A Spatial Database for South Asia

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

I. Introduction

A spatially granular lens can improve our understanding of development and better anchor public policy and investment design. In the past, the limited availability of spatial data has proved to be a constraint, particularly with respect to developing countries. Reliable and frequent data collection for nationally representative statistics has been a challenge for many developing countries, not to mention sub-national or local level statistics. Recently, with the growth of remote sensing data and efforts to integrate geographic information system (GIS) technics, this constraint has been relaxed. For example, applying GIS technology to census population and remote sensing–based geographic data, LandScan™ High Resolution Global Population Data Set provides population count data at the level of 1 km-by-1 km–gridded cells for all countries on an annual basis. Similarly, spatial data on a set of topics, including agriculture, climate, geography, land use, and population distribution, are proliferating rapidly.

Against this backdrop, various geospatial portals have been set up to help curate the growing number of spatial data for the development community. Efforts at the global level include Climate Change Knowledge Portal of the World Bank, Environmental Data Explorer of the United Nations Environmental Programme, and WFPGeoNode of the World Food Programme, among others. These portals allow different types of mapping experiences, and they have greatly improved the public's access to geospatial data. Partly because of their mandates, however, these portals have focused on specific topics of interest. Even when their coverage goes beyond their topics of interest, indicators of the other topics (e.g., living standards) are of much less spatial granularity and generally are available at the national level.

Most of these portals function as data catalogs. National- or sub-national–level statistics are listed together with indicators at the gridded-cell level. The former often are based on more traditional forms of data, and the latter generally are derived from remote sensing data. Efforts at effectively integrating these different data sources are lacking. Subsequently, a simple framework for harmonizing indicators of multiple topics and of different levels of spatial disaggregation often is absent. And although access to these spatial data is available to the public, users without GIS knowledge face a deep learning curve to use them.

* The authors are with the office of the Chief Economist for South Asia, at the World Bank. The findings, interpretations, and conclusions expressed in this paper are entirely theirs and do not necessarily represent the views of the World Bank. The corresponding author is Yue Li, at yli7@worldbank.org. The authors gratefully acknowledge the research assistance provided by Mihir Prakash and Jiarui Wang, the contribution by private partners RMSI Private Ltd. and Matrix-Geo Solutions Pvt. Ltd., and the support received from Barjor Mehta, Rinku Murgai, and other colleagues of the India country team. The project benefited from close collaboration with India Poverty and Shared Prosperity Cluster Program, South Asia Regional Urbanization Flagship Report and Global Research Program on Spatial Development of Cities. It was funded by the World Bank, 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, and the Trade Multi-Donor Trust Fund.

Some regional geospatial portals are better at presenting a framework. Arab Spatial portal for the Middle East and North Africa region and HarvestChoice's Mappr application for the sub-Saharan Africa region are two examples. Despite these two portals' focus on agriculture and food security, indicators from other topics are curated and organized into a consistent framework across countries. However, even in these portals, indicators available at the gridded-cell level are not available at the sub-national levels defined by their spatial hierarchy. There remains a demarcation between statistics at different administrative levels and indicators at the gridded-cell level.

The Spatial Database for South Asia is a repository of spatially disaggregated indicators for all countries of South Asia. Sharing the mandate of other geospatial portals, the Spatial Database for South Asia aims to bring more granularity into data for development. This is accomplished by integrating diverse socioeconomic and geographic variables from scattered sources into one focal point for the region. Particular emphasis is put on establishing a consistent organizational framework, so that experts in geographic information systems (GIS) as well as researchers, policy makers, and the general public—who have no previous GIS knowledge—can easily make use of spatial data.

To that end, this spatial database relies on a relatively comparable spatial hierarchy across countries, including four administrative levels plus gridded cells or tiles. The four levels correspond to states or provinces, districts, sub-districts, and towns or villages. Paper maps of selected administrative levels are digitized to ensure map consistency across countries in terms of granularity, vintage, and quality. Both traditional sources of data (e.g., administrative records, census data, and surveys) and more modern forms of data (e.g., crowd sourced and remote sensing data) are curated. Regardless of the source, all indicators are organized around a single framework, covering a dozen themes. These themes range from environment, infrastructure, and urban extent (the topics of interest for most geospatial portals) to jobs, economic activity, business, and living standards (the topics receiving much less or little coverage). Whenever possible indicators are built out of primary data to ensure their consistency across countries, years and sources.

Most important, indicators are georeferenced along the spatial hierarchy described above. For every indicator, an attempt was made to represent the data at the most disaggregated level possible. Indicators available at a certain spatial level also are available at all higher levels (or more aggregated levels). For example, population density derived from LandScan™ High Resolution Global Population Data Set is available not only for the gridded-cell level but also for the selected administrative levels of the spatial hierarchy. In the case of India, all twelve themes have indicators available at the state and district level, ten have indicators available at all four administrative levels down to villages/towns, and four have indicators at all five spatial levels. In terms of indicators, 475 are available for states, 432 for both states and districts, 197 for all four administrative levels, and 128 for all five spatial levels.

This working paper explains in detail each of the steps taken to construct the database. Accompanying documents explain the structure of the database, describe the data sources, and present metadata for all indicators. However, this spatial database is a work in progress. The organizational framework is far from perfect. Right now, the component for India is the most complete. As the country coverage expands over time, different aspects of the organizational framework are expected to improve. This paper will be updated regularly to reflect the changes.

II. The spatial hierarchy

The spatial hierarchy forms the backbone of this spatial database. It builds on the administrative or geographic levels used by the population and housing census of each country in the region and brings them together into a relatively comparable spatial hierarchy. In this section, we describe how paper maps of these selected administrative or geographic levels are digitized to set the basis for the visualization of indicators and spatial analysis. Finally, we discuss the lowest spatial level, "tiles," which preserve data in the form of gridded cells.

Select administrative levels

In terms of national and sub-national relationship, the government structures in South Asia are quite diversified and have been undergoing adjustments over time. For example, the government of Bangladesh is highly centralized, both administratively and fiscally. The Constitution of Bangladesh specifies the country as a unitary republic. Much of the administrative power and fiscal authority reside with the central government. Since 1972, the Constitution has incorporated the legal basis and responsibilities of sub-national governments in Articles 59 and 60 (The Constitution of the People's Republic of Bangladesh 2015). However, the structure of sub-national governments, especially the structure of rural local governments, has been undergoing periodic modifications largely following regime changes in the central government.

In particular, the British colonial rule and Pakistani period set the foundation for the post-independence sub-national administrative system: a four-tier hierarchy system existed, including (in descending order) divisional councils, district councils, thana councils (or municipal committees), and union councils. Since the independence of Bangladesh in 1971, the tiers, their nomenclature, the composition of their government bodies, and their powers and mandates have experienced rounds of changes. At present, immediately below the central government, divisions are placed under divisional commissioners. Each division is further subdivided into districts (or zilas) and lower administrative units. For rural areas, a three-tier system is in place (in descending order): zila parishads at the district level, upazila parishads at the sub-district level (or upazila/thana level), and union parishads at the union level. Below unions come mouzas and villages, but their administrative arrangements are not as well established. For urban areas, two types of governments exist: city corporations and municipalities (or paurashavas). City corporations often expand several districts, but municipalities generally belong to one district. Cities are further divided into thanas, wards, and mohallas, depending on their size. The administrative arrangements of these lower levels are not as well established (Mollah 2007, Panday 2011, and Rahman 2012).

As another example, the government of India is federal in nature with some salient unitary features. According to the Constitution of India, the country is a Union of States. All states have their own directly elected legislatures and appointed Governors and Chief Ministers in the executive role, with the exception of some Union Territories. In its Seventh Schedule, the Constitution delineates the powers and functions of the central and state governments, although the residuary powers are assigned to the central government. Fiscally, the constitutional assignment is inherently imbalanced to give the central government a role in sub-national redistribution. States undertake most public service provisions but are assigned relatively

limited tax collection powers. However, fiscal decentralization from the central to state governments has increased (The Constitution of India 2015, Khemani 2007, Kochanek 2007, Rao and Singh 2001).

Regarding local governments below the state level, the 73rd and 74th amendments to the Constitution have provided constitutional status to the panchayati raj system and municipalities since 1993. Under the panchayati raj system, three tiers of rural local government have been established; in descending order, they include district panchayats, sub-district (or block) panchayats, and village panchayats. Urban areas are governed by three types of administrative institutions: municipal corporations for a large urban area, municipal councils for a small urban area, and nagar panchayats for an area in transition from rural to urban. However, states have the discretionary authority to define the criteria of each type and notify statutory towns. Municipal corporations often expand over several sub-districts (and sometimes even several districts), while municipal councils and nagar panchayats generally belong to one sub-district. Larger cities are further divided into wards. Over time, each level of administrative unit (even the state level) and the units' geographic territories have changed substantially in number and area (Bhattacharyya 2005, The Constitution of India 2015, Kochanek and Hardgrave 2007, Rao and Singh 2001).

Other countries in the region also have distinct administrative hierarchy structures. Thus, perfectly matching administrative levels across countries, in terms of their political, administrative, and fiscal power and responsibilities, is a difficult undertaking.

We rely on relatively comparable spatial levels across countries from a statistical point of view using information from countries' population and housing censuses. Population and housing census data provide fundamental statistical information on demographics, human settlements, and economic and social issues. Censuses also play a critical role in all elements of the national statistical system. Among other functions, censuses are the principal source of records used as a sampling frame for sample surveys; the national statistical system of almost every country relies on these surveys for efficient and reliable data collection. Procedures for collecting vital statistics and migration statistics are also generally coordinated with census data collection procedures (in terms of coverage, concepts, and classification) to allow benchmarking and to improve complementarity. Information on economic characteristics of individuals and housing is collected through census, and that information is also used to prepare listings of establishments (or their proprietors) for censuses and surveys of economic establishments (United Nations 2008).

Given their essential role in statistics, we select the administrative or geographic levels available from countries' population and housing censuses to define the highest four spatial levels. For example, in the case of Bangladesh, the levels reported from the Population and Housing Census 2011 are divisions, districts, upazilas/thanas, unions/wards, mouzas/mohallas, and villages. Information for cities (either city corporations or municipalities) can be aggregated from thanas, wards, and mohallas (Bangladesh Bureau of Statistics 2011, 2014). For India, states/union territories, districts, sub-districts, villages/towns, and wards are the available administrative or geographic levels in the Census of India 2011 (Office of the Registrar General and Census Commissioner 2011a). Sub-districts are reported with different names by different states, such as tehsils, talukas, community development blocks, police stations, mandals, and revenue circles. The highest four levels of both censuses are selected to establish a relatively comparable spatial hierarchy (i.e., divisions, districts, upazilas/thanas, and unions/wards from Bangladesh and states/union territories, districts, sub-districts, and villages/towns from India) (table 1).

