 1	▨	▨	▨	▨	
Theme 2	▨	▨			
Theme 3	▨	▨	▨	▨	▨
…					

Remote sensing



Crowd-sourced

<-back

Framework: unified across themes and levels

	Level 1 (States/provinces)	Level 2 (Districts)	Level 3 (Subdistricts)	Level 4 (Villages/towns)	Tiles
Urban extent	CN, RS, RT	CN, RS, RT	CN, RS, RT	CN, RS, RT	RS, RT
Demographics	CN, HS	CN, HS	CN	CN	
Jobs	CN, HS	CN, HS	CN	CN	
Economic activity	AR, EC, ES, NA, RS, RT	AR, EC, ES, NA, RS, RT	RS, RT	RS, RT	RS, RT
Infrastructure	CN, CS, EC, RS, RT	CN, CS, EC, RS, RT	CN, CS, RS, RT	CN, CS, RS, RT	CS, RS, RT
Information technology	AR, CN, ES	AR, CN, ES	CN	CN	
Finance	AR, CN, ES	AR, CN, ES	CN	CN	
Business	EC, ES	EC, ES			
Living standards	CN, HS	CN, HS	CN	CN	
Health	FS, HS	FS, HS			
Education	CN, FS, HS	CN, FS, HS	CN	CN	
Environment	RS, RT	RS, RT	RS, RT	RS, RT	RS, RT

Note: AR: administrative records; CN: census (population and housing census); EC: economic census; ES: establishment/firm surveys; FS: facility surveys: HS: household/labor force surveys; NA: national accounts; CS: crowdsourced data; RS: remote sensing data; RT: combining remote sensing and traditional data; GT: geo-referenced/geo-coded traditional data.

<-back

The spatial hierarchy

Spatial hierarchies differ across countries

- Bangladesh: highly centralized administratively and fiscally
  - *Constitution:* a unitary republic where much of the administrative power and fiscal authority reside with the central government.
  - *1972 reform:* legal basis for accrued responsibilities by sub-national governments (Articles 59 and 60).

- India: federal in nature with some salient unitary features
  - *Constitution:* a union of states with directly elected legislatures and appointed Governors and Chief Ministers in the executive role.
  - *1993 reform:* constitutional status to the panchayati raj system and municipalities (73rd and 74th amendments).
  - *Inter-fiscal transfers:* increased devolution of resources to the states (14th Finance Commission Award).

The meaning of the levels is county-specific

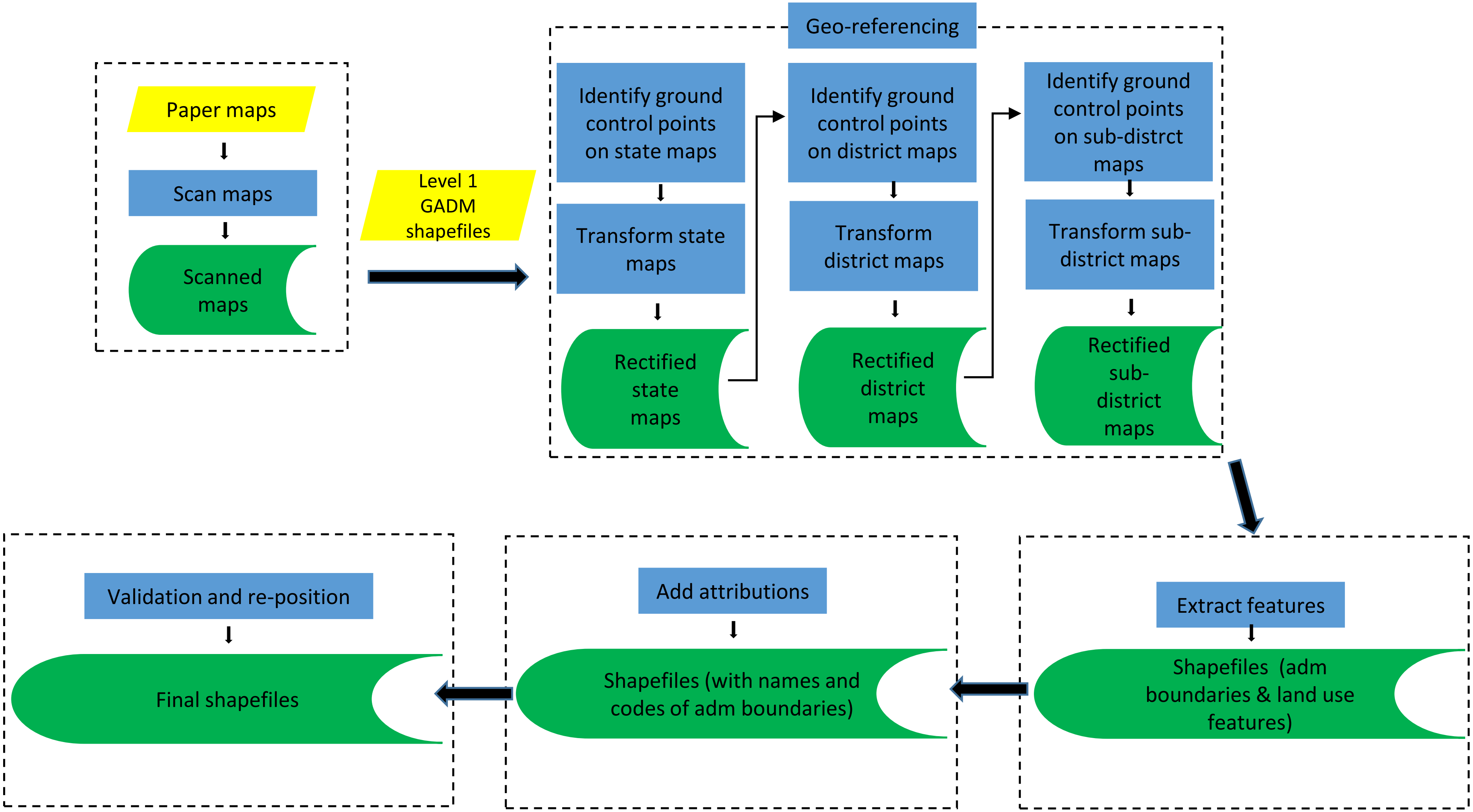
Spatial levels	Administrative levels			
	Bangladesh		India	
1	Divisions		States/Union Territories	
2	Districts (Zilas)		Districts	
3	Sub-districts (Upazilas)	Thanas	Sub-districts	
4	Unions	Wards	Villages	Towns
	Mouzas	Mohallas		Wards
	Villages			

Note: the highlighted levels are selected for the spatial hierarchy of the database.

Setting boundaries across the spatial hierarchy

- Data from different sources needs to be “mapped” to administrative units but that requires setting unequivocal boundaries.
- Boundaries are not always digitized by government agencies, especially at lower levels in the spatial hierarchy.
- Global Administrative Areas (GADM), an unofficial sources, provides digitized maps of administrative units for a number of countries.
- For India, GADM-digitized maps are outdated by a decade or so. Substantial administrative changes have taken place since then.
- Additionally, GADM-digitized maps are available only for states (level 1), districts (level 2), and sub-districts (level 3).
- To improve accuracy, for this spatial database we digitize official maps consistent with population censuses, which serve as anchors.

Key steps in the digitization of maps



Paper maps

Scan maps

Scanned maps

Level 1 GADM shapefiles

Geo-referencing

Identify ground control points on state maps

Transform state maps

Rectified state maps

Identify ground control points on district maps

Transform district maps

Rectified district maps

Identify ground control points on sub-distrct maps

Transform sub-district maps

Rectified sub-district maps

Extract features

Shapefiles (adm boundaries & land use features)

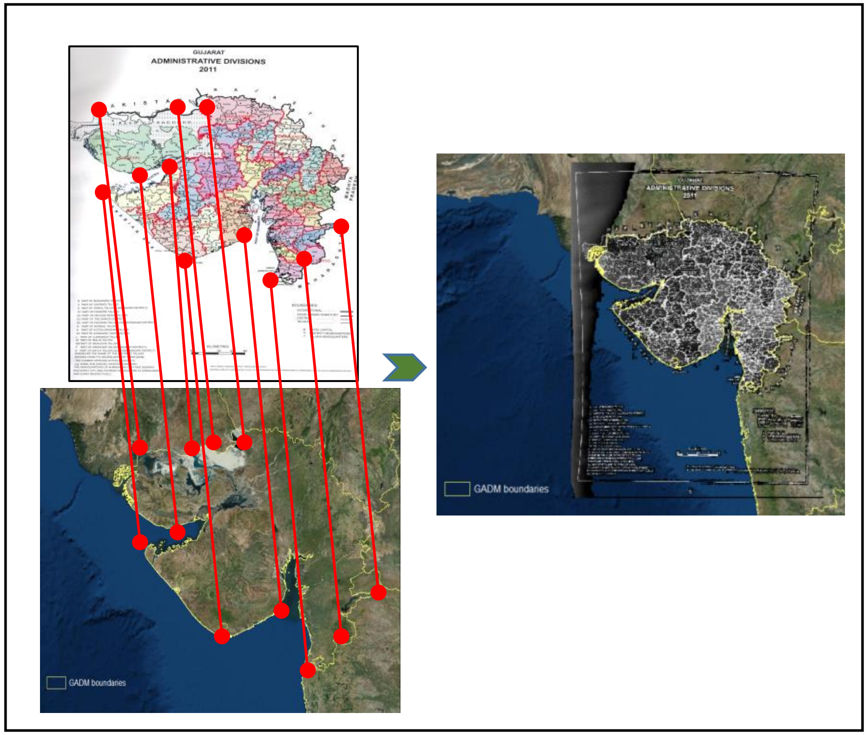
Add attributions

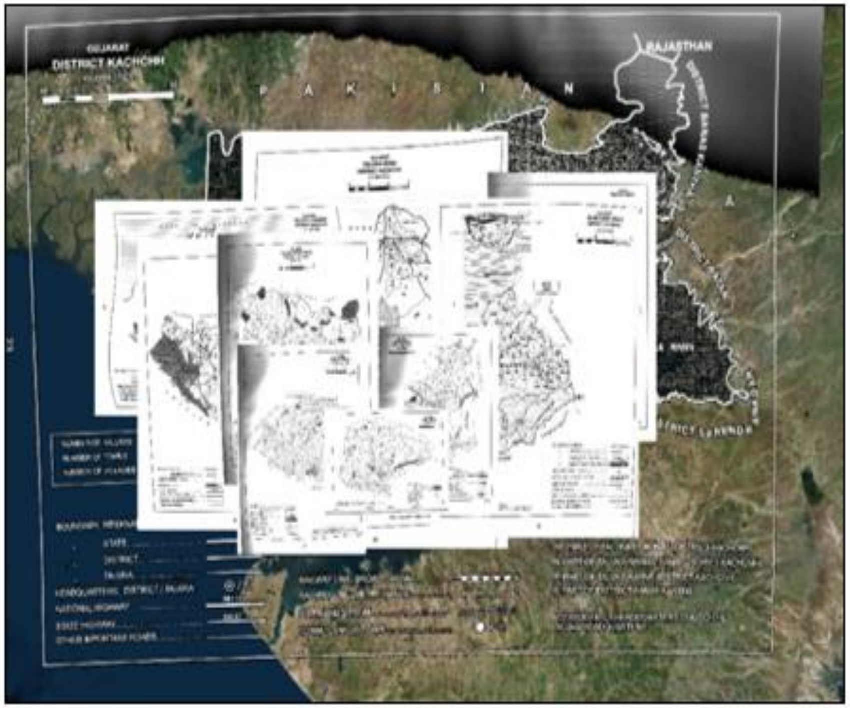
Shapefiles (with names and codes of adm boundaries)

