**Village Level Dataset Schema**

**Introduction**

Development of countries depends on success or failure policies framed by the government. Policies are framed based on census data collected which provide a sense of problems areas that need to be addressed. So, framing of good policies is dependent on how good and accurate the collected census data is. However, conducting census data in a country like ours is a very expensive and time-consuming process. Hence it is conducted every 10 years. However, we need more real time data to proactively shape our policies. So, we are exploring the idea of using satellite image data to get a measure or index on development to track the development of villages over different years. Satellite data is cost effective and available in real time. The index is built on indicators like Asset Ownership, Bathroom Facility, Fuel of Cooking, Minimum Source of Water, and Literacy. These indicators were computed for the year 2011 based on census data and CNN models were trained to correlate village satellite image with these indicators. We have predicted indicators for the years 2001 and 2011.

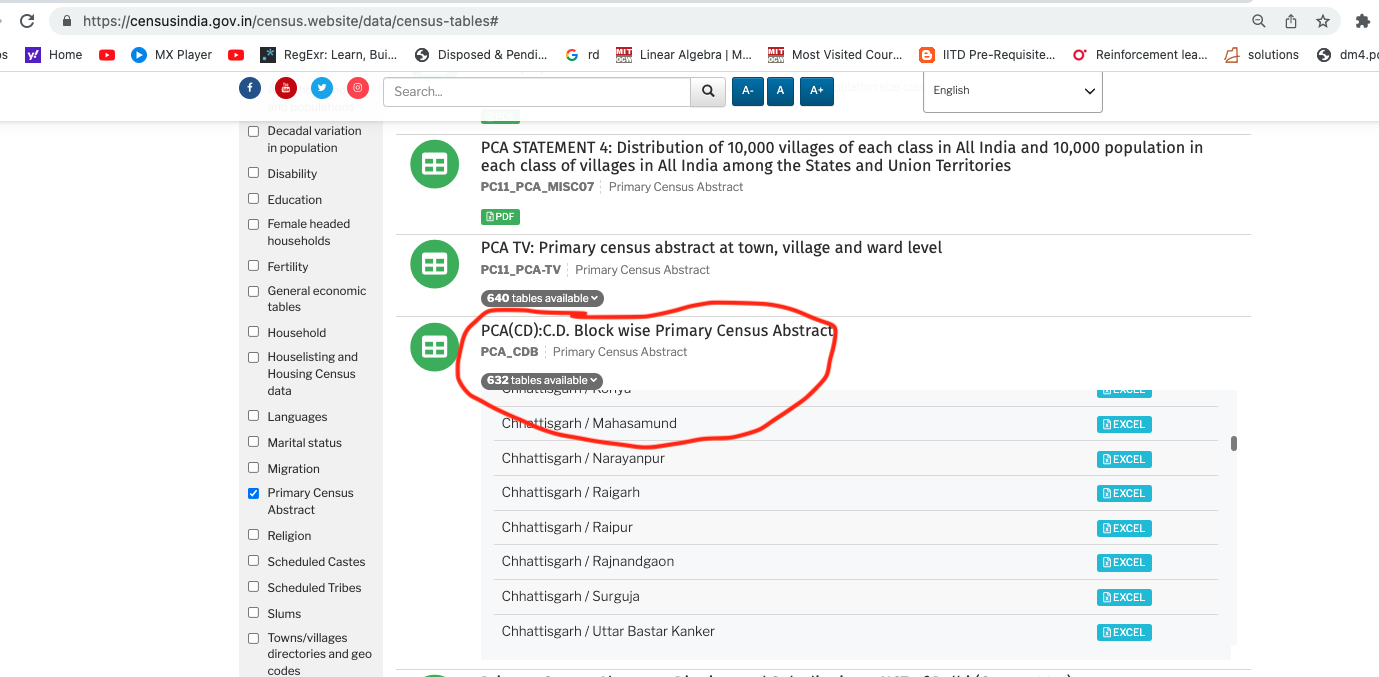
**To access Github Repository for this project click** [**here**](https://github.com/amangupt01/Village_Development_Model)

**Section 1: Raw Census Data For 2011**

**Table 1.1: Primary Abstract Data (select year 2011):**

[**https://censusindia.gov.in/census.website/data/census-tables#**](https://censusindia.gov.in/census.website/data/census-tables)