Some of these levels may not play an important role administratively or fiscally, and they may serve only as a geographic division. However, because they are adopted by the census, they can serve as a reasonably good spatial framework for us to incorporate other data sources, especially those of more traditional forms. For example, in Bangladesh, the Household Income and Expenditure Survey (among others data sources) develops its sampling frames from the census. In India, National Sample Surveys rely mainly on the census to derive their sampling frames; these surveys cover a wide range of topics, from household consumption expenditure to employment, and from community infrastructure to enterprises. Economic Census, which is a complete count of non-agriculture enterprises, also takes villages defined by the census as its primary sampling unit in rural areas.

Digitize paper maps

Determining which levels of administrative units to include is only the first step in establishing the spatial hierarchy. Next, we need to map these administrative units, including their locations, sizes, and shapes, and thus their geographic relationships to each other. Official maps of administrative units are generally available as paper maps. They are not easily integrated with socioeconomic indicators, nor are they easily used for spatial analysis. The general practice is to digitize these paper maps. During the process, maps are transformed from flat papers into digital vector storage formats that match the locations, sizes, and shapes of the administrative units under consideration on Earth. Only then can the indicators for each administrative unit be easily visualized on a map. More important, spatial analysis can be conducted for these indicators to address questions such as: Do two villages next to each other have similar poverty rates? Do villages near a metropolitan area have a higher consumption level than those far away? Is output of one sub-district affected by road networks in neighboring sub-districts?

Outside of countries' official agencies, most spatial databases use digitized maps acquired through open sources. One disadvantage of this practice is that the accuracy, vintage, and level of spatial disaggregation of the digitized administrative maps is constrained by the source used. For example, Global Administrative Areas (GADM) is one of the most widely used open sources (Global Administrative Areas 2011). It provides digitized maps of administrative units for a relatively large number of countries. However, for India, GADM-digitized maps are outdated by a decade or so. Their representation of India's administrative boundaries is close to what it was in reality in 2005. Substantial changes have taken place in the country at the district, sub-district, and village/town level since then. Additionally, GADM-digitized maps are available only for states (level 1), districts (level 2), and sub-districts (level 3). Overall, the vintage and level of spatial disaggregation of GADM-digitized maps vary across countries.

To have more control over the accuracy, vintage, and level of spatial disaggregation of maps, we digitize the official maps that are consistent with countries' 2010 or 2011 censuses. We do so for all four selected administrative levels for each country. When possible, we curate such digitized maps from official agencies. Because shapefile is the de facto standard storage format of vector geospatial data in mainstream information technology practices, we digitize the maps into shapefiles. The digitization process consists of five broad steps: scanning paper maps, geo-referencing images of maps, extracting features from geo-referenced maps and creating shapefiles, adding attributions, and validating against satellite imageries and repositioning (illustration 1).

In the case of India, for example, we use all physical maps of the Administrative Atlas of India 2011 as the source maps (Office of the Registrar General and Census Commissioner 2011b). These maps are designed

to be consistent with the Census of India 2011. Thirty-five state/union territory maps, 640 district maps, and 5,923 sub-district maps are used. The sub-district maps present the administrative boundaries of towns and villages belonging to them, the district maps present those of sub-districts, and the state maps present those of districts. At first, these 6,598 physical maps are scanned into high resolution and quality images to prepare for geo-referencing. Errors such as line stripping and blurred images are corrected to ensure quality.

Map geo-referencing, the second step, is the most critical, as it sets the foundation for later work. The process associates the raw image of a map with areas on Earth. In contrast to paper maps, the surface of Earth is not flat; rather, Earth is an ellipsoid or oblate spheroid. An accurate representation of Earth must establish a mathematical model of the shape of its surface and defining point coordinates (e.g., latitude and longitude). Geo-referencing assigns coordinates of such a representation of the Earth to points on the raw image of a map. It establishes links between distinctive points on the map and corresponding points on the Earth. These points are called ground control points. Based on these links, geo-referencing manipulates the geometric features of a raw image by scaling, bending, rotating, and so on. As a result, the image is transformed to match the particular location, size, and shape of the area under consideration on Earth.

Establishing "true" ground control points on the surface of the Earth for such a large number of maps from scratch is difficult. Following the general practice, we choose to geo-reference the source maps with respect to destination maps—shapefiles of maps that have been geo-referenced with relatively good accuracy. We choose version 2.0 of the level 1 Global Administrative Areas (GADM) shapefiles for India as the destination maps (Global Administrative Areas 2011). While the boundaries of districts, sub-districts, and villages/towns changed substantially between 2005 and 2011, the demarcation of states remained the same during that period. Therefore, vintage is not a major concern at this level. GADM shapefiles have been geo-referenced with datum D_WGS_1984 and Geographic Coordination System GCS_WGS_1984. Comparing key features of the level 1 GADM shapefiles against high resolution satellite imageries from Google Earth suggests that the geo-rectification is of high accuracy. Therefore, using the level 1 GADM shapefiles as the destination maps can help us match the source maps to the correct corresponding areas on Earth.

A sequential approach is taken. At first, the state-level maps are geo-referenced with respect to the level 1 GADM shapefiles (illustration 2a). At least 50 ground control points are identified for each state map. In principle, the ground control points should be features that are common to both the source and the destination maps, which helps to characterize the size and shape of the source maps. For example, identifiable intersections and bending features serve as good control points. Three points are required to geo-locate any map. Additional points with distinct characteristics will improve the accuracy of the geo-referencing. These control points establish linkages between the source and the destination maps. Based on these linkages, the destination maps are geometrically transformed, such that the control points overlap with their corresponding points on the destination maps. During the process, the source maps are re-scaled, bent, and even rotated. Ultimately, the source maps match the areas on the destination maps in terms of locations, sizes, and shapes. For this step, the source maps match the destination maps reasonably well, because they are both at the state level. However, a discrepancy arises, mainly because of differences in map vintage, map scale, and possibly the original source. The discrepancy is generally small and does not affect the matching in terms of locations and sizes, but rather the shapes of states at the margins. To improve quality and refine results, additional ground control points are taken, based on high resolution satellite imageries.

Next, the district level maps are geo-referenced with respect to the level 1 GADM shapefiles and the rectified state level maps. Finally, the sub-district level maps are geo-referenced with respect to the level 1 GADM shapefiles, the rectified state level maps, and the rectified district level maps. The procedures are similar. One main difference is the destination maps from which control points are chosen. For the district level maps, the main destination maps are the rectified state level maps; for sub-district maps, the main destination maps are the rectified district level maps (illustrations 2b, 2c). Based on the selected control points, physical maps of the same size are transformed in different ways to match their corresponding areas on the destination maps. When putting all sub-district maps of one district together, the process resembles putting puzzle pieces together to form a certain broad shape. Additionally, the sizes and shapes of the puzzle pieces may also be transformed during the process (illustration 2d).

Once all of the paper maps are geo-referenced, we begin feature extraction, which is the process of storing the features of the maps' raster images into vector format and preserving the information on location, size, and shape acquired during the geo-referencing process. For this spatial database, the main features of interest are the boundaries of administrative units. They are of the form of polygons. Feature extraction thus mainly involves delineating these polygons (illustration 3a). In some cases, villages are presented as points on the source maps, and their boundaries are partially available or unavailable. This situation often occurs for mountainous areas and large forest areas. To generate boundaries for these villages, we combine consulting satellite imageries with proximity analysis.

The source maps also present other land use information, such as forests, lakes, rivers, and mountains. They are of the form of polylines and/or polygons. In many cases, the polylines or polygons of different land uses intersect with the polygons representing administrative boundaries of towns and villages. The boundaries of towns and villages and the features of other land uses are extracted separately to create two corresponding layers of shapefiles at level 4. This way, we can avoid mistaking other land uses as towns and villages while also preserving key features of the source maps. The two layers can be edited, viewed, and analyzed independently or overlaid (illustration 3b).

The level 3 shapefiles are created as an aggregation of these two layers of level 4 shapefiles. The two layers are overlaid, the polylines or polygons within the boundary of each sub-district are dissolved, and the polylines on the boundary of each sub-district are preserved to form the boundary of the level 3 shapefile. Within a district, the boundaries of sub-districts are not independent from each other but share common boundaries. Despite best efforts in the geo-referencing step, some mismatches between adjacent sub-district boundaries may arise, mainly because of differences in map scale and map details. A process called edge matching and "mosaicking" is used to identify and resolve the mismatch. In principle, boundaries that are based on larger map scales (i.e., the same length on the map represents a shorter distance on Earth) or that have more details are relied upon to define the final common sub-district level boundaries. When a distinct geographic or land use feature exists along the border, satellite images are used to help the delineation.

The level 2 shapefiles are created as an aggregation of the level 3 shapefiles within the boundary of each district. The same process of edge matching and mosaicking is used to fine-tune the common boundaries of adjacent districts. At last, the level 1 shapefiles are created as an aggregation of the level 2 shapefiles within the boundary of each state. The same process of edge matching and mosaicking is used, and the GADM shapefiles are also consulted to define the final state level boundaries.

Attributions are added to the shapefiles for both administrative units and other land use areas. For administrative units, attributions include names, location codes, and administrative types. The information is drawn primarily from the source maps (i.e., the Administrative Atlas of India 2011) and is designed to be consistent with the Census of India 2011. However, these source maps sometimes miss information or make mistakes. We rely on location codes of the Meta Data and Data Standards (MDDS) adopted by the Census of India 2011 to correct these errors to the greatest possible extent. For example, in the case of missing information for a village, the MDDS name and code for a village belonging to the same sub-district will be assigned to the village after we conduct a logic check and consult satellite imageries.

In India, some areas in the outskirts of cities are administrative rural, but they are contiguous to cities in terms of their economic and social characteristics. The Census of India 2011 defines these areas as "outgrowths" of cities, but the Administrative Atlas of India 2011 does not consistently identify and present outgrowths across states. In this case, information from MDDS helps to bring more consistency. We also match the population information of Census of India 2011 with the shapefiles to validate and revise the attributions assigned to all administrative units.

These three steps constitute the main production process. Next, the shapefiles are further validated against high resolution imageries, mainly Google Earth imageries. The validation process focuses on two features that can provide meaningful information to assess the quality of the shapefiles: major geographic features, such as coastlines, and built-up images of cities. The shapefiles are repositioned when significant discrepancies are identified.

The validation and reposition step improves the quality of the shapefiles substantially. For example, before the reposition, the mismatches between coastlines depicted by the shapefiles and those visible from Google Earth imageries were as great as 1–2 kilometers. After the reposition, the mismatches are reduced to 100 meters or smaller (illustration 4a).