Validation and re-position

Final shapefiles

A more intuitive version of the process

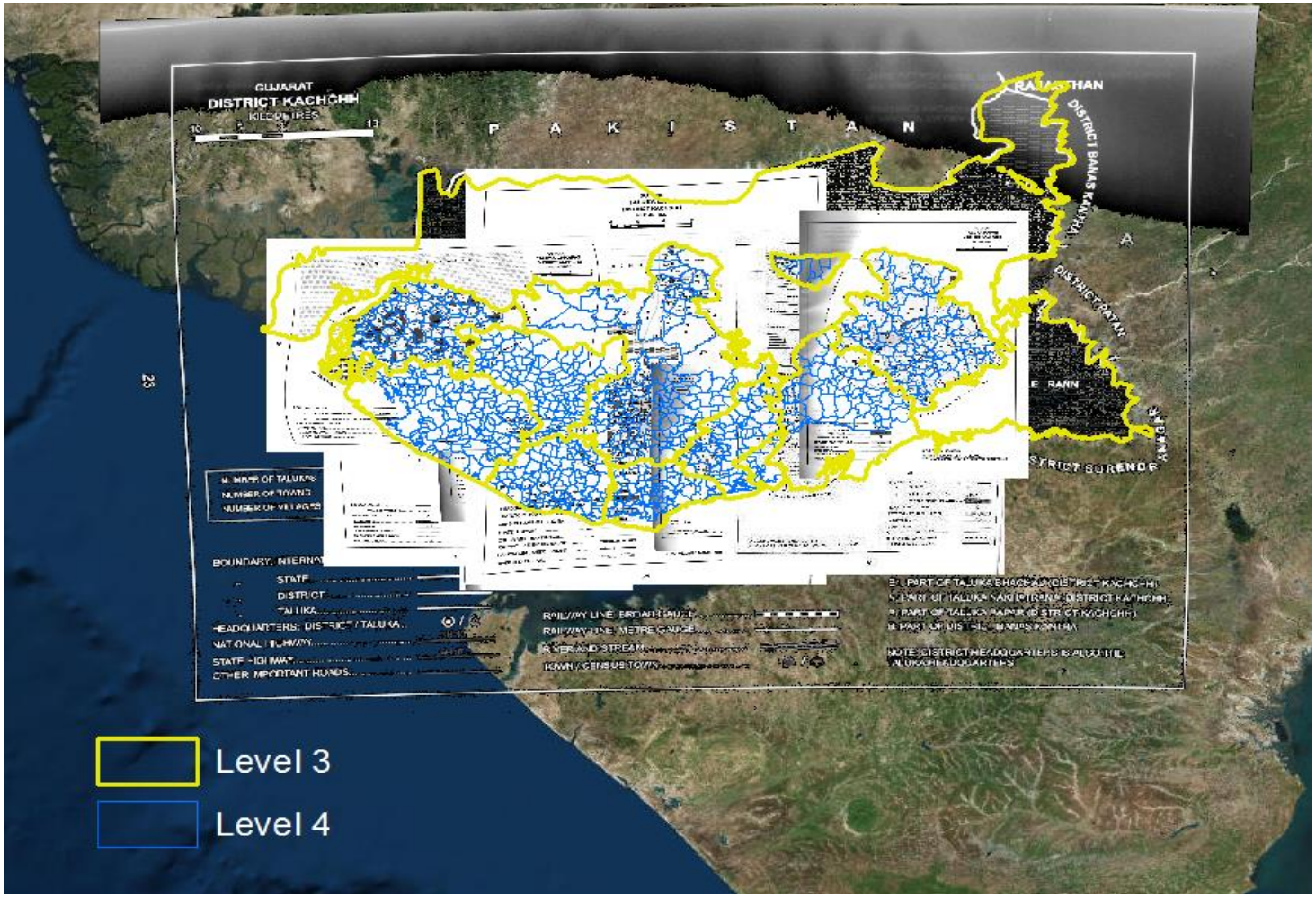




Geo-referencing scanned maps (based on features with well-know coordinates)

Assembling the puzzle

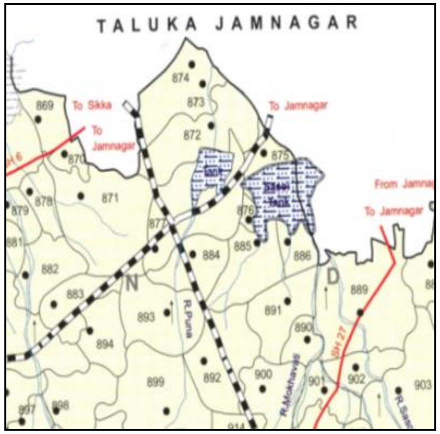
Extracting two key features from maps



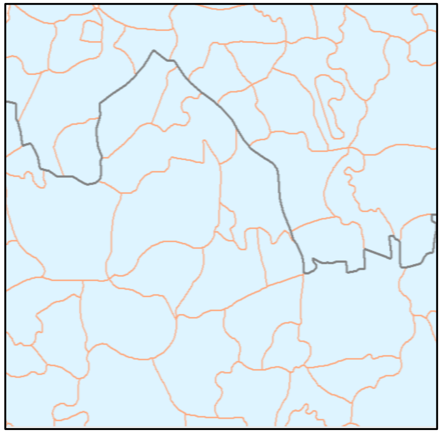
Administrative boundaries and physical markers

Two layers of data extracted in the process

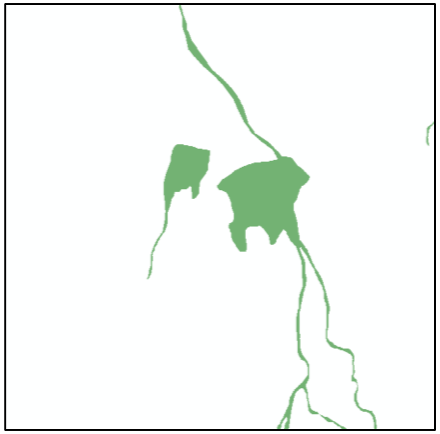
Level 3 source map



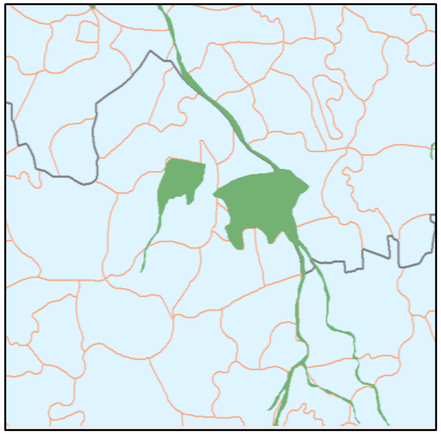
Administrative boundaries layer



Land use layer

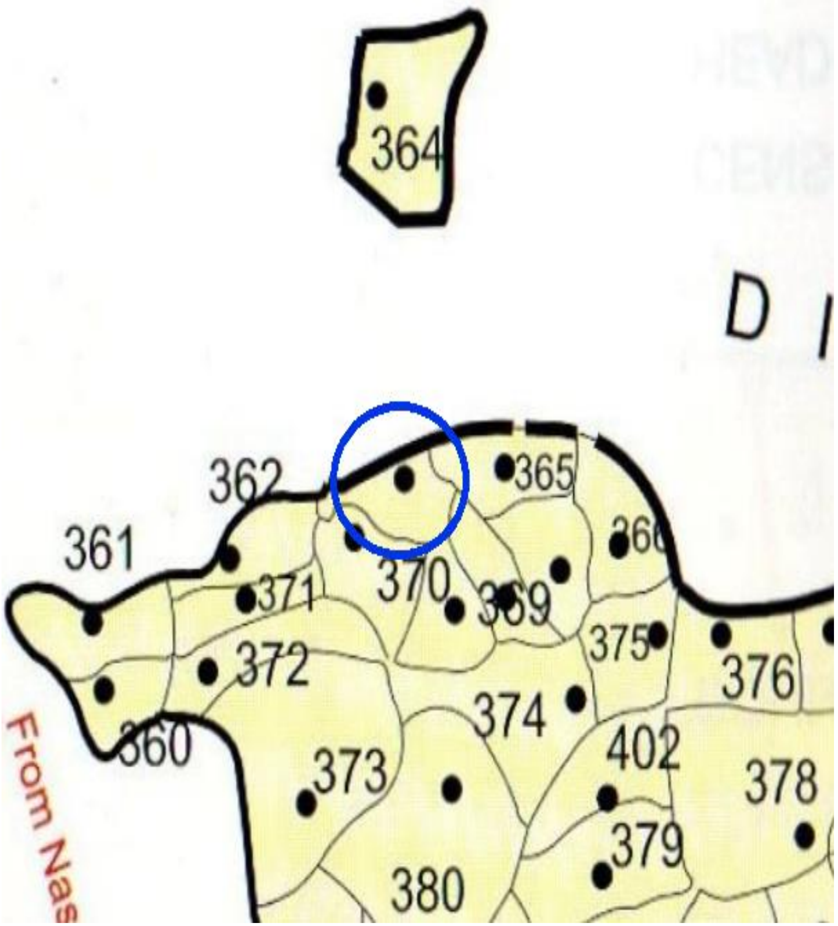


Two layers, superimposed



Adding attributions: how many Ashok Nagar?

- Attributions include:
  - Names
  - Location codes
  - Administrative status.
- Source maps are not perfect
  - Missing information
  - Mistakes
  - Outgrowths
- We rely on Meta Data and Data Standards (MDDS) adopted by the Census of India 2011 to correct these errors.