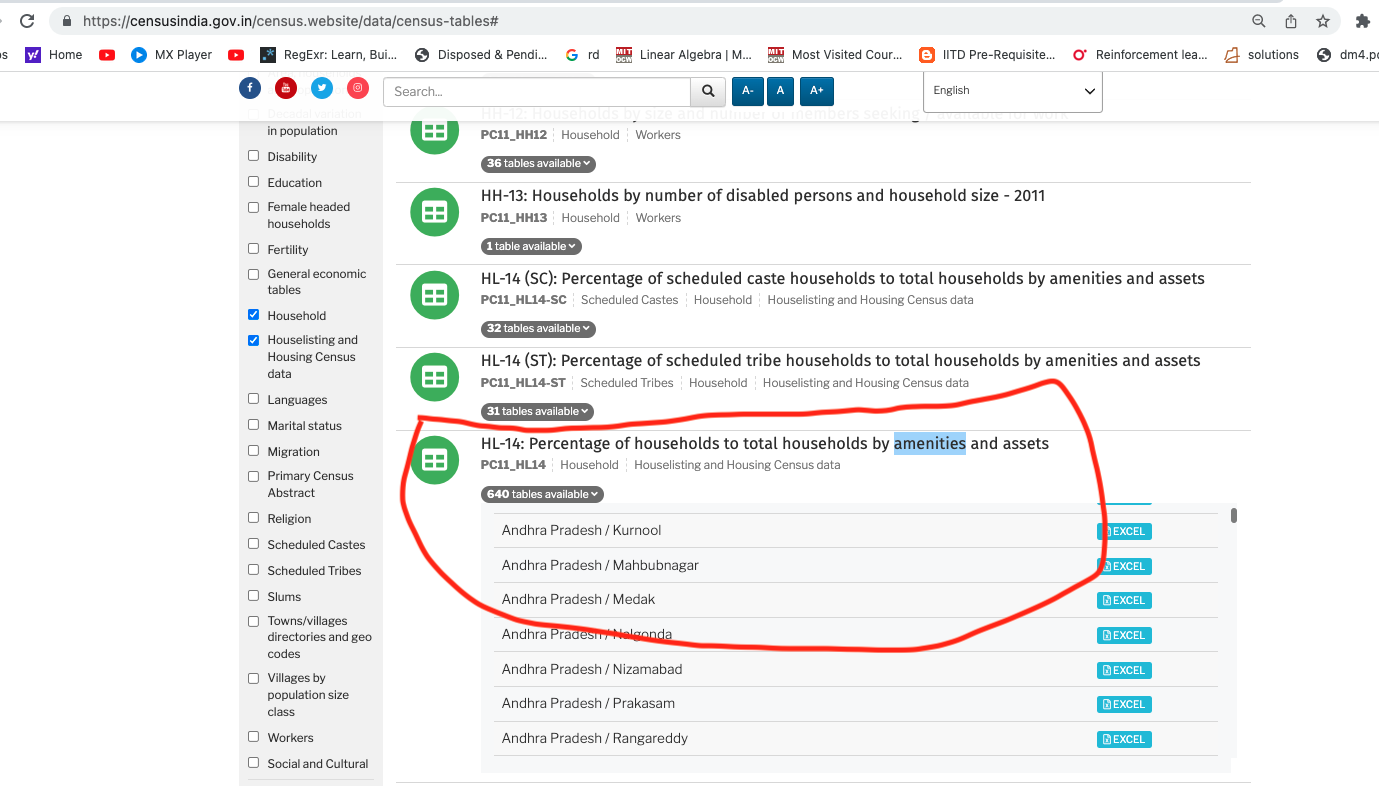
Select the year 2011 and download the raw dataset from PCA(CD): C.D. Block-wise primary census abstract (PCA\_CBD)

****

**Table 1.2: Household Data (select year 2011):**

[**https://censusindia.gov.in/census.website/data/census-tables#**](https://censusindia.gov.in/census.website/data/census-tables)

Select the year 2011 and download the raw dataset from HL-14: Percentage of households to total households by amenities and assets (PC11\_HL14)

****

We combined the data from the above two output files and considered only rows which have Village Level -> ‘Rural’. From this combined data, below are the specific columns that we have extracted.

**Table 1.3: Village\_level\_population\_enumeration\_2011.csv**

**[**[**Link**](https://drive.google.com/file/d/16MqXDDiDD4qqUXAkHZiekWQq2OEdewnM/view?usp=drive_link)**]**

This is the population enumeration data for the year 2011 at village level in India.