Regarding validation against built-up images of cities, city expansions do not necessarily follow the master plan. Having boundaries depicted by the shapefiles that do not match the boundaries of the built-up images perfectly is acceptable. However, in some cases, before the validation and reposition process, the city boundary from the shapefiles does not overlap with the built-up image of the corresponding city at all or overlaps poorly. In these cases, this step helps bring the position of shapefiles' city boundaries more in line with the position of corresponding built-up images (illustration 4b). The boundaries of all administrative units are interconnected. Therefore, shifting or modifying the boundary of one will affect the boundary or boundaries of other units, often beyond just the adjacent ones. To avoid overshooting, only cities with populations of 30,000 or more are considered for reposition, and a set of guidelines is followed in the process. For each state, for cities with populations of 50,000 or more, 90 percent of data need to be "reasonably aligned" and 10 percent of data need to be "partially aligned"; for cities with a population between 30,000 and 50,000, 75 percent of data need to be "reasonably aligned." Partially aligned means that between 50 and 75 percent of the urban built-up image falls within the digitized urban administrative boundary; reasonably aligned means that more than 75 percent of the urban built-up image is covered within the digitized urban administrative boundary.

Illustrations 5a–5b are an example of the final shapefiles of all four levels. Illustration 5a shows the digitized administrative boundaries of level 1 to level 3 administrative units for the states of Gujarat and Madhya Pradesh. Illustration 5b shows a more detailed map of the digitized administrative boundaries of level 3 and

level 4 administrative units and land uses of sub-district Vadodara of Gujarat. For level 4 administrative units, their names, their location codes consistent with the location codes of MDDS, and four broad administrative types (statutory towns, census towns, outgrowths, and villages) are accessible as attributions.

Add tiles as the lowest level

In addition to spatial levels based on administrative units available in censuses, we also establish a lower spatial level, "tiles," to incorporate more refined information provided by more modern sources. For example, many remote sensing–based data are available for tiles or gridded cells that cover areas smaller than those defined by even the lowest administrative levels. MODIS Land Cover Type I provides data characterizing five global land cover classification schemes at a resolution of approximately 500 meters. LandScan™ High Resolution Global Population Data Set presents population numbers at a resolution of 30 arc-seconds or 0.008333333 decimal degrees, representing approximately 1 km by 1 km at the equator. To preserve the information at these finer areas, we compute and present indicators at the level of tiles that are appropriate for these datasets.

III. Data sources, themes, and indicators

Socioeconomic and geographic indicators are the substance of this spatial database, just as the spatial hierarchy is its backbone. This section describes the data sources selected to derive these indicators, the thematic structure used to organize these indicators, and the harmonization of these indicators to improve inter-temporal and interspace comparability.

Curate data

The data used by this spatial database can be classified into three broad categories: traditional sources, modern sources, and "mixed sources" (data sources combining traditional and modern data or created by geocoding traditional information). These three broad categories of data can be further classified into eleven types. In the case of India, 30 data sources are tapped into. Table 2 groups these sources according to their data type, title and acronym.

Among traditional sources, Agricultural Prices from India (API), District Crop Production Statistics (DCPS), and Farm Harvest Prices of Principal Crops in India (FHP) are administrative records. Houselisting and Housing Census (PHC—HH), Primary Census Abstract (PHC—PCA), and Population Enumeration (PHC—PE) of the Census of India are census data. Economic Census (EC) is economic census data. Annual Survey of Industries (ASI) and Unincorporated Non-agricultural Enterprises Survey of the National Sample Survey (NSS—ENT) are establishment/firm surveys. District Information System for Education (DISE) and District Level Household and Facility Survey (DLHS) are facility surveys. Annual Health Survey (AHS), Annual Status of Education Report (ASER), DLHS, Household Consumption Expenditure Survey of the National Sample Survey (NSS—HCE), and Employment and Unemployment Survey (NSS—EUE) of the National Sample Survey are household/labor force surveys. State-Wise District Domestic Product (DDP) belongs to National Accounts.

Regarding modern sources, Open Street Maps (OSM) is crowdsourced data. DMSP-OLS Radiance Calibrated Nighttime Lights (RCNTL), Global Land Area with Soil Constraints (GASC) data, MODIS Land Cover Type I (MODIS) product, the four climate-related products of NASA Earth Observations (NEO), and Shuttle Radar Topography Mission - DEM v.2.1 (SRTM) data primarily are remote sensing data.

Mixed sources include Climate Research Unit Database v.3.22 (CRU), Global Map of Irrigation Areas (GMIA), and LandScan™ High Resolution Global Population Data Set (LANDSCAN), which combine both remote sensing and traditional information. Mineral Facilities of Asia and the Pacific (MFAP) and World Database on Protected Areas (WDPA) are geo-referenced traditional data. For example, MFAP is a geo-referenced database collected by the United States Geological Survey.

Data sources are selected based on a combination of factors, including quality, level of spatial disaggregation, spatial coverage, and temporal coverage. One case in point on quality is land use data. To the best of our knowledge, three sources provide land cover information at the global scale as open source data: Global Land Cover-SHARE, GlobCover, and MODIS (table 3). GlobCover is of the highest resolution at 300 meters. However, MODIS applies a more stringent process for conducting its classification and achieves higher accuracy rates both globally and for urban land (Friedl et al. 2010). Because of its higher quality, we select MODIS as the data source for land use–related indicators.

Source selection also is affected by the level of spatial disaggregation of data, because it determines the lowest spatial level of the indicators derived from the source. In the case of India, table 4 lists the level of spatial disaggregation for each data source. For a traditional source, the level of disaggregation is affected by its design and the information disclosed by its originator. As an example, EC 2005 is conducted at the village level in rural areas; however, its originator discloses complete information for the district in which an enterprise locates, but not information about the sub-district or the village. Therefore, the indicators derived from it can be accurately geo-coded only to the district level. Modern and mixed sources generally come as raster data with a resolution (e.g., 500 meters) and can be geo-referenced to any spatial level. However, when the original resolution of a data source is larger than the sizes of most administrative units of a spatial level, the derived indicators will have smaller variation across units.

Regarding temporal coverage, we decide to capture the state of each country around two points in time, 2001 and 2011. We do this mainly because censuses are our organization framework, and a majority of countries in the region conduct their censuses around these two years. However, other data sources are not always available for the same points in time. Data sources available for other years are allocated to 2001 and 2011, respectively. Table 4 presents the details of the time allocation for each data source for the India component. For sources available for one year, we allocate the data to either 2001 or 2011, depending on the time differences. For example, DLHS is available for 2007—08, and we allocate it to 2011. For sources available for multiple years, we select the year of data that is closer to 2011 (or 2001), and with better quality, lower levels of spatial disaggregation, and broader spatial coverages (e.g., more districts than other years).

To improve quality and spatial coverage, we also triangulate the same data source of different years. Data from ASER is available for 2009, 2010, and 2011, but each round presents information for only a fraction of districts. We use 2011 round data as the primary source but also incorporate information from 2009 and 2010 rounds to cover more districts. For some indicators, the data source is available annually, and a longer

reference period is preferred. One example is climate-related indicators derived from CRU data. We use a few decades of data to compute the indicators of interest, such as decadal average temperature and precipitation abnormality.

Harmonize themes and indicators

By curating such a wide range of data sources, our aim is to cover key socioeconomic and geographic topics and make the spatial database multi-thematic. For that, we construct and organize indicators into twelve broad themes. To ensure cross country comparability of these themes, we refer to the classifications used by recent statistical year books of South Asian countries (Statistical Year Book of Bangladesh 2010, Statistical Year Book India 2014, Statistical Pocket Book Nepal 2010, and Pakistan Statistical Year Book 2011), the 2012 United Nations Statistical Year Book, and the World Development Indicators and Data page of the World Bank. The resulting twelve broad themes are (following the order of their appearance in the database): *urban extent, demographics, jobs, economic activity, infrastructure, information technology, finance, business, living standards, health, education, and environment*.

Each theme is further divided into sub-themes. For example, the *urban extent* theme presents the footprint of urbanization in terms of area, population, and population density. It includes three sub-themes: administrative, built-up, and lit at night. These sub-themes adopted different concepts of "urban" and define it based on census statistics, built-up images, and images of lit at night, respectively. As another example, the *jobs* theme consists of four sub-themes: labor force, unemployment, employment, and earnings. The labor force sub-theme provides indicators on working age population and labor force participation; the unemployment sub-theme reports unemployment and underemployment rates; the employment sub-theme describes employment structure by work status and by sector; and the earnings sub-theme reports labor earnings by gender and by sector.

Another major undertaking of this spatial database is constructing indicators such that they are relatively harmonized over time and across countries in the region. National statistical systems in countries of the region have made notable progress in terms of coverage, reliability, standardization, and efficiency. However, each country's statistical system is still in the process of development and has its own idiosyncrasies (e.g., questionnaires on the same topic are defined differently over time and across countries). We rely on international standards to consolidate data collected by these different questionnaires into harmonized indicators.

A case in point on harmonization over time is the indicator on improved source of sanitation (illustration 6). According to the World Health Organization (WHO) and United Nations Children's Fund (UNICEF) Joint Monitoring Programme for Water Supply and Sanitation, among different types of latrines, only "ventilated improved pit latrine" and "pit latrine with slab" are considered a source of improved sanitation; "pit latrine without slab" and "open pit" are considered unimproved. For India, the PHC—HH 2011 questionnaire distinguishes between these different types of pit latrines. However, the PHC—HH 2001 questionnaire does not make such a distinction and only asks whether a pit latrine exists within the premises of the household. To reconcile this difference, we construct two indicators: one on enhanced improved sanitation (which is consistent with the WHO/UNICEF criteria and is available only for 2011) and one on improved sanitation (which treats all types of pit latrines as improved and is available for both 2001 and 2011 for India). The first indicator allows comparison between India and other countries, and the second one allows intertemporal comparison within India.

The indicator on improved source of water illustrates the point on harmonization across countries (illustration 7). Among different types of wells and springs, only a protected or covered well or spring is considered an improved source of water by the WHO/UNICEF Joint Monitoring Programme for Water Supply and Sanitation. For Bangladesh, the questionnaire of the Housing and Population Census 2011 does not distinguish between protected and unprotected well or spring. For India, the PHC—HH 2011 questionnaire distinguishes between protected and unprotected well but does not do so for spring. To ensure consistency between the two countries, we consider both well and spring as an unimproved source. The criteria is more stringent than the international standard but allows comparison across countries within the region.