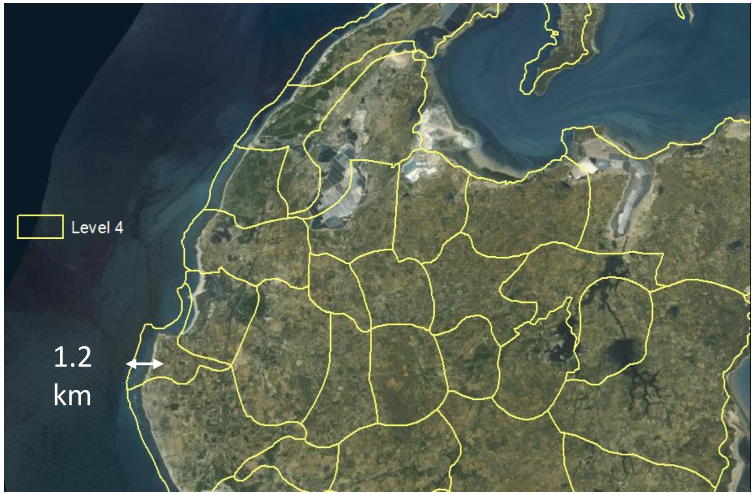


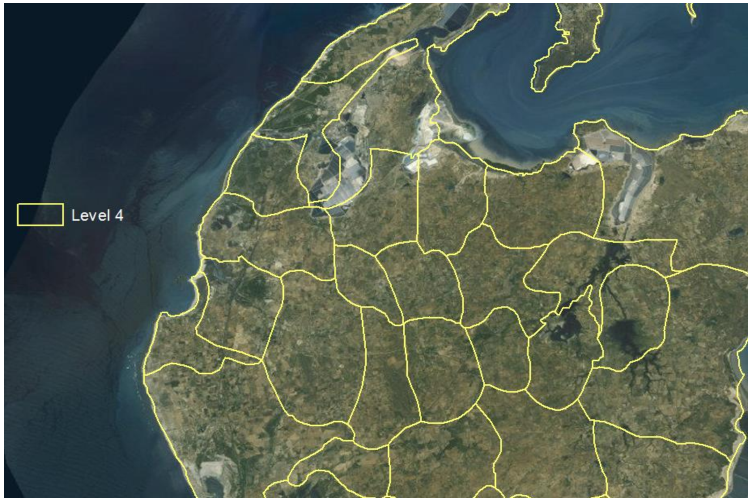
Validation and repositioning of shape files

Before

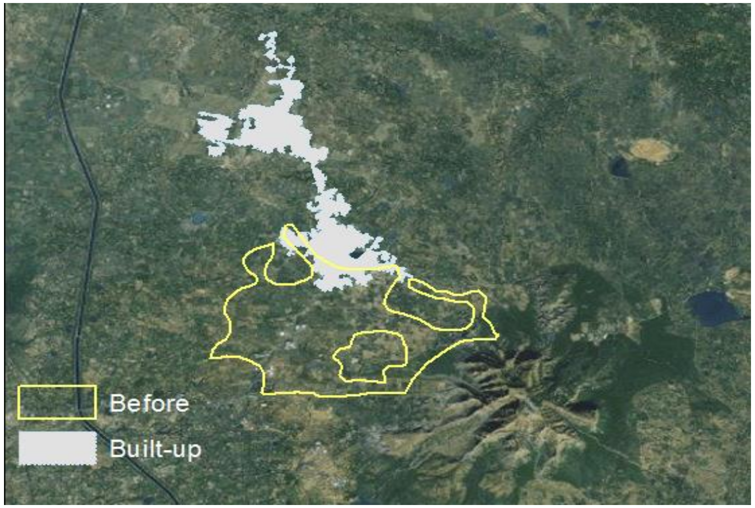
After

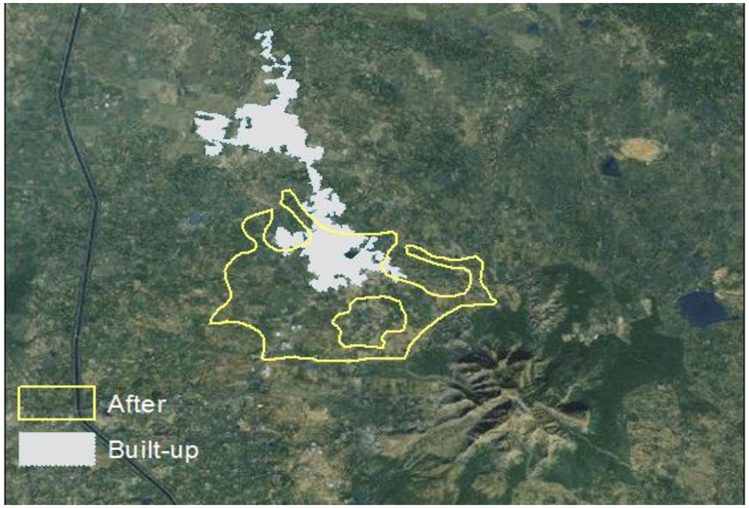
Coast lines



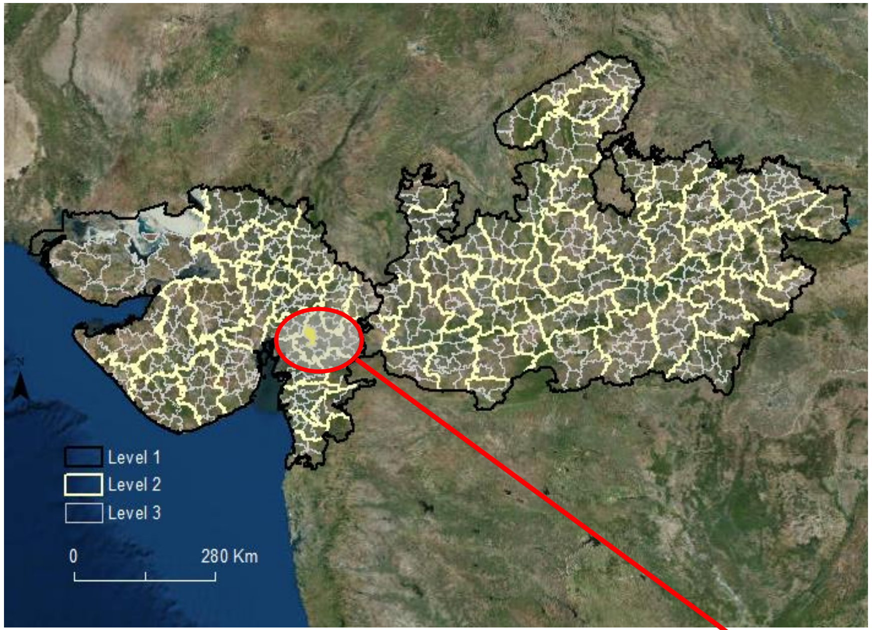


Built-up



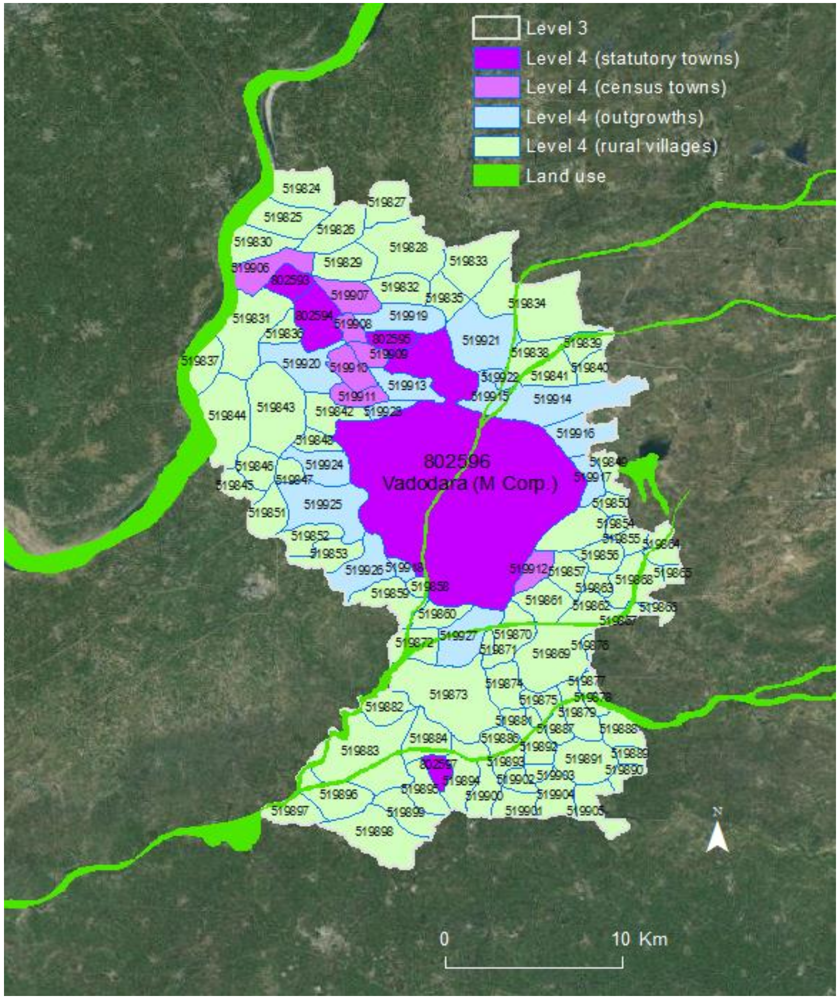


The final product



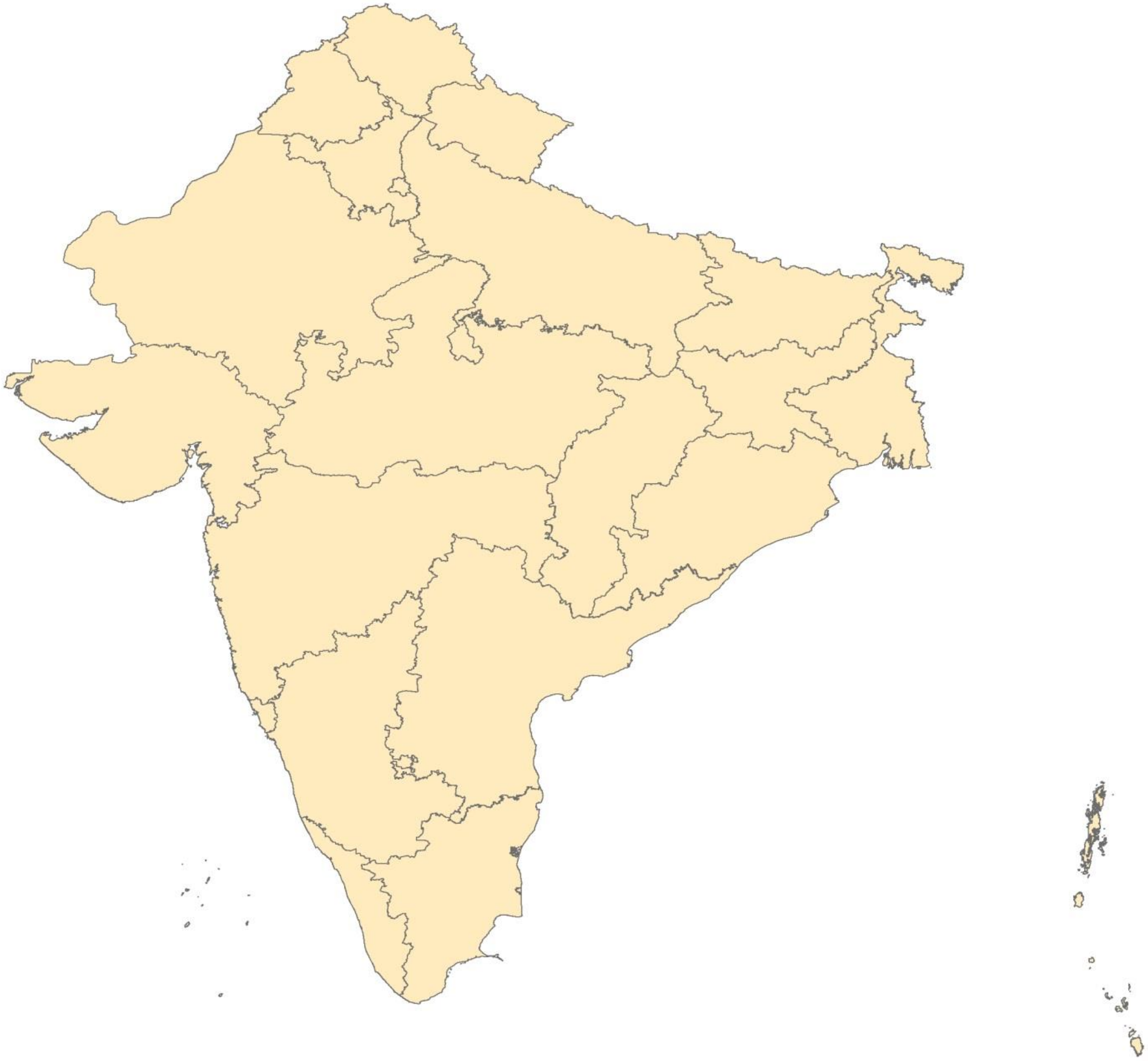
Gujarat and Madhya Pradesh

Gujarat: Vadodara (Level 3)