| **S.no** | **Column Name** | **Data Type** | **Description** |
| --- | --- | --- | --- |
| 1 | STATE | alphanumeric | State census id |
| 2 | DISTRICT | integer | District census id |
| 3 | SUB-DISTT | integer | Sub-District census id |
| 4 | TOWN/VILLAGE | integer | Town/village census id |
| 5 | WARD | integer | Ward census id |
| 6 | EB | integer | Enumeration block census id |
| 7 | LEVEL | string | Whether it is a district/mandal/village |
| 8 | NAME | string | Name of the village |
| 9 | TRU | Total/rural/urban | Description of the count if it is total population or rural population or urban |
| 10 | No\_HH | integer | Total number of households |
| 11 | TOT\_P | integer | Total population |
| 12 | TOT\_M | integer | Total male population |
| 13 | TOT\_F | integer | Total female population |
| 14 | P\_06 | integer | Total population in age group 0-6 |
| 15 | M\_06 | integer | Total male in age group 0-6 |
| 16 | F\_06 | integer | Total female in age group 0-6 |
| 17 | P\_SC | integer | Total schedule caste population |
| 18 | M\_SC | integer | Total schedule caste male |
| 19 | F\_SC | integer | Total schedule caste female |
| 20 | P\_ST | integer | Total schedule tribe population |
| 21 | M\_ST | integer | Total schedule tribe male |
| 22 | F\_ST | integer | Total schedule tribe female |
| 23 | P\_LIT | integer | Total literate population |
| 24 | M\_LIT | integer | Total literate male |
| 25 | F\_LIT | integer | Total literate female |
| 26 | P\_ILL | integer | Total illiterate population |
| 27 | M\_ILL | integer | Total illiterate male |
| 28 | F\_ILL | integer | Total illiterate female |
| 29 | TOT\_WORK\_P | integer | Total working population |
| 30 | TOT\_WORK\_M | integer | Total working male |
| 31 | TOT\_WORK\_F | integer | Total working female |
| 32 | MAINWORK\_P | integer | Total main worker population |
| 33 | MAINWORK\_M | integer | Total main worker male |
| 34 | MAINWORK\_F | integer | Total main worker female |
| 35 | MAIN\_CL\_P | integer | Total main cultivator population |
| 36 | MAIN\_CL\_M | integer | Total main cultivator male |
| 37 | MAIN\_CL\_F | integer | Total main cultivator female |
| 38 | MAIN\_AL\_P | integer | Total main agricultural laborer population |
| 39 | MAIN\_AL\_M | integer | Total main agricultural laborer male |
| 40 | MAIN\_AL\_F | integer | Total main agricultural laborer female |
| 41 | MAIN\_HH\_P | integer | Total main household industry population |
| 42 | MAIN\_HH\_M | integer | Total main household industry male |
| 43 | MAIN\_HH\_F | integer | Total main household industry female |
| 44 | MAIN\_OT\_P | integer | Total main other work population |
| 45 | MAIN\_OT\_M | integer | Total main other work male |
| 46 | MAIN\_OT\_F | integer | Total main other work female |
| 47 | MARGWORK\_P | integer | Total marginal worker population |
| 48 | MARGWORK\_M | integer | Total marginal worker male |
| 49 | MARGWORK\_F | integer | Total marginal worker female |
| 50 | MARG \_CL\_P | integer | Total marginal cultivator population |
| 51 | MARG\_CL\_M | integer | Total marginal cultivator male |
| 52 | MARG\_CL\_F | integer | Total marginal cultivator female |
| 53 | MARG\_AL\_P | integer | Total marginal agricultural laborer population |
| 54 | MARG\_AL\_M | integer | Total marginal agricultural laborer male |
| 55 | MARG\_AL\_F | integer | Total marginal agricultural laborer female |
| 56 | MARG\_HH\_P | integer | Total marginal household industry population |
| 57 | MARG\_HH\_M | integer | Total marginal household industry male |
| 58 | MARG\_HH\_F | integer | Total marginal household industry female |
| 59 | MARG\_OT\_P | integer | Total marginal other work population |
| 60 | MARG\_OT\_M | integer | Total marginal other work male |
| 61 | MARG\_OT\_F | integer | Total marginal other work female |
| 62 | MARGWORK\_3\_6\_P | integer | Total marginal worker population of age 3-6 |
| 63 | MARGWORK\_3\_6\_M | integer | Total marginal male population of age 3-6 |
| 64 | MARGWORK\_3\_6\_F | integer | Total marginal female population of age 3-6 |
| 65 | MARG\_CL\_3\_6\_P | integer | Total marginal cultivator population of age 3-6 |
| 66 | MARG\_CL\_3\_6\_M | integer | Total marginal cultivator male of age 3-6 |
| 67 | MARG\_CL\_3\_6\_F | integer | Total marginal cultivator female of age 3-6 |
| 68 | MARG\_AL\_3\_6\_P | integer | Total marginal agricultural labourer population of age 3-6 |
| 69 | MARG\_AL\_3\_6\_M | integer | Total marginal agricultural labourer male of age 3-6 |
| 70 | MARG\_AL\_3\_6\_F | integer | Total marginal agricultural labourer female of age 3-6 |
| 71 | MARG\_HH\_3\_6\_P | integer | Total marginal household industry population of age 3-6 |
| 72 | MARG\_HH\_3\_6\_M | integer | Total marginal household industry male of age 3-6 |
| 73 | MARG\_HH\_3\_6\_F | integer | Total marginal household industry female of age 3-6 |
| 74 | MARG\_OT\_3\_6\_P | integer | Total marginal other work population of age 3-6 |
| 75 | MARG\_OT\_3\_6\_M | integer | Total marginal other work male of age 3-6 |
| 76 | MARG\_OT\_3\_6\_F | integer | Total marginal other work female of age 3-6 |
| 77 | MARGWORK\_0\_3\_P | integer | