Chapter 8

Other 2001 Settlement-Specific Spatial Data

8.1 Meiyappan (2018) polygon settlement files for 2001

Research leading to the publication of Meiyappan et al. (2016) included the development and posting to SEDAC of spatial boundary data for settlements in the 2001 Indian census: for the spatial data, see Meiyappan et al. (2018). These data were downloaded from the SEDAC server on 23 July 2018,[1] The downloaded documentation is in `/Programs/Spatial/Meiyappan/Data/SEDAC_Villages_1991_2001/Documentation`.

The 2001 data cover urban as well as rural settlements, but do not include within-urban wards. To quote from the SEDAC website, the settlement boundaries were "developed by digitizing village/town level boundaries from the official analog maps published by the Survey of India for 2001. This data set also utilized tabular data for 1991 and 2001 from the Primary Census Abstract (PCA) and Village Directory (VD) data series of the Indian census." More detail is provided in the documentation, which explains how control points were used to improve the spatial consistency of the digitizing. Apparently the research team did not actually digitize any 1991 maps. Also, given limits on its time and resources, they were not able to enforce topological consistency for the 2001 polygons.

Meiyappan et al. (2018) further caution that

> For four states in northeast India (Arunachal Pradesh, Nagaland, Manipur, and Meghalaya), the authors used sub-district/taluka level boundaries because village/-town level cadastral maps were difficult to acquire and/or unavailable.

What this means is that in fact there are no settlement-level boundaries provided for Arunachal Pradesh, Nagaland, Manipur, and Meghalaya in the Meiyappan et al. (2018) collection.

Read_SEDACVillages_2001.R The program reads the 2001 village and town boundaries and data (census abstracts as well as village amenities) prepared by Meiyappan et al. (2018). The data include urban as well as rural settlements, although a large block of the attributes data pertains only to villages.

[1] `http://sedac.ciesin.columbia.edu/data/set/india-india-village-level-geospatial-socio-econ-1991-2001`.

Most of the data-files are shapefiles, but three come in the form of `.gdb` files. They are in a UTM projection, `EPSG 32644`.

The program is easily edited to handle the 1991 files as well—as I understand it, the spatial data for villages and towns are simply copied from the 2001 maps, but 1991 PCA and so-called village amenity attributes are presumably linked in. So far the program has only been run on the 2001 files.

One of the 2001 files holds aggregated data for seven different states and territories: Chandigarh, Delhi, Daman and Diu, Dadra and Nagar Haveli, Lakshadweep, Pondicherry, and Andaman and Nicobar Islands. The only 3 settlements covered are in Pondicherry—Ozhukarai, Pondicherry, and Karaikal municipalities; no other towns or rural villages are spatially represented. This file's data is not especially useful. Curiously, **this is not mentioned in the documentation**.

To read the `.gdb` files into R format, we first extract the directory ending with `.gdb` to a temporary directory `PathTemp` and then use this code:

```
Layers <- st_layers(dsn=list_files_with_exts(PathTemp, exts="gdb"))
Data <-   st_read(dsn=list_files_with_exts(PathTemp, exts="gdb"),
layer=Layers$name)
```

I think that the `SID` variable is the standard state code used in other programs and datasets. There are usually 213 attribute variables.

A few odd things noticed in an initial pass through the files:

- `STCode 02` (Himachal Pradesh) has 3638 NA values on the state identifier `SID`. None of these has any attribute data;
- `STCode 14` (Manipur) has 218 variables instead of the usual 213;
- `STCode 15` (Mizoram) has 214 variables.
- All three `.gdb` files (AR, ML, NL) contain 216 variables.
- The data for Chhattisgarh contain no statutory or census towns.

For now the R-format data are placed in a temporary holding-bin for further processing.

Directory with program: `/India/Programs/Spatial/Meiyappan/Programs`

Input files: Zipped files in `/Programs/Spatial/Meiyappan/Data/SEDAC_Villages_1991_2001/2001`

Output file: The `.RData` files are stored in `/Programs/Spatial/Meiyappan/Data/SEDAC_Villages_1991_2001/R_Data_Temp`, named as `State_yyyy_xx.RData` where "yyyy" indicates the year "xx" is the `STCode` of the state. The file `State_2001_99.RData` holds aggregated, tehsil-level data and boundaries, but except for three municipalities in Pondicherry that were mentioned above, the file contains no other settlement-level information.