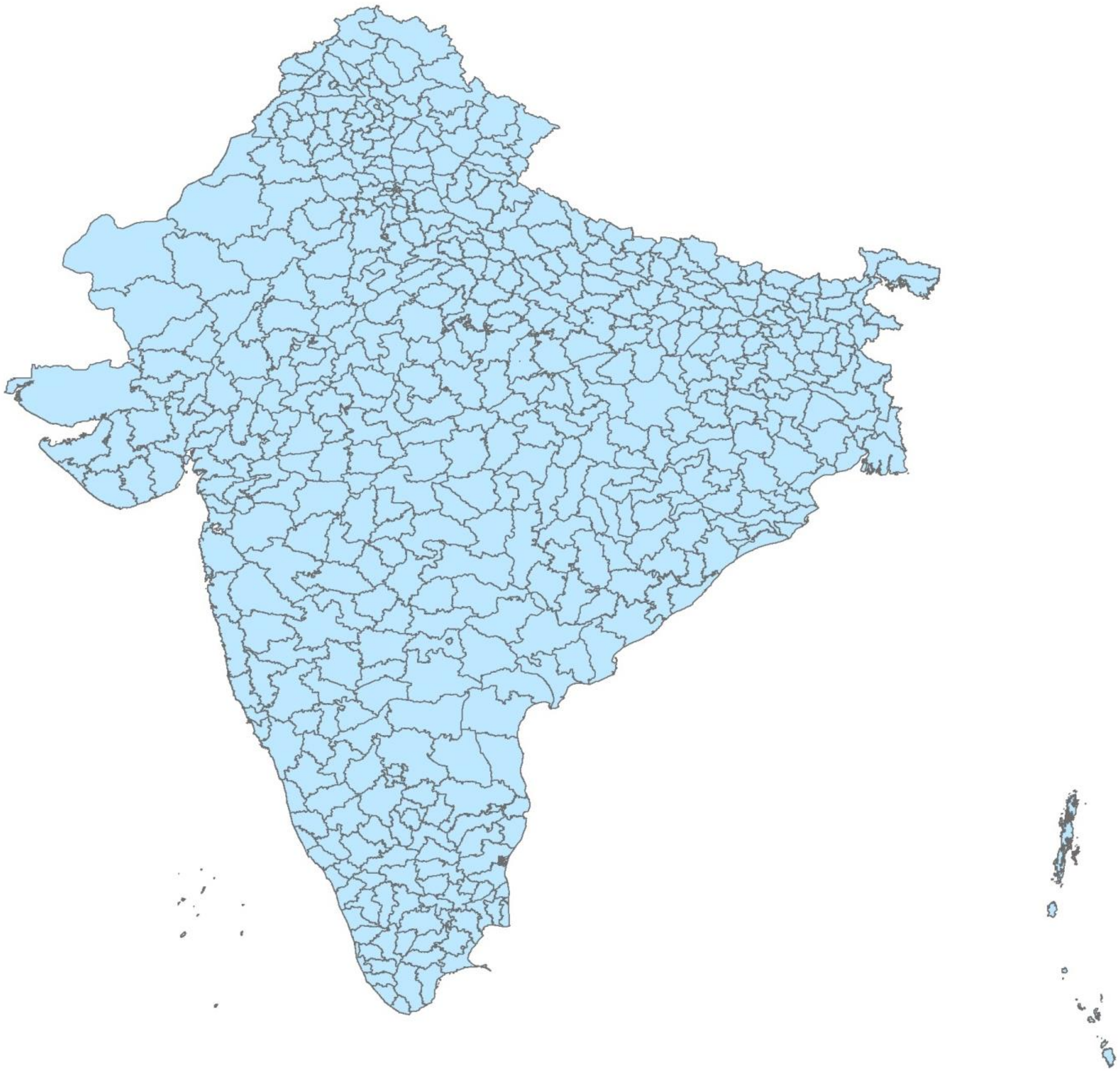
Shape files, as of today

Level 1



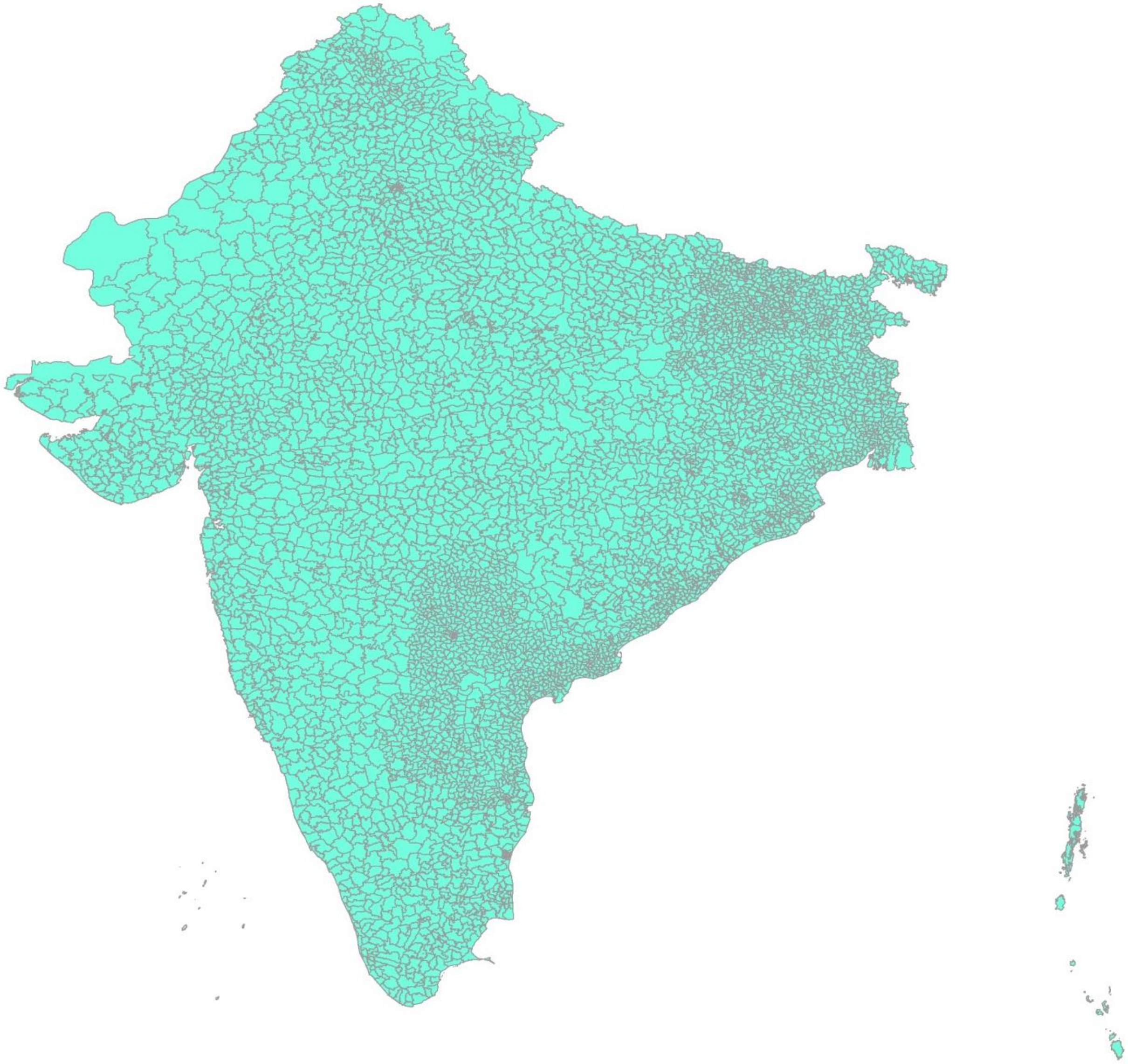
Shape files, as of today

Level 2



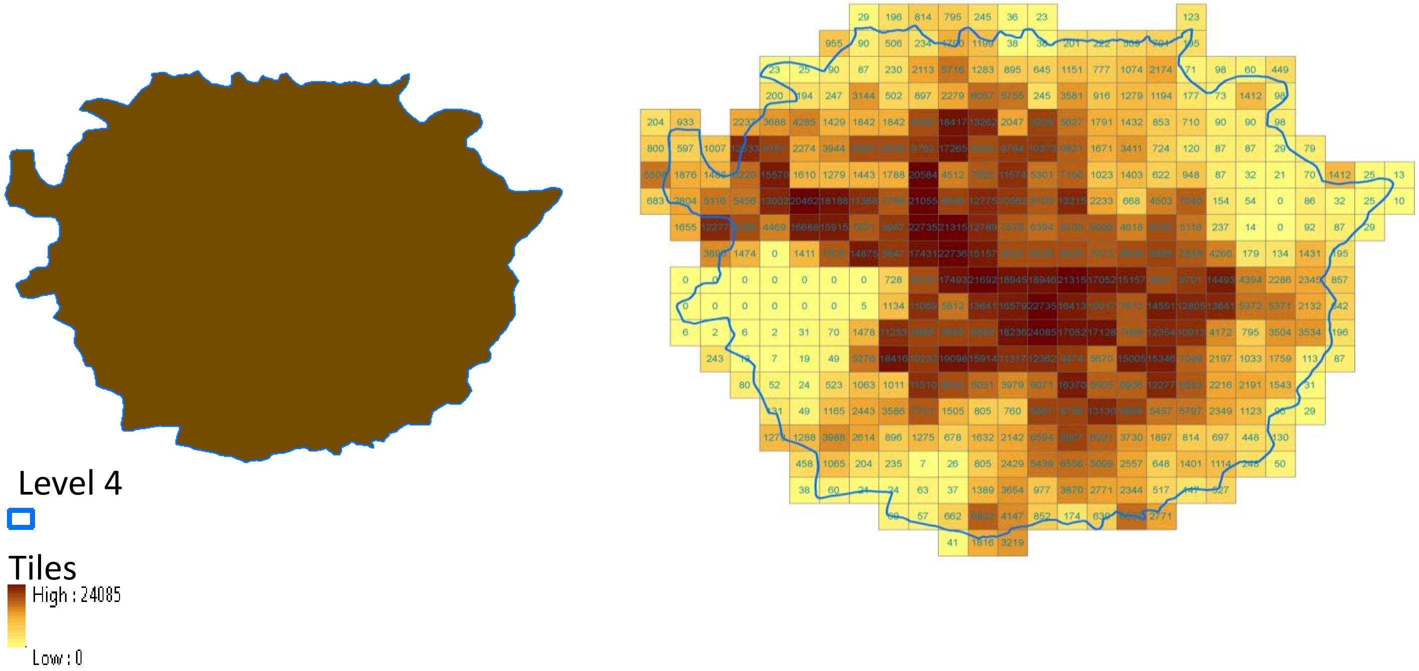
Shape files, as of today

Level 3



Tiles as the lowest level of disaggregation

- Remote sensing data is often available for areas (tiles or gridded cells) much smaller than the lowest administrative level:
  - LandScan™: population numbers at a resolution of approximately 1 km by 1 km at the equator.
  - MODIS Land Cover Type I product: five global land cover classification schemes at a resolution of approximately 500 meter.



Themes and indicators

Three main classifications of the data

- By type:
  - Traditional (administrative records, censuses, surveys).
  - Modern (remote sensing, crowd-sourcing)
  - Mixed (combining sources or geocoding traditional data).

- By provider:
  - For India, 30 data sources are tapped into.

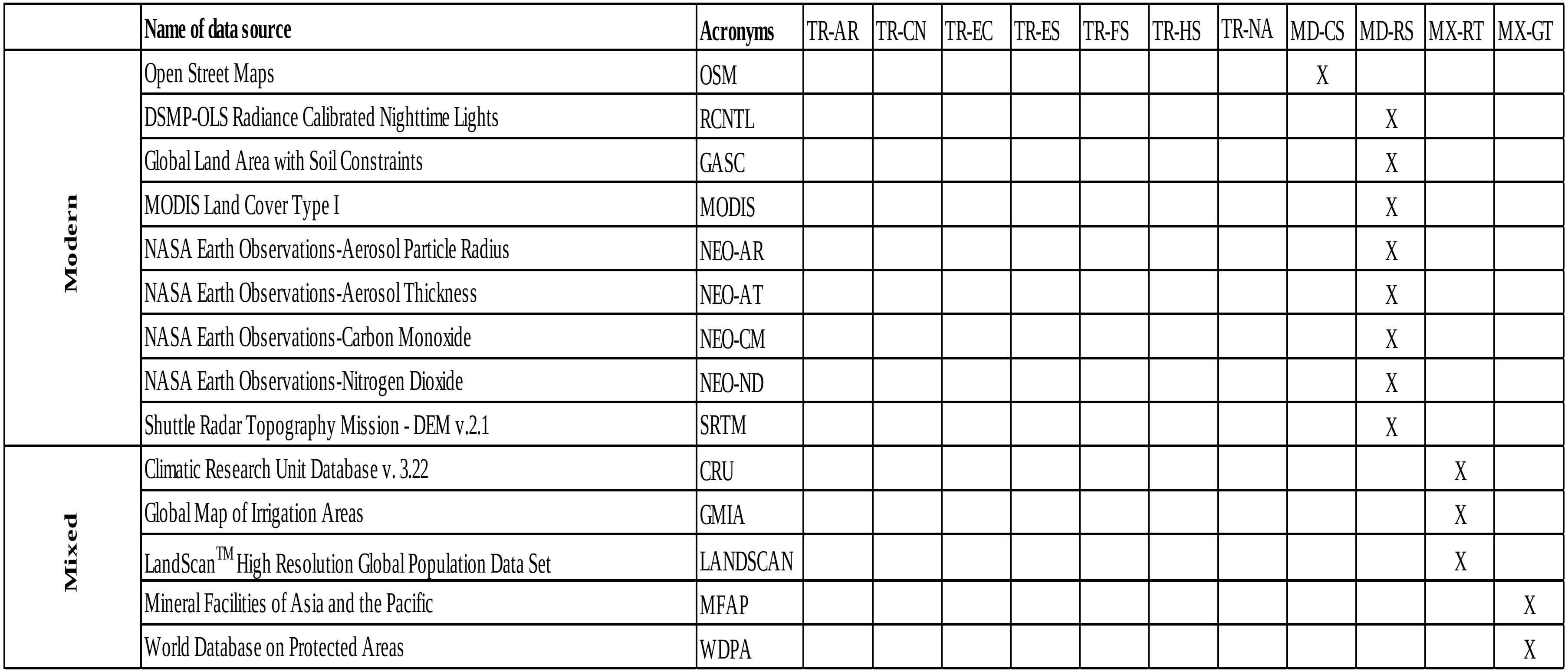
- By topic:
  - 12 major themes (e.g. education, infrastructure...)
  - Each disaggregated into sub-themes
  - Several indicators by sub-theme

The data used for India (1)

	Name of data source	Acronyms	TR-AR	TR-CN	TR-EC	TR-ES	TR-FS	TR-HS	TR-NA	MD-CS	MD-RS	MX-RT	MX-GT
Traditional	Agricultural Prices from India	API	X										
	District Crop Production Statistics	DCPS	X										
	Farm Harvest Prices of Principal Crops in India	FHP	X										
	Population and Housing Census_Houselisting and Households Census	PHC-HHC		X									
	Population and Housing Census_Primary Census Abstract	PHC-PCA		X									
	Population and Housing Census_Population Enumeration Data	PHC-PED		X									
	Economic Census	EC			X								
	Annaul Survey of Industries	ASI				X							
	National Sample Survey _Enterprises	NSS-ENT				X							
	District Information System for Education	DISE					X						
	District Level Household Survey	DLHS					X	X					
	Annual Health Survey	AHS						X					
	Annual Status of Education Report	ASER						X					
	National Sample Survey _Household Consumption Expenditure	NSS-HCE						X					
	National Sample Survey _Employment and Unemployment	NSS-EUE						X					
	State Wise District Domestic Product	DDP							X				

Note: TR-AR: administrative records; TR-CN: census (population and housing census); TR-EC: economic census; TR-ES: establishment/firm surveys; TR-FS: facility surveys: TR-HS: household/labor force surveys; TR-NA: national accounts; MD-CS: crowdsourced data; MD-RS: remote sensing data; MX-RT: combining remote sensing and traditional data; MX-GT: geo-referenced/geo-coded traditional data

The data used for India (2)



	Name of data source	Acronyms	TR-AR	TR-CN	TR-EC	TR-ES	TR-FS	TR-HS	TR-NA	MD-CS	MD-RS	MX-RT	MX-GT
Modern	Open Street Maps	OSM								X			
	DSMP-OLS Radiance Calibrated Nighttime Lights	RCNTL									X		
	Global Land Area with Soil Constraints	GASC									X		
	MODIS Land Cover Type I	MODIS									X		
	NASA Earth Observations-Aerosol Particle Radius	NEO-AR									X		
	NASA Earth Observations-Aerosol Thickness	NEO-AT									X		
	NASA Earth Observations-Carbon Monoxide	NEO-CM									X		
	NASA Earth Observations-Nitrogen Dioxide	NEO-ND									X		
	Shuttle Radar Topography Mission - DEM v.2.1	SRTM									X		
Mixed	Climatic Research Unit Database v. 3.22	CRU										X	
	Global Map of Irrigation Areas	GMIA										X	
	LandScan^{TM} High Resolution Global Population Data Set	LANDSCAN										X	
	Mineral Facilities of Asia and the Pacific	MFAP											X
	World Database on Protected Areas	WDPA											X

Note: TR-AR: administrative records; TR-CN: census (population and housing census); TR-EC: economic census; TR-ES: establishment/firm surveys; TR-FS: facility surveys: TR-HS: household/labor force surveys; TR-NA: national accounts; MD-CS: crowdsourced data; MD-RS: remote sensing data; MX-RT: combining remote sensing and traditional data; MX-GT: geo-referenced/geo-coded traditional data

Selecting data when multiple sources exist (1)

Data sources on land use

	Name	Originator	Resolution	Year	No. of Land Classes	Method	Accuracy
Selected	**MODIS**	NASA	500 m	2001–2012	17	Use a supervised decision-tree classification method.	75% globally; 93% urban land
	GlobCover	ESA	300 m	2004–06, 2009	22	Combine supervised and unsupervised algorithms (stratified clustering) with land cover class labeling based on experts' knowledge.	67% globally; 70% urban land
	Global Land Cover-SHARE	FAO	1000 m	2014	11	Synthesize existing global information sources, and incorporate the best available national and sub-national land cover information.	80% globally; 70% urban land

Source: Authors based on Bicheron et al. (2008), Friedl et al. (2015) and Latham et al. (2014).

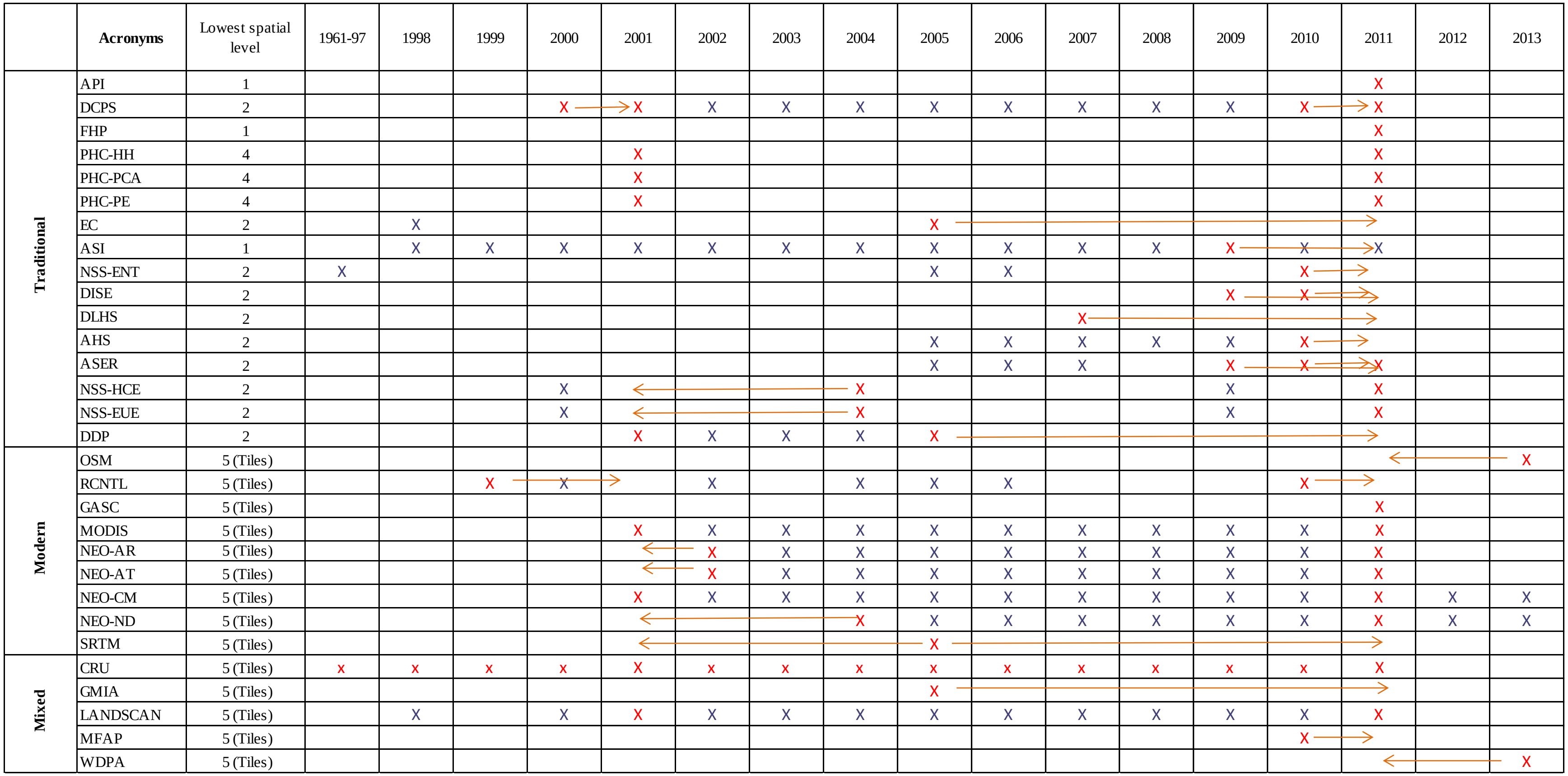
Selecting data when multiple sources exist (2)

Data sources on night lights

	Products	Resolution	Year	Screen out cloud cover, lightning, moonlit	Stable lights	Correct saturation at bright cores	Inter-satellite calibrated	Inter-annual calibrated
	DMSP-OLS Nighttime Lights	1000 m	1992–2013	Yes	Yes	No	No	No
Selected	**DMSP-OLS Radiance Calibrated Nighttime Lights**	1000 m	1996, 1999, 2000, 2002, 2004, 2006, 2010	Yes	Yes	Yes	Yes	No
	VIIRS Nighttime Lights	500 m	2014 (10 months), 2015 (5 months)	Yes	No	Yes	Yes	No

Source: authors based on Hsu et al. (2015) and Wu et al. (2013).

Matching source years data with database years

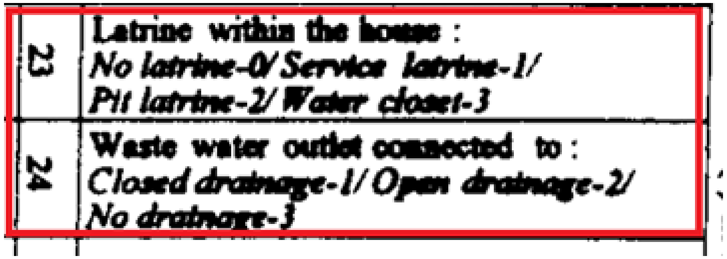


	Acronyms	Lowest spatial level	1961-97	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
Traditional	API	1															X		
	DCPS	2				X →	X	X	X	X	X	X	X	X	X	X →	X		
	FHP	1															X		
	PHC-HH	4					X										X		
	PHC-PCA	4					X										X		
	PHC-PE	4					X										X		
	EC	2		X							X →						→		
	ASI	1		X	X	X	X	X	X	X	X	X	X	X	X →	X	→ X		
	NSS-ENT	2	X								X	X				X →	→		
	DISE	2													X →	X →	→		
	DLHS	2											X →				→		
	AHS	2									X	X	X	X	X	X →	→		
	ASER	2									X	X	X		X →	X →	→ X		
	NSS-HCE	2				X	←			← X					X		X		
	NSS-EUE	2				X	←			← X					X		X		
	DDP	2					X	X	X	X	X →						→		
Modern	OSM	5 (Tiles)															←		← X
	RCNTL	5 (Tiles)			X →	X	→	X		X	X	X				X →	→		
	GASC	5 (Tiles)															X		
	MODIS	5 (Tiles)					X	X	X	X	X	X	X	X	X	X	X		
	NEO-AR	5 (Tiles)					←	X	X	X	X	X	X	X	X	X	X		
	NEO-AT	5 (Tiles)					←	X	X	X	X	X	X	X	X	X	X		
	NEO-CM	5 (Tiles)					X	X	X	X	X	X	X	X	X	X	X	X	X
	NEO-ND	5 (Tiles)					←			← X	X	X	X	X	X	X	X	X	X
	SRTM	5 (Tiles)					←				← X →						→		
Mixed	CRU	5 (Tiles)	x	x	x	x	X	x	x	x	x	x	x	x	x	x	X		
	GMIA	5 (Tiles)									X →						→		
	LANDSCAN	5 (Tiles)		X		X	X	X	X	X	X	X	X	X	X	X	X		
	MFAP	5 (Tiles)														X →	→		
	WDPA	5 (Tiles)															←		← X

Note: Red color indicates what is used in the spatial database.

Harmonizing indicators over time

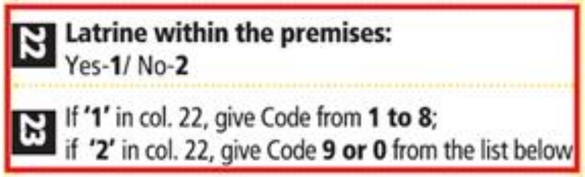
Census of India 2001:



23 Latrine within the house :
No latrine-0/ Service latrine-1/
Pit latrine-2/ Water closet-3

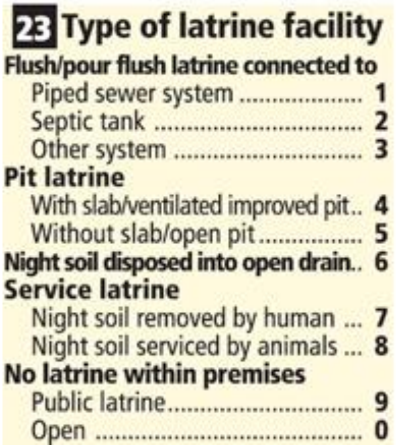
24 Waste water outlet connected to :
Closed drainage-1/ Open drainage-2/
No drainage-3

Census of India 2011:



22 **Latrine within the premises:**
Yes-**1**/ No-**2**

23 If **'1'** in col. 22, give Code from **1 to 8**;
if **'2'** in col. 22, give Code **9 or 0** from the list below



23 Type of latrine facility

Flush/pour flush latrine connected to	
Piped sewer system	1
Septic tank	2
Other system	3
Pit latrine	
With slab/ventilated improved pit..	4
Without slab/open pit	5
Night soil disposed into open drain..	6
Service latrine	
Night soil removed by human	7
Night soil serviced by animals	8
No latrine within premises	
Public latrine................................	9
Open	0

Example:
improved
sanitation

Solution: two indicators.

- Enhanced improved sanitation, consistent with WHO/UNICEF standard
- Improved sanitation, less strict than WHO/UNICEF standard; all types of pit latrine are considered as improved

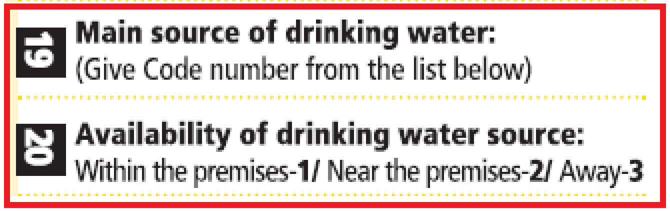
Harmonizing indicators across countries

Census 2011, Bangladesh

Source of drinking water – 2011 long form
[] Tap
[] Tube-well / Deep tube-well
[] Well
[] Pond
[] River/ Ditch / Canal
[] Other

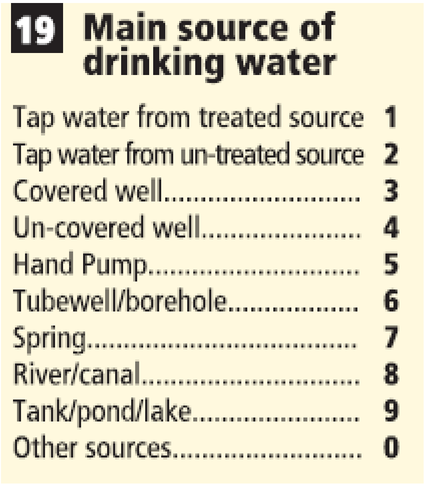
Example: improved water

Census of India 2011:



19 **Main source of drinking water:**
(Give Code number from the list below)

20 **Availability of drinking water source:**
Within the premises-**1**/ Near the premises-**2**/ Away-**3**



19 Main source of drinking water

Tap water from treated source	1
Tap water from un-treated source	2
Covered well..........................	3
Un-covered well......................	4
Hand Pump............................	5
Tubewell/borehole.................	6
Spring..................................	7
River/canal............................	8
Tank/pond/lake......................	9
Other sources.........................	0

Solution:

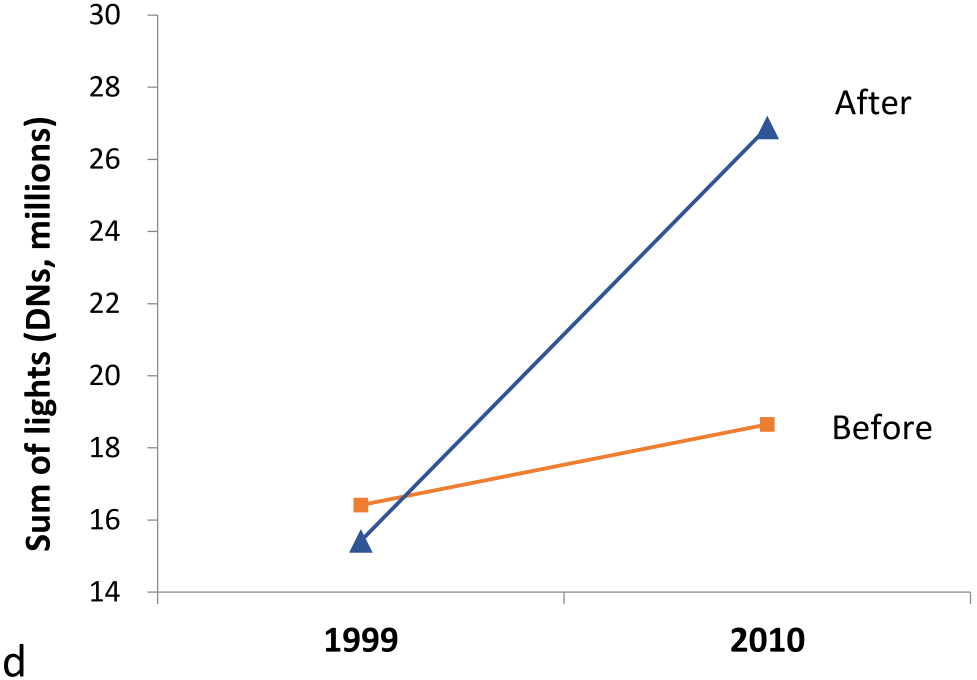
Both well and spring are considered unimproved water, stricter than WHO/UNICEF standard.