Original data of modern and mixed sources are also treated to improve intertemporal and interspace comparability. Treating nighttime lights data is a good example. The National Oceanic and Atmospheric Administration (NOAA) provides three products for nighttime lights: DMSP-OLS Nighttime Lights, DMSP-OLS Radiance Calibrated Nighttime Lights (RCNTL), and VIIRS Nighttime Lights. Table 5 presents the key characteristics of all three products. In a nutshell, imagery collected by the sensor of DMSP-OLS satellites can detect visible band light sources such as city lights, gas flaring, shipping fleets, and so on. NOAA developed a methodology to identify stable lights at approximately 1 square kilometer and first treated the data by removing sunlit, glare, moonlit, and other temporal lights. This methodology is applied to DMSP-OLS Nighttime Lights and RCNTL but not VIIRS Nighttime Lights. This makes the first two products better candidates for intertemporal and interspace comparison of stable nighttime lights.

One problem related to both intertemporal and interspace comparison is saturation at bright cores. Because data collected by DMSP-OLS satellites is at a high setting, the sensor cannot accommodate bright sources. Bright sources become saturated, and the heterogeneity within the bright core is not captured. As a result, the intensity of light is top coded or truncated from the top for DMSP-OLS Nighttime Lights. This prevents accurate comparison of places with bright lights, over time and across locations. To address this issue, NOAA uses images taken at different gain settings and blends them with DMSP-OLS stable lights to recover more nighttime lights details, especially for the bright sources. This has been applied to RCNTL. In addition, satellites have different sensitivities and are deployed in different years. An inter-satellite procedure is applied to address this issue for RCNTL but not for DMSP-OLS Nighttime Lights (Wu et al. 2013). Given these reasons, we select RCNTL for this spatial database.

Finally, the DMSP-OLS satellite's sensors degrade over time. Even if the same intensity of light is obtained for the same location for two different years, it does not necessarily mean the light intensity has remained the same. To address this problem, we apply an inter-annual calibration adjustment using parameters computed by Hsu and others (2015) on the data of RCNTL. After this adjustment, the total intensity of lights in India shows a much steeper increase between 1999 and 2010 (figure 1). By using RCNTL and applying inter-annual calibration, we ensure greater intertemporal and interspace comparability of indicators based on nighttime lights data.

Crowdsourced data is a special case where the comparability over time and across space is limited by the nature of this type of information. The term crowdsourcing first was coined to capture the idea of outsourcing to the crowd. It is a distributed problem-solving and production process that involves outsourcing tasks to a network of people in the form of an open call. With the growing prominence of the Internet, the process becomes more related to online open call and participative online activity. A particular

crowdsourced data is, thus, the result of a participative online data collection activity. Inevitably, the quality of such data depends on individual contributors' knowledge, experience, and dedication. The intertemporal and interspace comparability of such data depends on the distribution of these individuals over time and across spaces. We use OSM to compute indicators related to connectivity through road and railway networks for 2011. Information on the distribution of contributors is not available. Therefore, we cannot comment on nor improve the comparability.

IV. Geo-reference indicators

As the last but not least critical step, organized and harmonized indicators need to be geo-referenced to be "spatial." Indicators are geo-referenced such that the levels of spatial disaggregation of all indicators are matched with the established spatial hierarchy. This section will describe the matching process. For indicators derived from traditional sources, the matching mainly entails creating concordance between the location codes of the indicators and those of the corresponding spatial units. For indicators based on modern or mixed sources, the matching mainly involves geo-referencing the original data and performing intersection between the original data and the shapefiles of spatial units. Both types of matching processes will establish a link between indicators and the spatial features of the corresponding spatial units (e.g., locations, sizes, and shapes) to enable visualization and spatial analysis.

Indicators from traditional sources

Geo-referencing indicators derived from traditional data sources is a relatively straightforward process. It can be classified into three steps: establishing a set of location codes for the selected spatial units as the destination codes, constructing a concordance between the location codes used by the original source and the destination codes, and computing indicators for the destination location codes based on indicator values for the original location codes.

Following the definition of the spatial hierarchy, we select the location codes associated with 2011 (or 2010) censuses as the destination location codes for spatial level 1 to level 4. Because these location codes are used by the official maps of administrative units, the linkage between indicators and the geographic features of the administrative units are naturally established. However, constructing the concordances between location codes of traditional sources and these destination codes proves to be another challenge, because the administrative structures of South Asian countries experienced substantial changes between 2001 and 2011, and different data were collected at different points in time along this changing process.

For India, for example, we use the location codes of the Meta Data and Data Standards (MDDS) adopted by the Census of India 2011 as the destination codes. Most traditional sources used by this database for the India component were created between 2001 and 2011. During this relatively short period, however, the number of administrative units had grown. 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. During the process, their geographic demarcations changed accordingly. Some districts were created by being carved out from one existing district, while others were created by merging a few parts of existing districts. The same applies to sub-districts and villages/towns.

The sampling frames of most traditional sources relied on information from the Census of India 2001 (including the location codes) but were adjusted to incorporate these changes. As a result, the location codes of these sources do not match with the location codes of the MDDS perfectly, and the extent of mismatches depends on the time that a data source was created and the design of its sampling frame.

To build concordances for these data sources, we track down the changes of administrative units between 2001 and 2011. Because the lowest spatial level of the indicators derived from a majority of traditional data sources is district level (level 2), we build the timeline of change for each district on an annual basis. We do so by referring to the official websites of districts, the maps of the Administrative Atlas of the Census of India 2011, and the MDDS, which provides a concordance between administrative units defined by the Census of India 2001 and those defined by the Census of India 2011. Based on these sources, we identify 51 of 593 districts that existed in 2001 that experienced some form of modification and 47 new districts that were created. These changes took place in 17 states/union territories. In the state of Punjab, for example, parts of Amritsar district were carved out to form Tarn Taran district in 2005; in the same year, parts of Ropar district and parts of Patiala were merged to form another new district, Sahibzada Ajit Singh Nagar (illustration 8). This timeline of changes at the district level constitutes the basis of the concordance by presenting the relationship between existing and new districts on an annual basis.

We also track the changes for sub-districts and villages/towns but on a decade basis. This is sufficient, because census is the only traditional source that allows us to compute indicators at these two levels. The MDDS provides a concordance between administrative units defined by the Census of India 2001 and those defined by the Census of India 2011 for these two levels. It constitutes the basis for tracking the changes and constructing the final concordance. But the concordance provided by MDDS is not perfect. Figure 2 shows the number of mismatched sub-districts and villages/towns between the location codes presented by MDDS for 2001 and the location codes extracted from the Census of India 2001 directly. We correct these errors by referring to the information from official websites (mostly of cities and towns), the maps of the Administrative Atlas of the Census of India 2011, and the data from the 2001 and 2011 censuses. The information available at these two spatial levels is more limited. We focus on making corrections for places with relatively large populations. Our correction efforts reduce the mismatched population to 8.6 million down from 22.1 million, or to 0.8 percent of total population down from 2.1 percent. Figure 3 presents the reduction in mismatched population by state.

Computing indicator values for the destination location codes is straightforward once the concordances are constructed. For indicators in the form of aggregates (e.g., population), we reallocate the values associated with the parent administrative units based on the original location codes to the child units based on the destination location codes, either through summation or subtraction. For indicators in the form of percentages, means, standard deviations, or other descriptive statistics, we compute the weighted average of the value associated with the parent administrative units for the child units when weights can be constructed (e.g., population as weights) and compute the simple average when weights are not available.

Indicators from modern and mixed sources

Geo-referencing indicators derived from modern or mixed sources starts with making sure the projection systems of these data are consistent with those of the shapefiles of the administrative units. In particular, we use D_WGS_1984 as the datum and GCS_WGS_1984 as the Geographic Coordination System for the shapefiles of administrative units. It is the most popular projection system at the global level and is also

used by Google Earth. When the projection system of the original data differs from this, such as the MODIS product, a transformation of the original projection system is performed.

As a next step, we perform intersection between the original data and the shapefiles of spatial units (illustration 9). We generate a vector version of the raster data source, which also consists of tiles or cells. We then overlay the shapefile of the administrative units for a spatial level over the vector file and create the intersection of both 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. Computing values of indicators for these administrative units are then confined to these tiles (and/or fractions of tiles). If the indicators are in the form of aggregates, such as areas or population, we compute the indicator for an administrative unit by summing up the values of all tiles (and/or fractions of tiles) within it. The same logic applies when the indicators are in the form of percentages, means, or other descriptive statistics. Overall, the intersection defines the spatial scope of the computation for each administrative unit at a spatial level.

To conclude this section, the geo-reference process ensures that the levels of spatial disaggregation of all indicators are consistent with the established spatial hierarchy. The complete spatial hierarchy of this spatial database includes five levels. Indicators available at a certain spatial level are also available at all higher levels (or more aggregated levels). For example, population density derived from LandScan™ High Resolution Global Population Data Set is not only available for the level of gridded cells but also for the selected administrative levels of the spatial hierarchy. In the case of India, all twelve themes have indicators available for states and districts, ten have indicators available at all four administrative levels down to villages/towns, and four at all five spatial levels. In terms of indicators, 475 are available for states, 432 for both states and districts, 197 for all four administrative levels, and 128 for all five spatial levels, including tiles.

V. Concluding remarks

This spatial database is a work in progress. In comparison with many spatial portals, it relies on a consistent framework in terms of spatial hierarchy and thematic areas. By doing so, it aims to better integrate data of different types and multiple topics and improve spatial data access by general users without GIS experience. We acknowledge that the framework is far from perfect. The India component of the database is the most complete, and the Bangladesh component is in progress. Taking lessons learned from these experiences, we are revising the framework and developing our methodologies. For instance, right now, only a subset of indicators can be geo-referenced to all five levels. We are developing methodologies to expand the subset. New geospatial data are emerging. We will curate new data while expanding the country coverage. Last but not least, we expect to learn from the valuable feedback of users of the current version of the database.

References

Bangladesh Bureau of Statistics. (2011). *Bangladesh Population and Housing Census Socioeconomic and Demographic Report*. Ministry of Planning, Government of the People's Republic of Bangladesh.

Bangladesh Bureau of Statistics. (2014). *Bangladesh Population and Housing Census: Zila Report: Barguna*. Ministry of Planning, Government of the People's Republic of Bangladesh.

Bhattacharyya, H. (2005). Federalism and Regionalism in India: Institutional Strategies and Political Accommodation of Identity. South Asia Institute, Department of Political Science, University of Heidelberg. Working Paper No. 27.

Bicheron, P., P. Defourny, C. Brockmann, L. Schouten, C. Vancutsem, M. Huc, S. Bontemps, M. Leroy, F. Achard, M. Herold, F. Ranera, and O. Arino. (2008). GLOBCOVER: Products Description and Validation Report. European Space Agency. Toulouse Cedex 9, France. http://due.esrin.esa.int/files/p68/GLOBCOVER_Products_Description_Validation_Report_I2.1.pdf

Friedl, M. A., Sulla-Menashe, D., Tan, B., Schneider, A., Ramankutty, N., Sibley, A., and Huang, X. (2010). MODIS Collection 5 global land cover: Algorithm refinements and characterization of new datasets. Remote Sensing of Environment, 114, 168–182.

Global Administrative Areas. (2011). http://www.gadm.org/

Hsu, F. C., Baugh, K. E., Ghosh, T., Zhizhin, M., and Elvidge, C. D. (2015). DMSP-OLS Radiance Calibrated Nighttime Lights Time Series with Intercalibration. *Remote Sensing*, 7(2), 1855-1876.

Khemani, S. (2007). Does delegation of fiscal policy to an independent agency make a difference? Evidence from intergovernmental transfers in India. *Journal of Development Economics*, 82(2), 464–484.

Kochanek, S., and Hardgrave, R. (2007). *India: Government and politics in a developing nation*. Cengage Learning.

Latham, John, Renato Cumani, Ilaria Rosati and Mario Bloise. (2014). FAO Global Land Cover SHARE (GLC-SHARE) Beta-Release Version 1.0, Land and Water Division, Food and Agriculture Organization of the United Nations (FAO), Rome, Italy. http://www.glcn.org/downs/prj/glcshare/GLC_SHARE_beta_v1.0_2014.pdf

Mollah, M. A. H. (2007). Administrative Decentralization in Bangladesh: Theory and Practice. *International Journal of Organization Theory and Behavior*, 10(1), 1.

Office of the Registrar General and Census Commissioner. (2001). Census of India 2001. Ministry of Home Affairs, Government of India. http://censusindia.gov.in/

Office of the Registrar General and Census Commissioner. (2011a). Census of India 2011. Ministry of Home Affairs, Government of India. http://censusindia.gov.in/

Office of the Registrar General and Census Commissioner. (2011b). Census of India 2011: Administrative Atlas of India. Ministry of Home Affairs, Government of India. http://censusindia.gov.in/

Panday, P. K. (2011). Local government system in Bangladesh: How far is it decentralized? *Lexlocalis—Journal of Local Self-Government*, 9(3), 205–230.

Rahman, M. S. (2012). Upazila Parishad in Bangladesh: Roles and Functions of Elected Representatives and Bureaucrats. *Commonwealth Journal of Local Governance*.

Rao, M. G., and Singh, N. (2001). Federalism in India: Political Economy and Reforms.

The Constitution of India. (2015). Government of India: New Delhi. http://indiacode.nic.in/coiweb/welcome.html

The Constitution of the People's Republic of Bangladesh. (2015). Government of Bangladesh: Dhaka. http://bdlaws.minlaw.gov.bd/pdf_part.php?id=367

United Nations. (2008). Principles and Recommendations for Population and Housing Censuses (Revision 2). United Nations: New York.

Wu, J., He, S., Peng, J., Li, W., and Zhong, X. (2013). Intercalibration of DMSP-OLS night-time light data by the invariant region method. *International journal of remote sensing*, 34(20), 7356-7368.

http://www.tandfonline.com/doi/abs/10.1080/01431161.2013.820365#preview

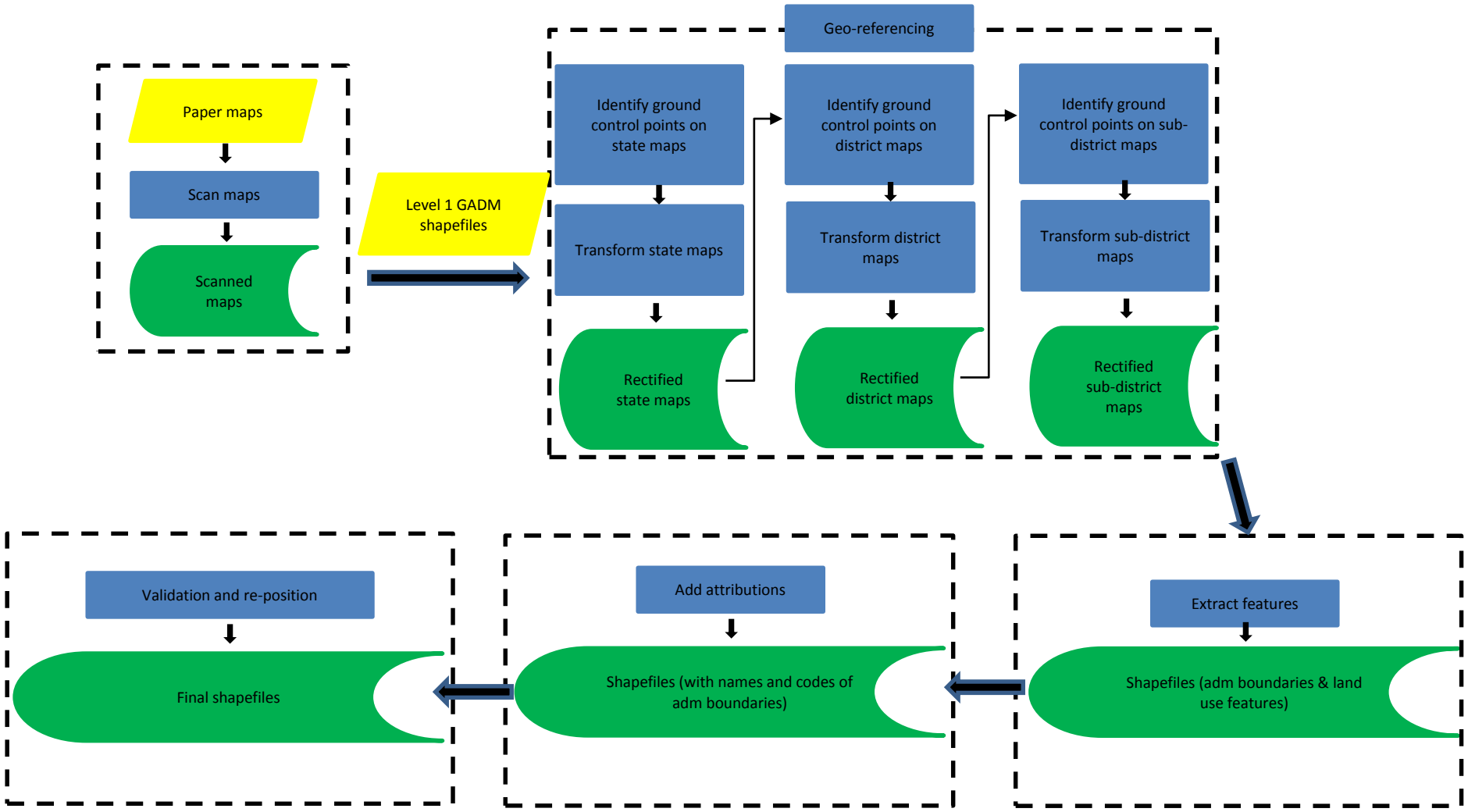
Figures, illustrations, and tables

Table 1. Selection of administrative levels based on censuses

<table>
<tr><th>Spatial levels</th><th colspan="3">Administrative levels</th></tr>
<tr><td></td><td colspan="2">Population and Housing Census 2011, Bangladesh</td><td>Census of India 2011</td></tr>
<tr><td>1</td><td colspan="2">Divisions</td><td>States/Union Territories</td></tr>
<tr><td>2</td><td colspan="2">Districts (Zilas)</td><td>Districts</td></tr>
<tr><td>3</td><td>Sub-districts (Upazilas)</td><td>Thanas</td><td>Sub-districts</td></tr>
<tr><td>4</td><td>Unions</td><td>Wards</td><td rowspan="3">Villages</td><td>Towns</td></tr>
<tr><td></td><td>Mouzas</td><td rowspan="2">Mohallas</td><td rowspan="2">Wards</td></tr>
<tr><td></td><td>Villages</td></tr>
</table>

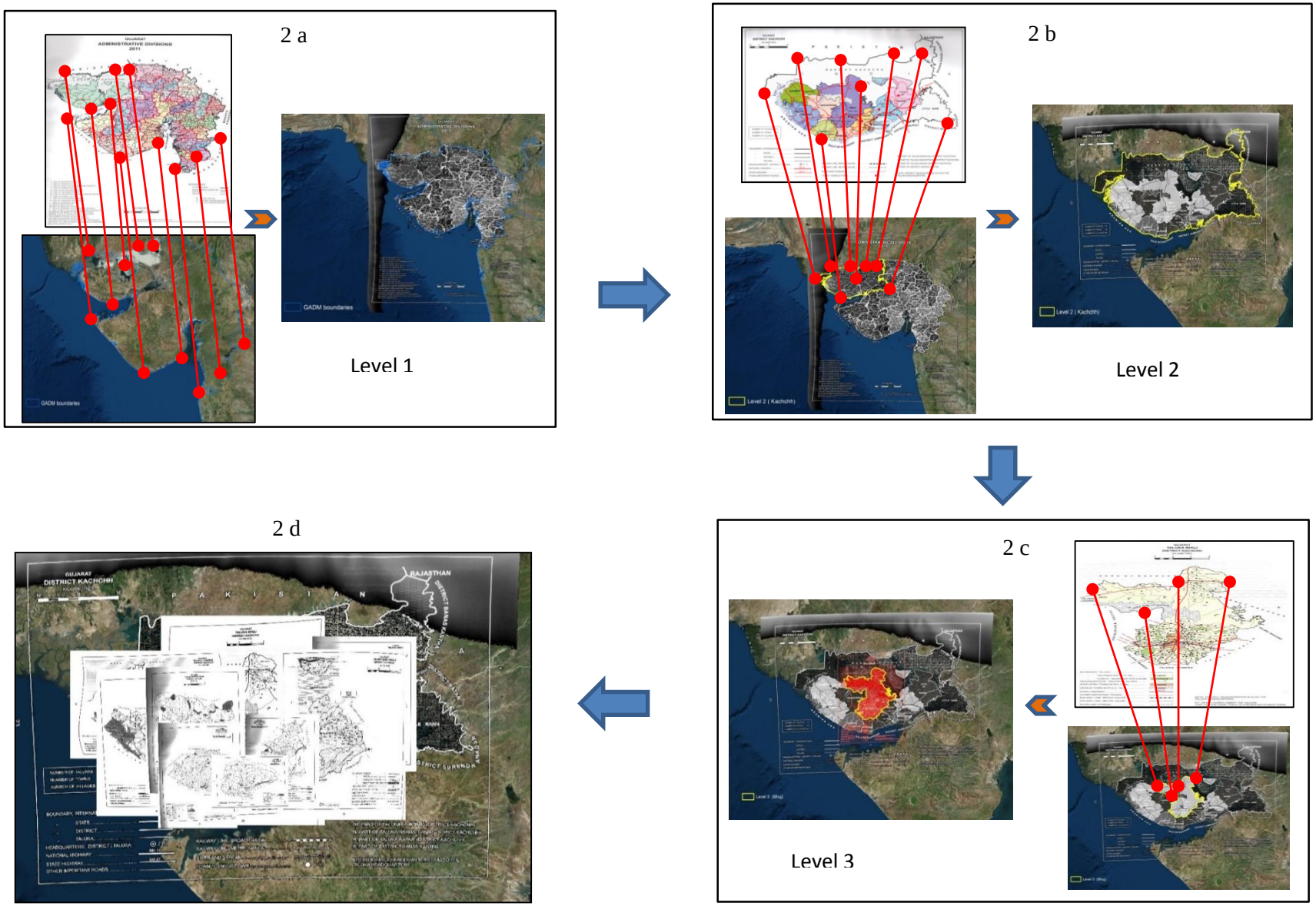
Source: authors, based on Bangladesh Bureau of Statistics (2011, 2014) and Office of the Registrar General and Census Commissioner (2011a).
Note: the highlighted levels are selected for the spatial hierarchy of the database.