Program last changed: 13 August 2018

Describe_SEDACVillages_2001.R This program converts the data to EPSG 4326 and creates our standard identifiers `STCode`, `DTCode`, `SDTCode`, and `TVCode` using the `C_CODE01` parsed according to the 2001 identifier scheme. It also checks and corrects the spatial validity of the polygons, using the *lwgeom* package and the `st_make_valid()` function.

I will leave for later—at the point when we try to merge the 2001 PCAs with these spatial data—the task of "unioning" cases of multiple polygons per settlement identifier.

The Meiyappan et al. (2018) data are not especially clean. The TRU variable contains many peculiar codes that appear to have crept in from other variables. The LEVEL variable identifies some spatial features as being higher-order administrative units (e.g., tehsils, districts, and so on). These non-settlement administrative units are dropped when LEVEL equals one of the following:

```
AdminNames <- c("DISTRICT", "CIRCLE", "MANDAL", "POLICE
    STATION",
"SUB DIVISION", "TALUK", "TEHSIL", "C.D.BLOCK", "R.D.BLOCK",
"DEVELOPMENT BLOCK", "COMMUNUE PANCHAYAT")
```

Very similar code was used in processing the 2001 PCAs.

When it is present, the key census code composite identifier C_CODE01 almost always has 16 digits. The exceptions are: 1 case in Bihar for which the PCAs allowed us to replace C_CODE01 entirely, 5 cases in West Bengal and 2 in Andhra Pradesh in which C_CODE01 needed a trailing "0" to be added.

However, C_CODE01 can be NA. It looks as if the absence of the 2001 census code C_CODE01 generally implies that the feature is an uninhabited forest, a lake, some other uninhabited natural area, or that the nature of the data is unknown. But this is not always the case, as evident in Himachal Pradesh (STCode 02) in which C_CODE01 can be missing and yet there exist research-team-derived codes for state, district, subdistrict, and settlement, with the population variable TOT_P being positive.

Based on some checking, the following code is applied to assign TRUE to a newly created ProbableNaturalFeature variable:

```
X$ProbableNaturalFeature <- with(X, is.na(C_CODE01) & (is.na(TOT_P) | TOT_P==0L
    ) )
```

Admittedly, this code does not handle all situations found in the data. We have quite a few cases remaining in which is.na(TVCode) is TRUE but ProbableNaturalFeature is FALSE. A cross-tabulation shows the numbers of such cases.

Directory with program: /India/Programs/Spatial/Meiyappan/Programs

Input files: Files in /Programs/Spatial/Meiyappan/Data/SEDAC_Villages_1991_2001/R_Data_Temp As noted above, the file State_2001_99.RData holds aggregated, tehsil-level data and boundaries, but except for three municipalities in Pondicherry that were mentioned above, the file contains no other settlement-level information.

Output file: For the states with usable information, files in /Programs/Spatial/Meiyappan/Data/SEDAC_Villages_1991_2001/R_Data

Program last changed: 24 February 2019

Describe_SEDACVillages_Part2_2001.R This program make multi-polygon records from the union of separate single-polygon records with the same identifier codes (including NAME and population TOT_P). I eliminated cases lacking our standard administrative identifiers (STCode, DTCode, SDTCode, TVCode) derived from the census variable C_CODE01. I consider augmenting these identifiers with NAME and TOT_P to be extra-sure, but found that the

researchers had entered variants on the settlement name for a given set of administrative identifers and population (TOT_P). (This may have been an effort to distinguish hamlets within revenue villages, or something of the sort.) In the end, I added only TOT_P to the admin identifiers define duplicates for spatial unions.

Some states (especially STCode 08, 09) have thousands of duplicates needing a spatial union, whereas others have relatively few.

The essence of the code uses *sf*-enhanced *dplyr* functions:

```
# Make the duplicates into MULTIPOLYGON features:
Test <- data.frame( subset(X_DT, subset=Duplicates, select=c(IDVars, "
    geometry")) )
Tst <- st_as_sf(Test, sf_column_name="geometry")

# Use the dplyr approach, enhanced for sf objects, to make spatial
    unions of the
# attribute-duplicated points. There is an option in summarize() to
    simply combine
# the points rather than unioning them.
Y <- Tst %>% select(STCode, DTCode, SDTCode, TVCode, TOT_P) %>%
group_by(STCode, DTCode, SDTCode, TVCode, TOT_P) %>% summarize()
cat("Are all unioned settlements spatially valid?", "\n")
print( all(st_is_valid(Y)) )
```

Directory with program: /India/Programs/Spatial/Meiyappan/Programs

Input files: Files in /Programs/Spatial/Meiyappan/Data/SEDAC_Villages_1991_2001/R_Data

Output file: For the states with usable information, files in /Programs/Spatial/Meiyappan/R_Data_Cleaned

Program last changed: 16 March 2019

8.2 University of Tokyo settlement point shapefiles for 2001

These data were downloaded on 24 December 2018, from the website: http://india.csis.u-tokyo.ac.jp/default/download. The site—it seems to have been intended to serve as kind of gazetteer for Indian villages, but is now only partly operational—provides the following information about the origins of the data:

> Place names and their positions are based on the GIS information on the 2001 Census, supplied by the courtesy of ML Infomap (http://www.mlinfomap.com/). We greatly appreciate Dr. Manosi Lahiri for her support to us.

There is a variable indicating the vintage of the paper map by which the village was located. Evidently a serious attempt was made to pinpoint "hamlets," which are the small settlements that are aggregated into revenue villages.

It seems that the objective of the mapping effort was to establish locations of villages as far back as the British colonial period, using old paper maps when available. (As just mentioned, the vintage of the map is coded in the data.) It may be, therefore, that some locations refer to villages and hamlets that no longer exist. Also, the village names extracted from historical maps may have changed beyond all recognition.

The suggested attribution is as follows: Mizushima Laboratory (2013), "India Place Finder", `http://india.csis.u-tokyo.ac.jp/`, Department of Oriental History, Graduate School of Humanities and Sociology, The University of Tokyo. The professor who headed the effort is Tsukasa Mizushima, Professor, Department of the Orient History, The University of Tokyo. E-mail: `zushima2010@gmail.com`. His on-line vita lists several publications and presentations that use GIS techniques in historical analyses of the colonial period, but nothing is evident on modern India; see `http://www.l.u-tokyo.ac.jp/~zushima9/profile_e.html`.

Although ML Infomap is cited as a source of these data, the village coordinates can differ significantly from our collection of 2001 settlement points from ML Infomap (which are not redistributable). My impression so far is that the University of Tokyo points are much better situated, to judge from their agreement with present-day OPENSTREETMAP imagery.

Read_UTokyo_Villages_2001.R The program simply unzips, reads, and transforms the data to R's *sf* format. Apart from the settlement coordinate, there is not much here on each village and hamlet other than its name. In particular, there are no locational administrative unit identifiers, which would have to be gleaned from a spatial overlay of the points on district and sub-district boundaries.

Directory with program: `Country_Data/India/Programs`

Input files: `/Boundaries_Other/U_Tokyo_Habitations_2001/Hamlets_all.zip`

Output file: The *sf*-format file is in `/R_Data/VillagePoints_UTokyo_2001.RData`

Program last changed: 24 December 2018

8.3 DATAMEET settlement polygons, selected states

These settlement boundaries are provided in `.geojson` files assembled by the remarkable DATAMEET group and downloaded from `http://projects.datameet.org/indian_village_boundaries`. They are available for the states of Bihar, Goa, Gujarat, Karnataka, Kerala, Maharashtra, Rajasthan, and Sikkim.[2]

The DATAMEET collection is mainly specific to 2001 and helpful in filling in the record for that year for the states of Bihar, Goa, Gujarat, Kerala, Maharashtra, and Sikkim. Except for Rajasthan (the newest addition to the DATAMEET collection) and Karnataka (whose boundaries are for 1991), the spatial files contain 2001 boundaries and identifier codes that are generally compatible with the 2001 PCA identifiers. The year-2001 focus is explicit and unambiguous for some states; for others, the documentation can be a little vague on the vintage of the maps that were digitized, although 2001 still seems like the safest guess. The spatial files are accompanied by `.csv` or semicolon-separated `.txt` files that link the 2001 to the 2011 village identifiers.

Karnataka is the exception: its text file provides 1991–to–2001 links and the documentation makes it clear that the boundaries refer to 1991. Since we have very little other spatial data for 1991 and are not yet very far along in coding the 1991 PCAs and related non-spatial data, there is little reason to process the 1991 Karnataka data right now. I am therefore leaving these data in a half-finished status. Rajasthan has only 2011 codes and presumably also contains 2011

[2]The suggested citation for the DATAMEET village data is: "Url: `http://projects.datameet.org/indian_village_boundaries/` Attribution: Villages Maps Provided by Indian Village Boundaries Project [`http://projects.datameet.org/indian_village_boundaries/`] by Data{Meet}. It's made available under the Open Database License (ODbL)[`http://opendatacommons.org/licenses/odbl/`]."

(or possibly even more recent) boundaries. The GITHUB repository for Rajasthan references a new government web-site, `https://nad.ncog.gov.in/login`, the National Asset Directory, whose user interface allows display of maps down to the village level. Since we already have ML Infomap boundaries for 2011 and are also exploring the Survey of India 2011 boundaries newly-issued via Justin Elliot Meyers' GITHUB, the Rajasthan file is a low priority at present.

Read_Datameet_Villages.R This program examines the `.txt` and `.csv` files, leaving the spatial data for the next program.

I check the identifiers in these text files against our PCA-derived state, district, and subdistrict codes. A number of these files include a `CEN_2001` variable that purports to concatenate all 2001 identifiers. Done correctly, such a composite identifier would have 16 characters, but the `CEN_2001` variable often contains either fewer or more characters than that. There is more discussion of this issue below with reference to the `.geojson` spatial files. Among other things, the following code listing provides references to the state-specific documentation provided by DATAMEET:

```
# Bihar documentation is here:
#  http://projects.datameet.org/indian_village_boundaries/br/
# and references the 2011 Administrative Atlas among the sources.

# Goa---all CEN_2001 cases have 16 characters

# Gujarat: Documentation is http://projects.datameet.org/indian_village_boundaries/gj/
# The documentation references the 2011 census and the Gujarat revenue
# department maps (pdfs, but lacking clear dates of creation)

# Kerala documentation is here:
#    http://projects.datameet.org/indian_village_boundaries/kl/
# It references a 2010 Delimitation Map, whose pdfs at http://delimitation.lsgkerala.gov.in/map.

# Maharashtra: http://projects.datameet.org/indian_village_boundaries/mh/
# which references 2011 census sources. However, the DataMeet project link leads
# to what is apparently the shapefile source:
#    http://mrsac.gov.in/en/projects/high-resolution-data-base-mapping/geo-referencing-village-map-
    project-gvmp
# whose dates of creation are not obvious. The data appear to have been cobbled together from
# multiple sources.

# The Odisha documentation is here:
#      http://projects.datameet.org/indian_village_boundaries/or/
# There is no clear indication as to the source and vintage of the maps.

# Sikkim documentation is:
#       http://projects.datameet.org/indian_village_boundaries/sk/
# It references the 2011 Administrative Atlas.

# Karnataka: http://projects.datameet.org/indian_village_boundaries/ka/
# The DataMeet documentation page indicates that the boundaries were digitized
# from the *1991* DCHB maps.

# The data have a lot of potentially helpful identifiers for both
# 2001 and 1991, but there are no 2011 identifiers. For 2001, the following
# qualify as identifiers based on their values and length:
#     VILL_NAME, DIST_CODE, DISTRICT_N, THSIL_CODE, V_CT_CODE
# There are 862 blank records on these identifiers.

# I believe LOC_CODE is for 1991, but will not attempt to define 1991 identifiers
# for now.

# Rajasthan .csv file: I have edited the original file to fix its header, which should be:
# tvcode_2011,village_name_2011, state_code_2011, district_code_2011, sub_district_code_2011,
# tvcode_2001, village_name_2001, state_code_2001, district_code_2001, sub_district_code_2001,
# CEN_2011
```