Harmonizing indicators over time

Example:
night
lights

Satellite sensors get gradually worn out and capture less light over time.

Solution:
Apply an inter-annual calibration adjustment using parameters computed by Hsu et al. (2015) on the DMSP-OLS Radiance Calibrated Nighttime Lights.



Source: authors, based on DMSP-OLS Radiance Calibrated Nighttime Lights.

Note: the figure shows the sum of light intensity for 1999 and 2010.

Geo-referencing

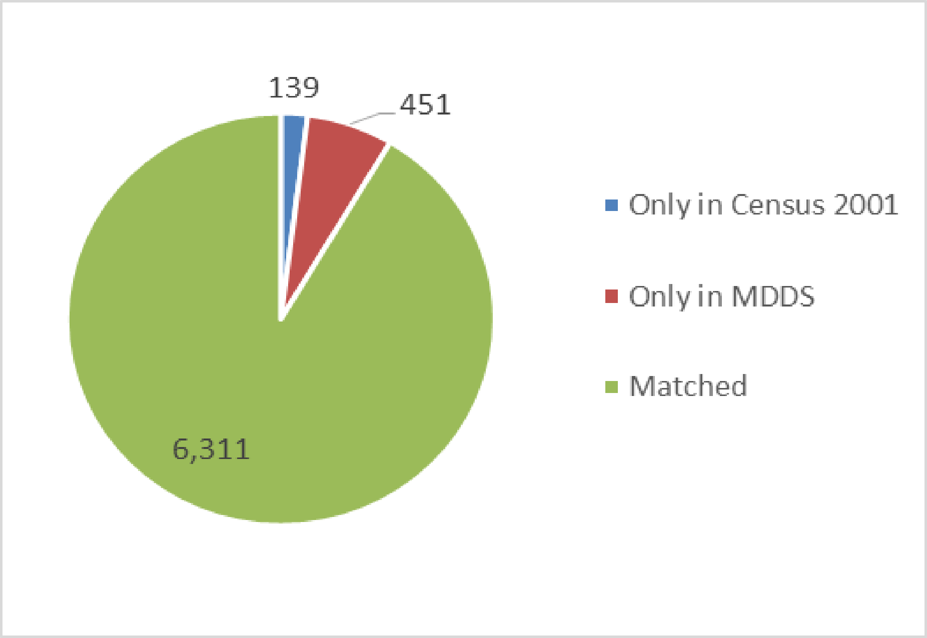
Administrative units change over time

- In principle, data from traditional sources can be “mapped” to shape files based on location codes (e.g. district-level code).
- In practice, administrative structures experience substantial changes over a decade, so the mapping is not unequivocal.
- Example: India between 2001 and 2011
    - the number of districts increased from 593 to 640,
    - the number of sub-districts from 5,463 to 5,924,
    - the number of towns from 5,161 to 7,935, and
    - the number of villages from about 639,000 to more than 649,000.
- Some administrative units were carved out of previously existing units, others were created by merging parts of existing units.

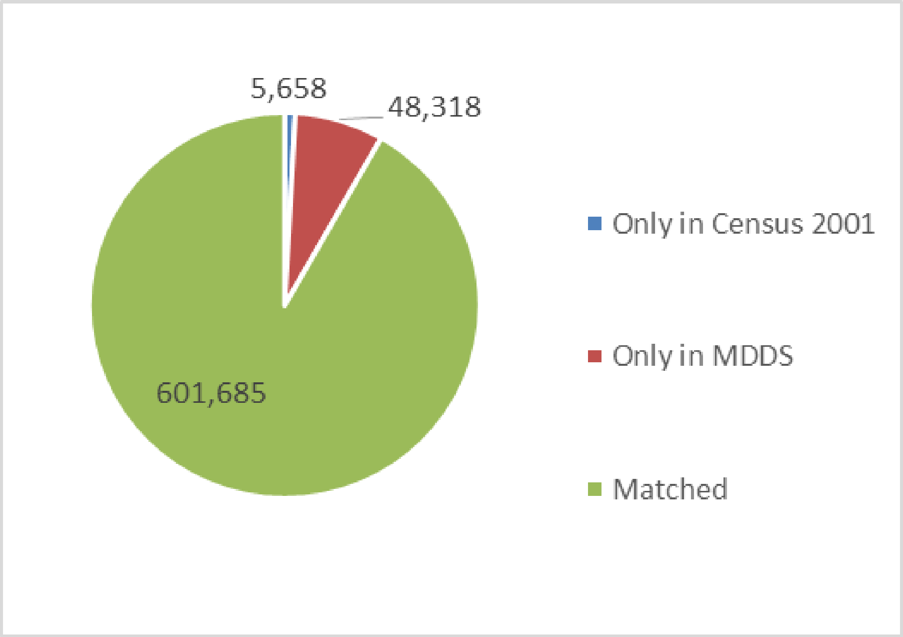
Location codes can differ across sources

The official concordances presented by the Meta Data and Data Standards (MDDS) adopted by the Census of India do not match location codes of 2001 Census perfectly.