Illustration 1. Paper map digitization process



Source: authors

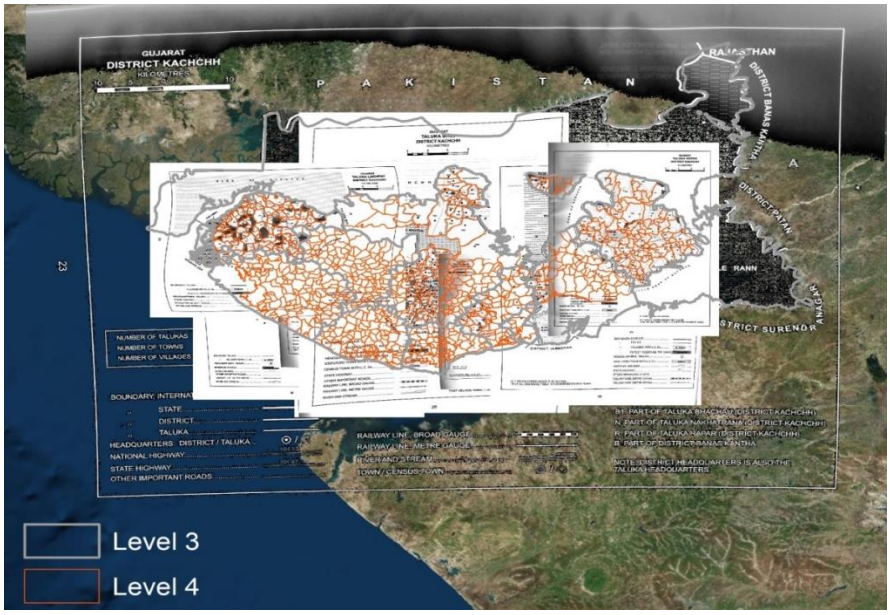
Illustrations 2a–2d. Geo-referencing scanned map



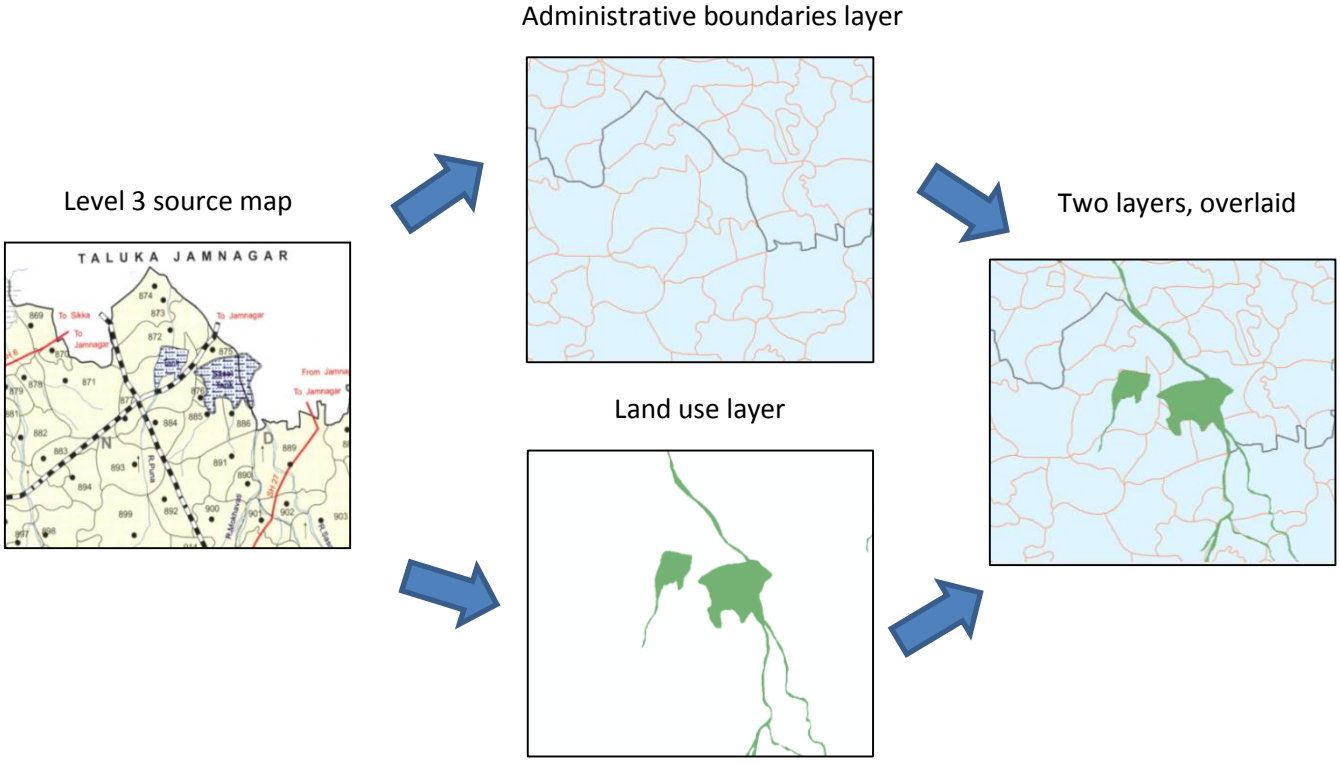
Source: authors

Illustrations 3a–3b. Extracting features

3 a

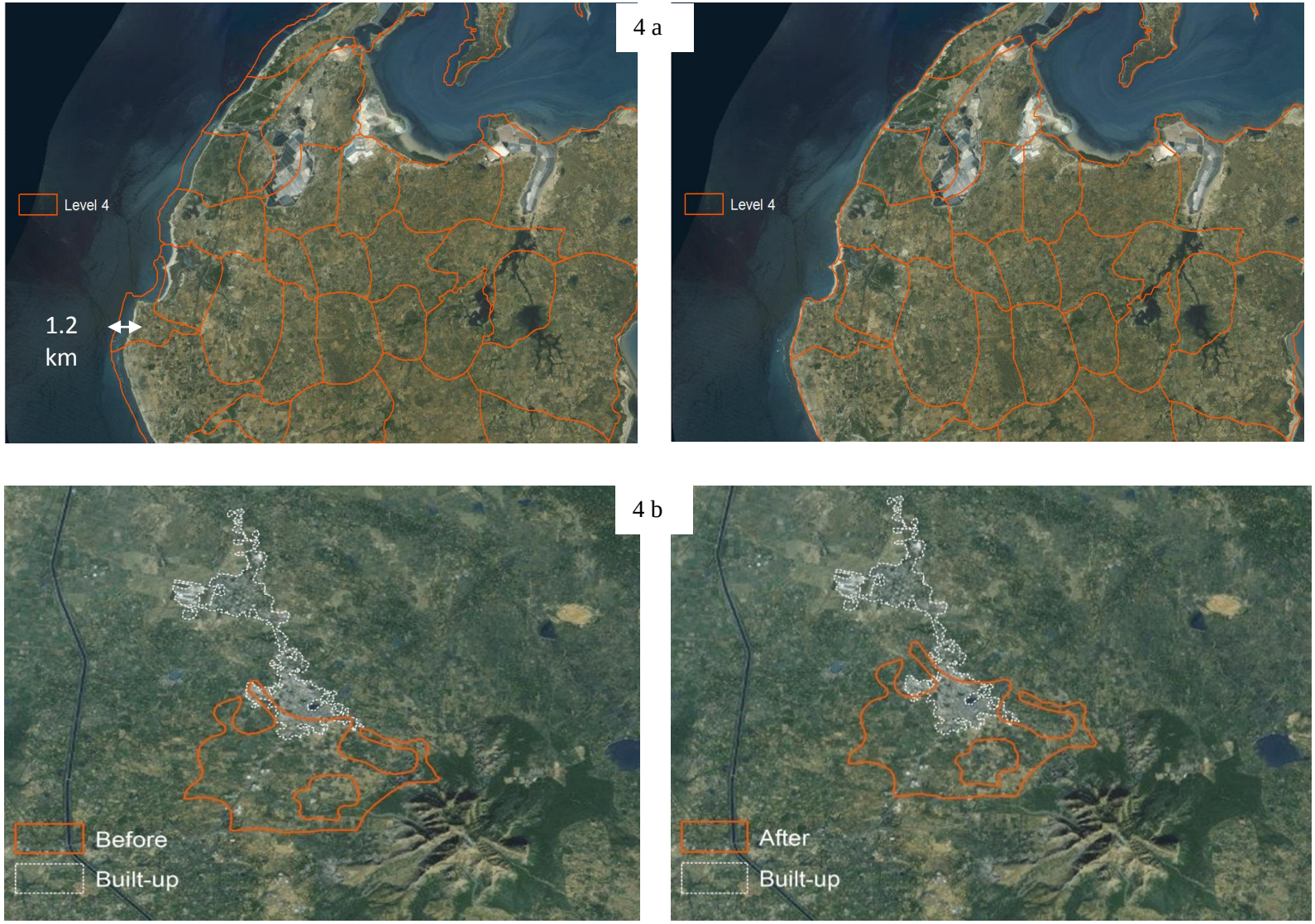


3 b



Source: authors

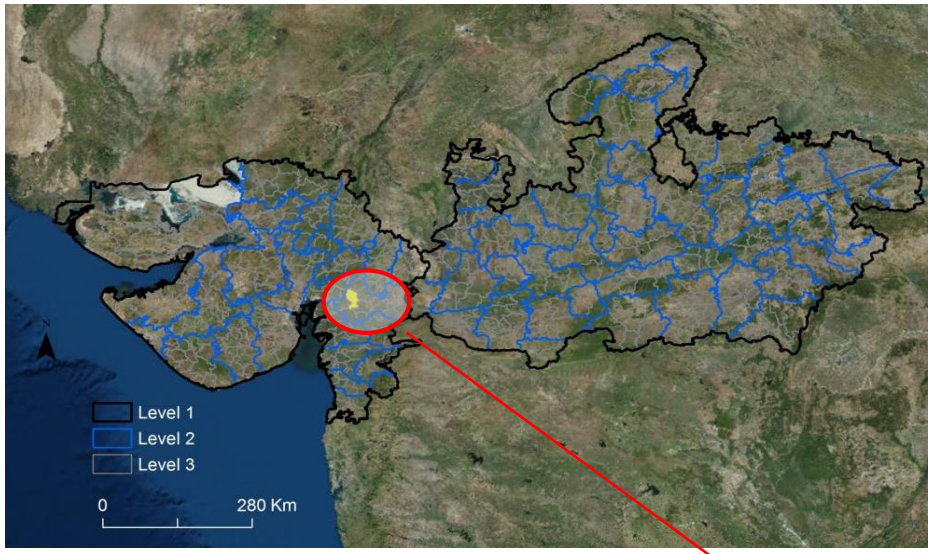
Illustrations 4a–4b. Validation and reposition



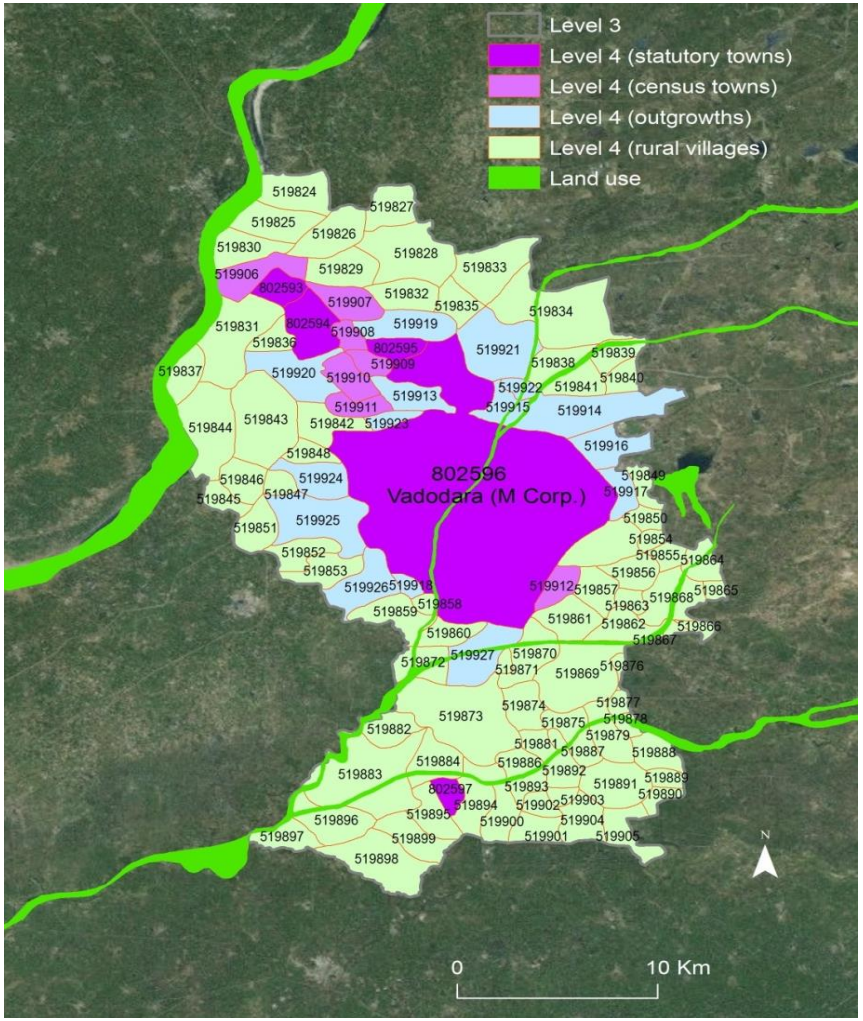
Source: authors

Illustrations 5a–5b. Final shapefiles of four levels of administrative units

5 a Gujarat and Madhya Pradesh



5 b Gujarat: Vadodara (Level 3)



Source: authors

Table 2. Data sources for India: acronyms and types

	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 Housing Census	PHC—HH		X									
	Population and Housing Census_Primary Census Abstract	PHC—PCA		X									
	Population and Housing Census_Population Enumeration	PHC—PE		X									
	Economic Census	EC			X								
	Annual Survey of Industries	ASI				X							
	National Sample Survey_Enterprises	NSS—ENT				X							
	District Information System for Education	DISE					X						
	District Level Household and Facility 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				

(Table 2 cont.)

	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™ 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

Table 3. Compare data sources on land use

Name	Originator	Resolution	Year	Number of Land Classes	Method	Accuracy
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).

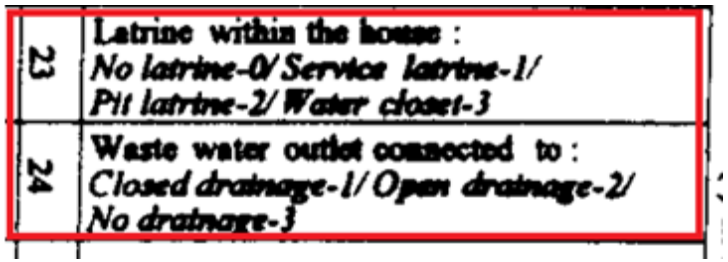
Table 4. Data sources for India: spatial levels and temporal coverages

	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.

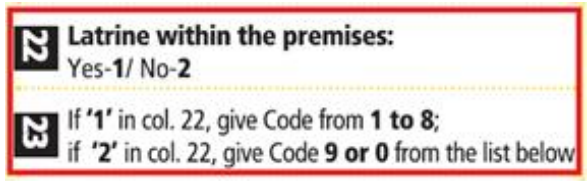
Illustration 6. Harmonize traditional indicators over time: source of improved sanitation

PHC—HH 2001 questionnaire

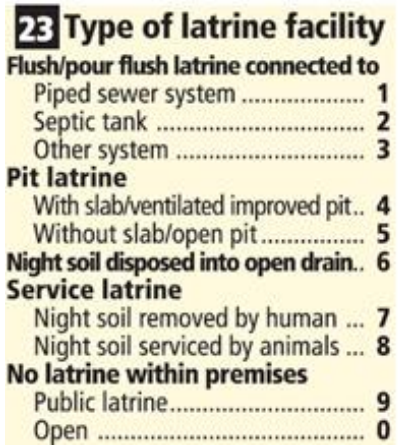


23	Latrine within the house : *No latrine-0/ Service latrine-1/ Pit latrine-2/ Water closet-3*
24	Waste water outlet connected to : *Closed drainage-1/ Open drainage-2/ No drainage-3*

PHC—HH 2011 questionnaire



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

Solution:
Define 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

Source: authors based on PHC-HH 2001, 2011.

Illustration 7. Harmonize traditional indicators across countries: improved source of water

Census 2011, Bangladesh, questionnaire

Source of drinking water – 2011 long form

[] Tap
[] Tube-well / Deep tube-well
[] Well
[] Pond
[] River/ Ditch / Canal
[] Other

PHC—HH 2011 questionnaire

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:
Define improved water: stricter than WHO/UNICEF standard, both well and spring are considered unimproved source

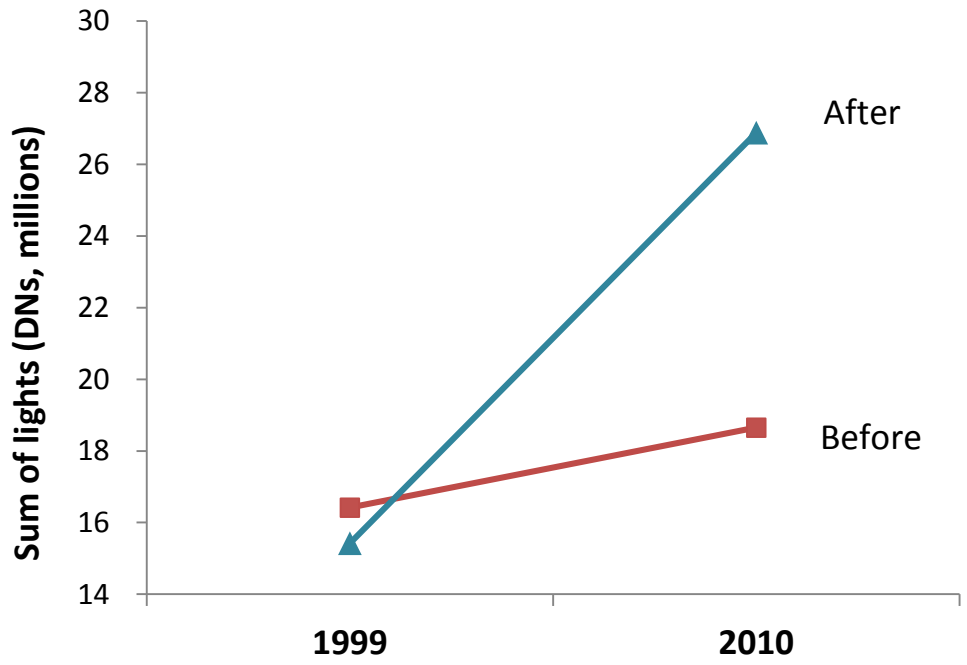
Source: authors based on PHC—HH 2011 and Census 2011, Bangladesh

Table 5. Harmonize modern indicators over time: compare data sources on nighttime lights

Products	Resolu-tion	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
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).

Figure 1. Harmonize modern indicators over time: inter-annual calibration of nighttime lights



Source: authors, based on DMSP-OLS Radiance Calibrated Nighttime Lights 1999 and 2010

Note: the figure shows the sum of light intensity for 1999 and 2010 after applying an inter-annual calibration adjustment using parameters computed by Hsu et al. (2015)

Illustration 8. The timeline of changes of districts in Punjab

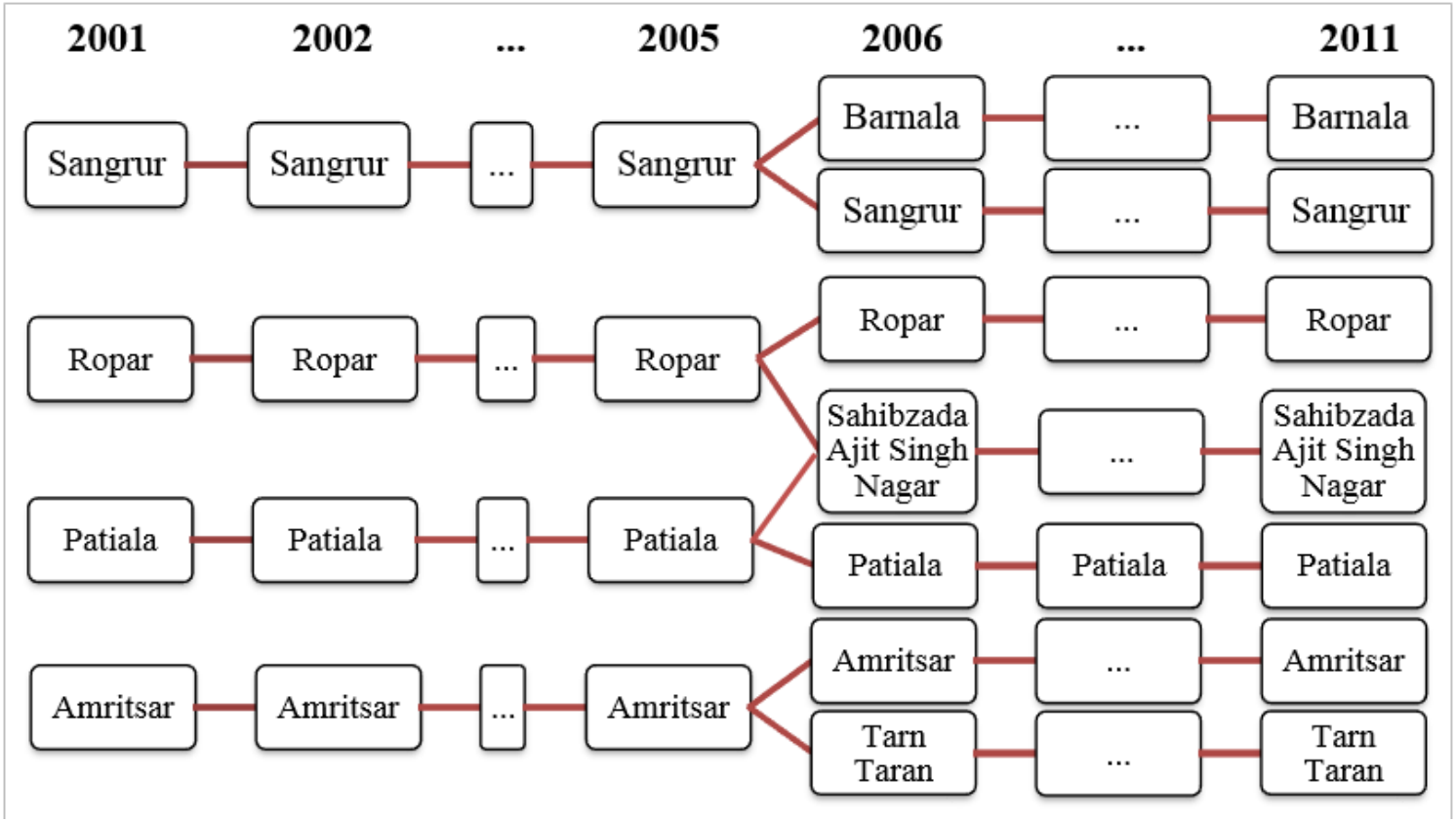
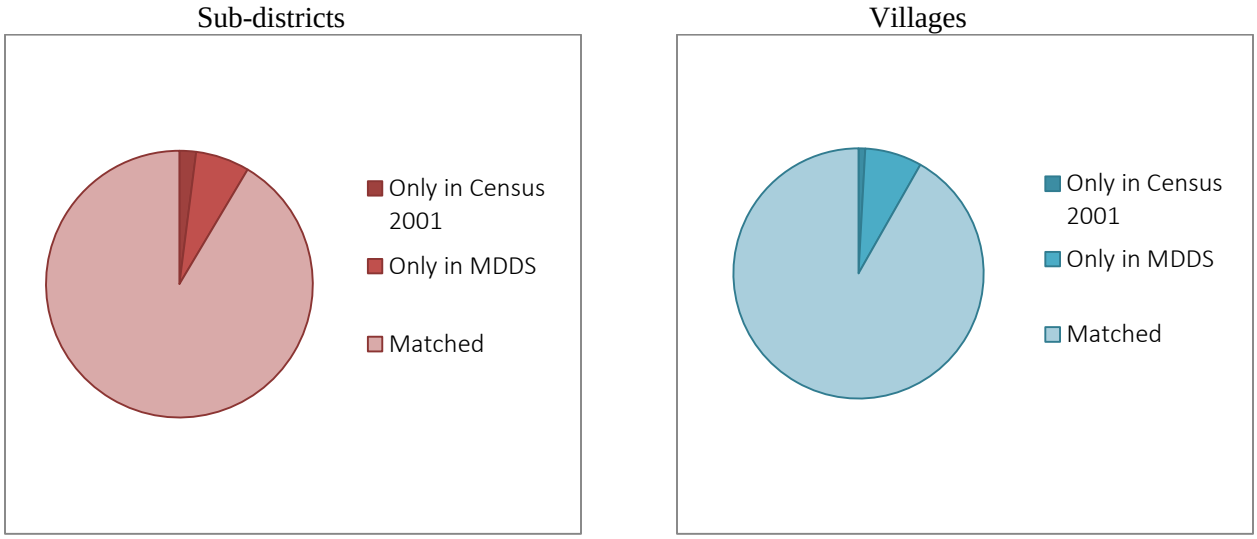
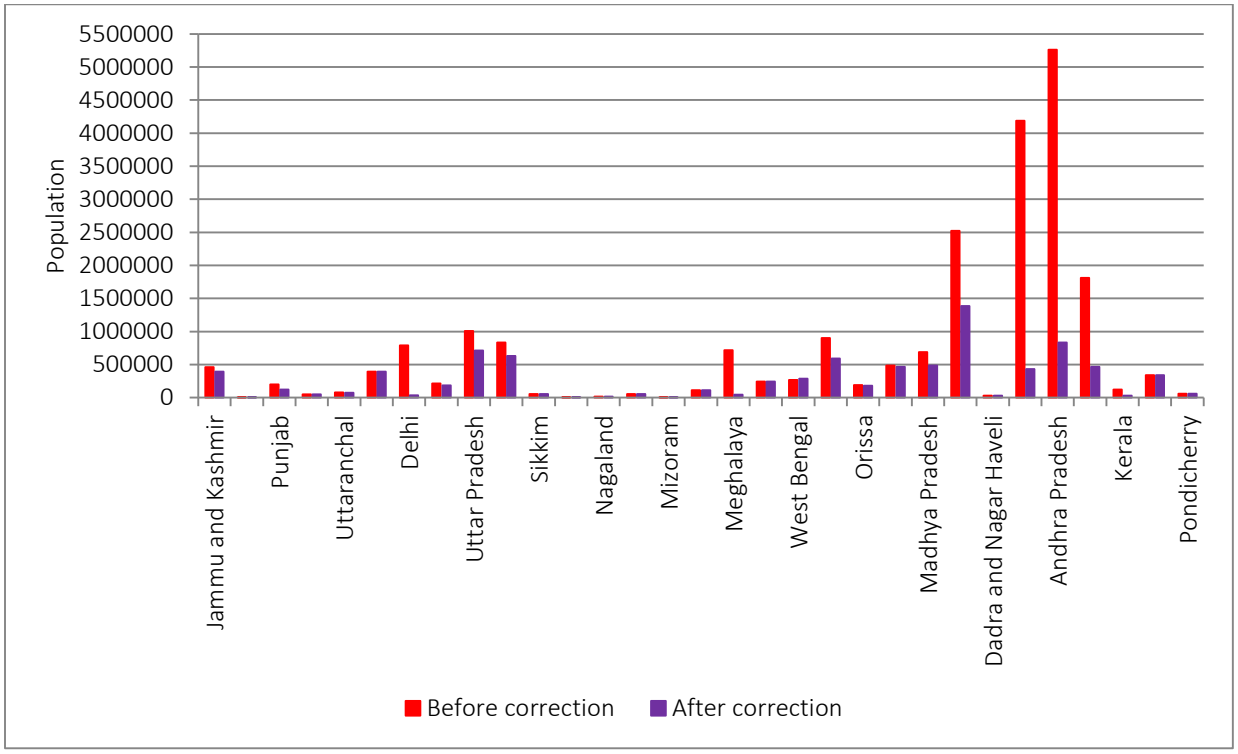


Figure 2. Mismatches of 2001 location codes between MDDS and Census of India 2001



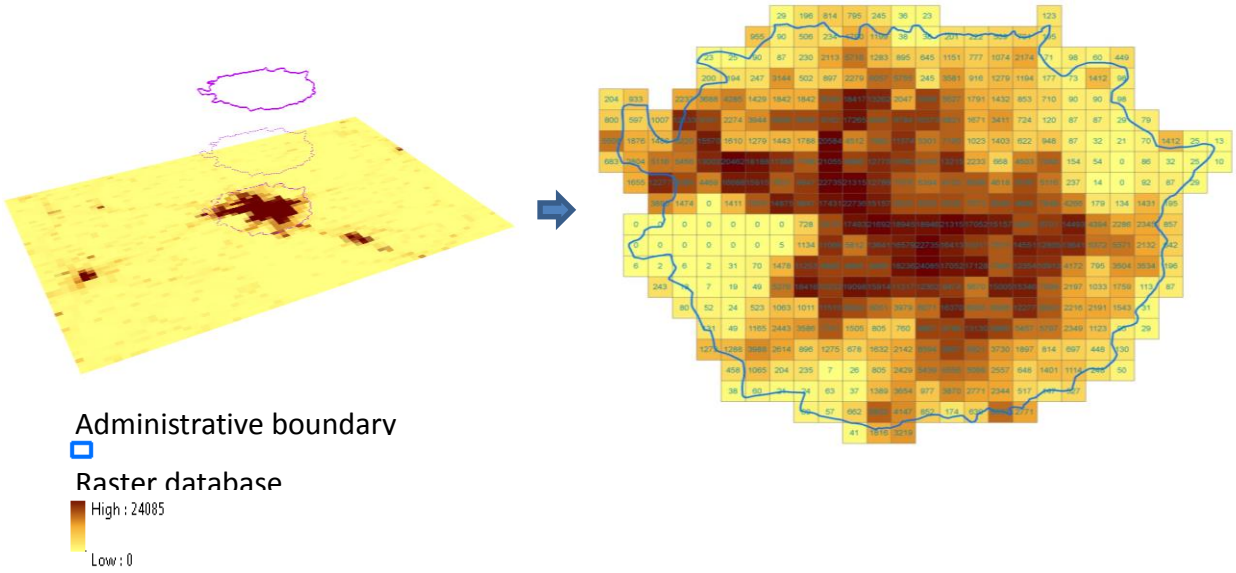
Source: authors, based on Census of India 2001, 2011

Figure 3. Improvement of mismatches between MDDS and Census of India 2001



Source: authors, based on Census of India 2001, 2011

Illustration 9. Geo-reference remote sensing data



Source: authors