Directory with program: `/India/Programs/Spatial/DataMeet/Programs`

Input files: The zip-file with the full collection (as of March 2019) is `/Programs/Spatial/DataMeet/Data/india_village_files_master.zip`. Also a `RajaRevised.txt` file which corrects the header on the Rajasthan file.

Output files: In the `/Programs/Spatial/DataMeet/R_Data_Temp` directory for now. Karnataka (with its 1991-era boundaries) is not included.

Program last changed: 27 May 2020

Read_Datameet_Villages_Spatial.R This program examines the `.geojson` data-files for the priority states (which for now exclude Karnataka and Rajasthan). See code extract above for links to the documentation.

Sikkim: `CEN_2001` looks more reliable than the district and subdistrict names provided in the `.geojson` data. However, all of its values are shorter than 16 in length: the lengths of `CEN_2001` are 8, 11, 12, and 13. As it turns out, all these except the length-8 case can be solved by left-padding with zeroes.

One case—whose `CEN_2001` is length 8—has no settlement name or entries in `CEN_2001` from which a `TVCode` can be derived. It might be a displaced polygon for Namchi. I'm leaving it in for now.

Maharashtra: The spatial files `mh1.geojson` and `mh2.geojson` have many cases in which `CEN_2001` is too short, as well as thousands of cases in which it is too long, a few having length 17 and many more with 18-character values. Upon investigating, I've found that the length-18 cases come from concatenating *state, district, and subdistrict codes for 2011* in one clump, and then appending to that the *2001 version of the `TVCode`*. See the program for details on how all this is sorted out. Fortunately, since the 2001 `TVCode` is unique within states, we can recover the 2001 administrative codes through the PCAs.

Odisha: In addition to `CEN_2001` codes of length 16, there are cases of length 2 and 8; see the program. The length-8 cases repeat the name of the subdistrict in what should be the variable for village name. Quite a few cases do not match the PCAs when the district and subdistrict identifiers are used, so I matched on the state and `TVCode` (successfully, given the uniqueness of 2001 `TVCodes` within state) and then assigned the PCA versions of district and subdistrict codes

Goa: Entirely without problems.

Gujarat: The file includes a polygon for Dadra, which might come in handy at some point, but I have excluded it for now. For four cases, the subdistrict code needed editing to conform with the version in the PCAs.

Bihar: A number of (resolvable) problems; see the program for details. One point deserves particular mention: In matching with the PCAs, we find quite a few cases involving a village with "Part in xx" in its name, with only 1 `TVCode` in the PCAs (it is supposed to be unique, after all) but with a discrepancy between the `SDTCode` of the PCA records and the "xx" of the village name in the PCA record. (The "xx" seems to refer to subdistrict.) The mis-match raises concerns about whether the PCAs actually have the correct subdistrict identifier; but this hinges on the meaning of "Part in xx"—should this be interpreted as the subdistrict of the settlement in 2001 or, possibly, the subdistrict where the settlement was formerly located before a pre-2001 change of subdistrict boundaries? Unclear. I think we have to stick with the PCA version of the subdistrict codes.

Kerala: Except for one Mahe polygon for Pondicherry (which was dropped), no problems are evident here.

Directory with program: `/India/Programs/Spatial/DataMeet/Programs`

Input files: The zip-file with the full collection (as of March 2019) is `/Programs/Spatial/DataMeet/Data/india_village_files_master.zip`. Also, `CensusAbstractsPlus_2001.`

`RData` in `/R_Data`.

Output files: In `/Programs/Spatial/DataMeet/R_Data_Temp` directory for now.

Program last changed: 29 May 2020