Sub-districts

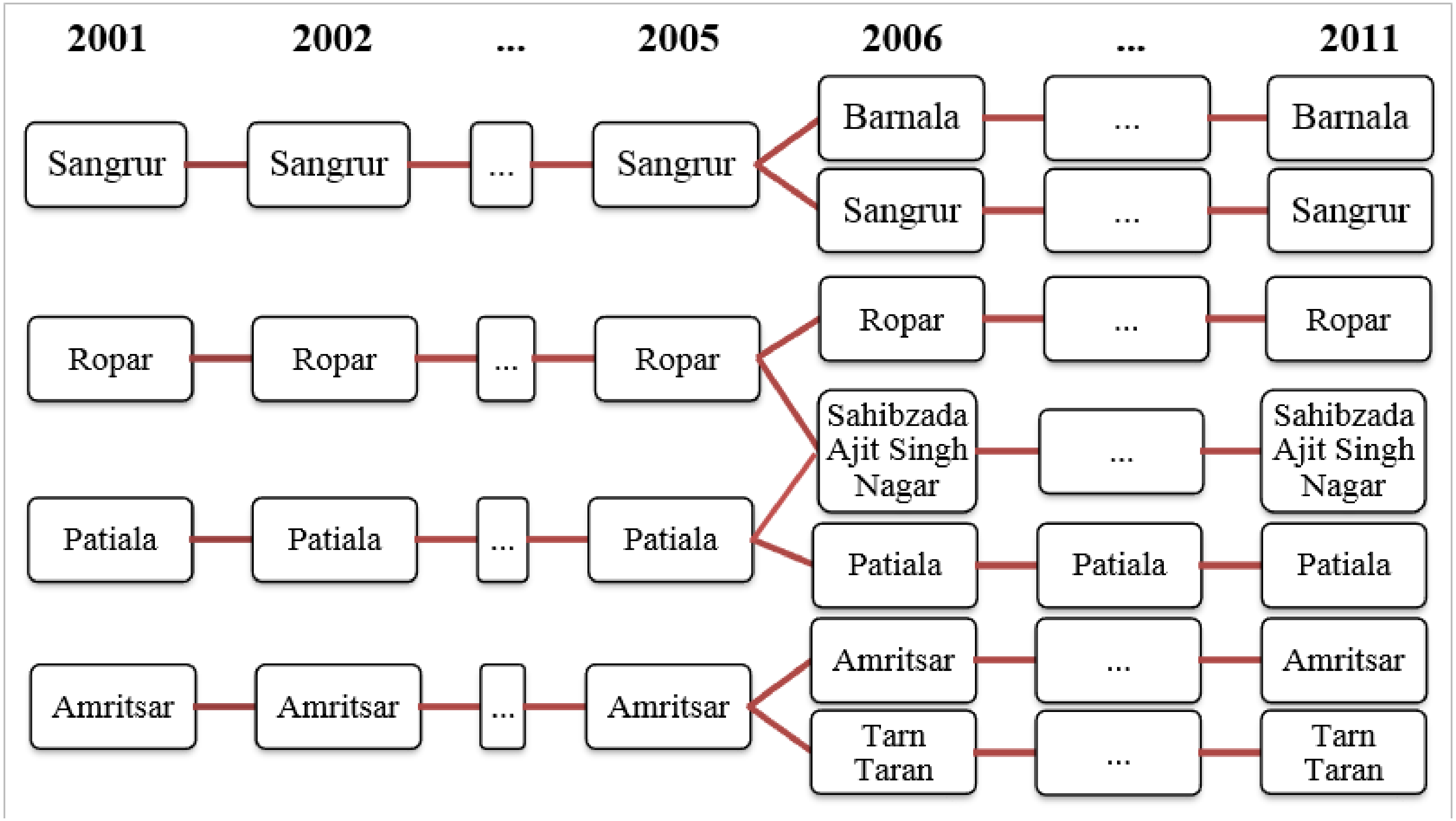


Towns/villages



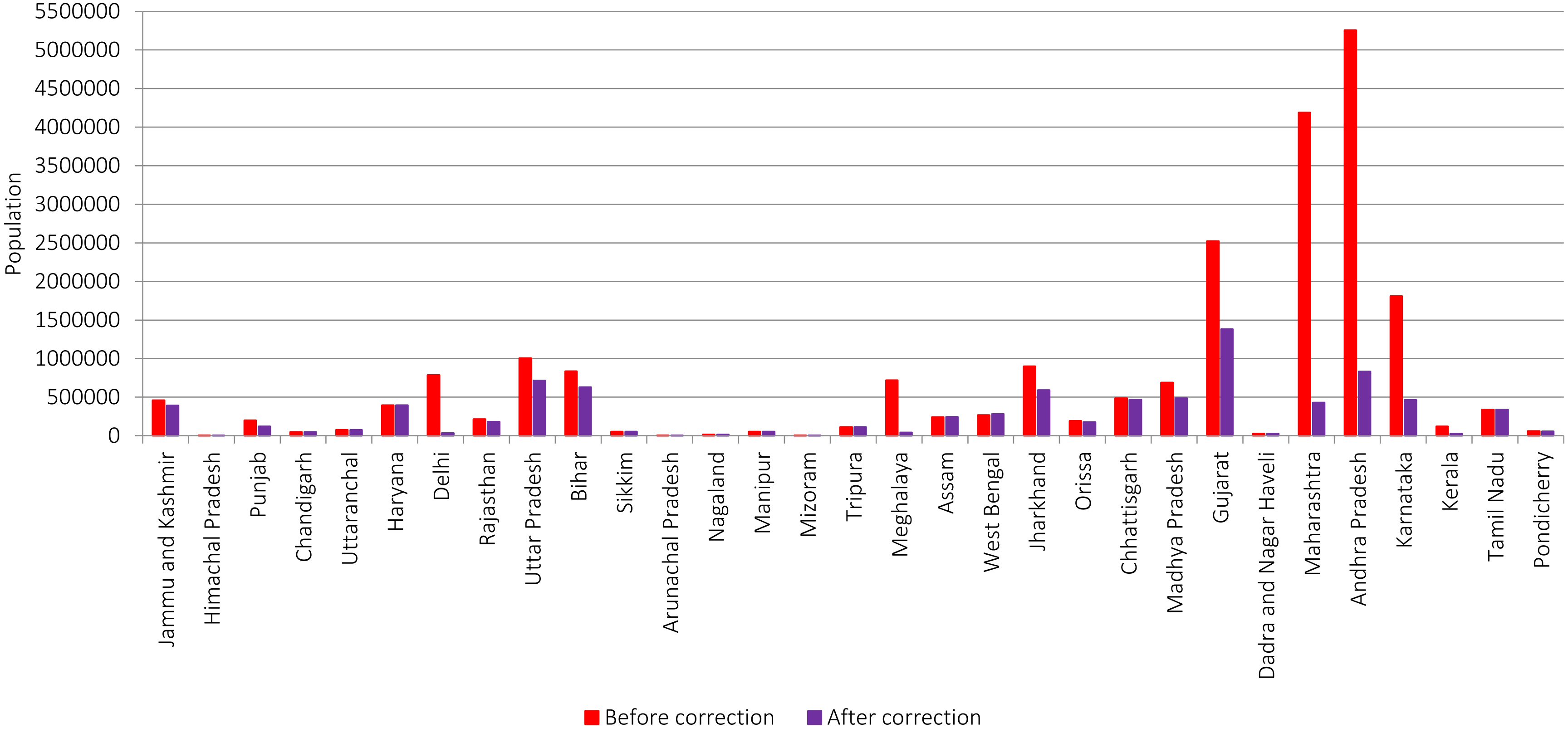
A detailed administrative timeline had to be built

Example: districts in Punjab (India)



Corrections made for units with large populations

Mismatches were reduced from 22.1 million people (2.1 percent of the population) to 8.6 million (0.8 percent).



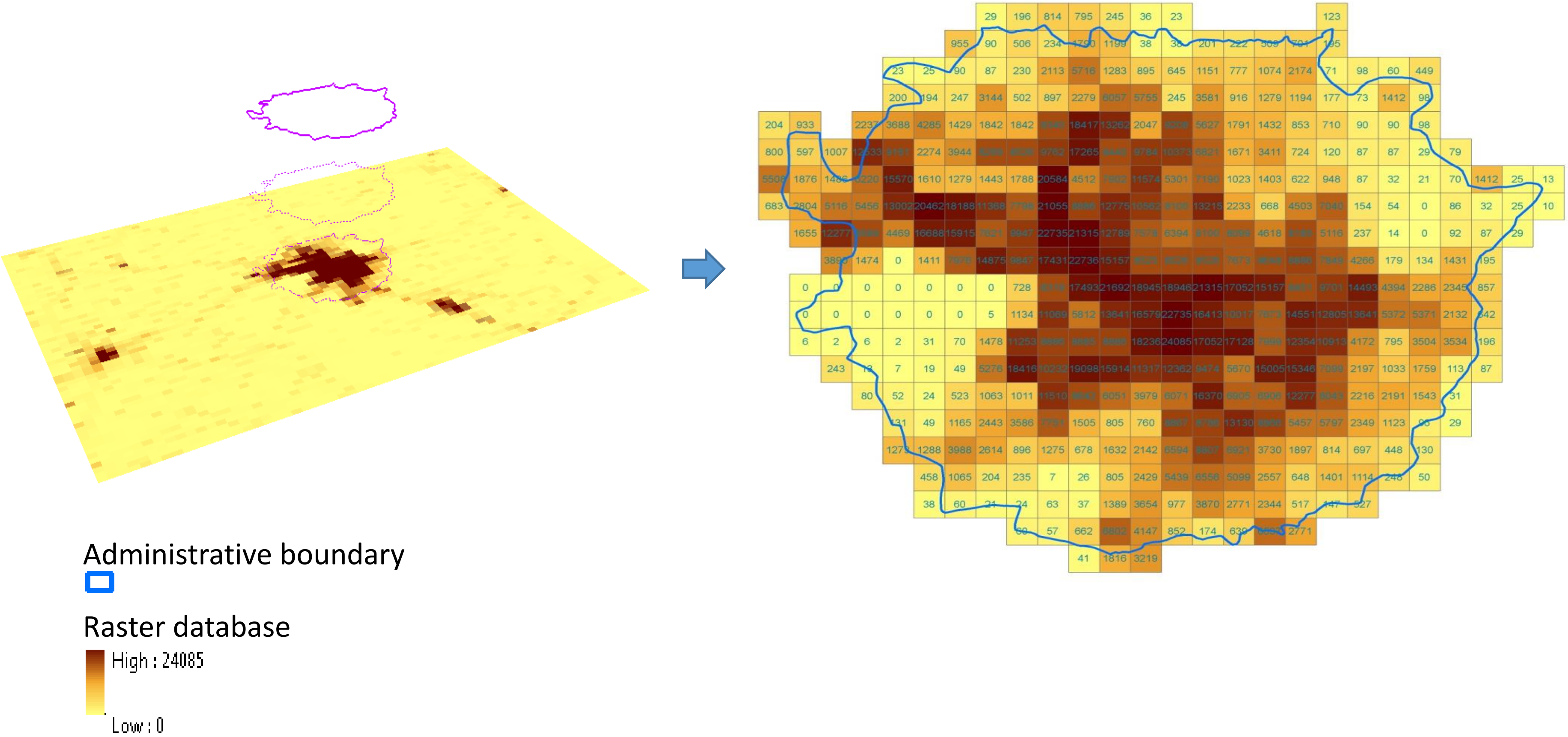
Note: States or Union Territories without mismatches are excluded from the figure. They include Andaman & Nicobar Islands, Daman & Diu, Goa, and Lakshadweep.

Geo-referencing modern and mixed sources

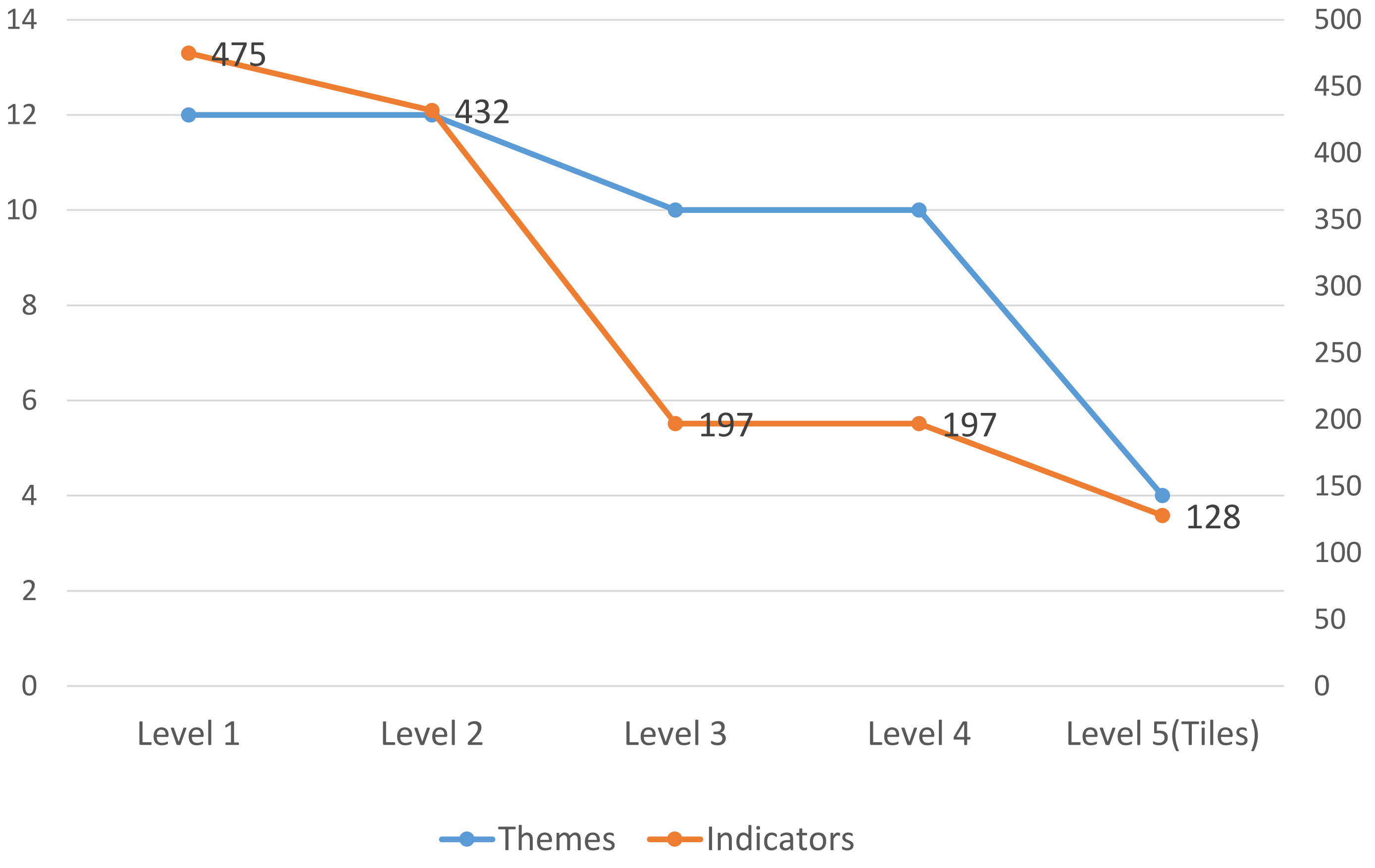
- The most popular system to project data to coordinates is the one used by Google Earth.
- In this system, D_WGS_1984 is used as the datum and GCS_WGS_1984 as the Geographic Coordination System for the shape files.
- There are cases where the projection system of the original data differs from this, as in the MODIS product.
- In these cases, a transformation of the original projection system is performed for our database.

Intersecting the original data and the shape files

The tiles (and/or fractions of tiles) that fall into each of the intersected areas are what belong to the specific administrative units at the corresponding spatial level.



The end-result: indicators across spatial levels



Explore

A work in progress

- The database is exclusively for internal World Bank use. We hope to make it available to the general public in 2016.
- The India component of the database is the most complete; work is under way for Bangladesh, and Nepal comes next.
- The portal can be accessed at:

  www.worldbank.org/spatialdatabase-southasia
- The portal is protected.

  Username: worldbank-southasia

  Password: ja4uKasp

- Data is still being uploaded to the portal and there may still be technical glitches. Please report them so that we can fix them.
- The portal is currently optimized for Google Chrome. It will be expanded to other web browsers gradually.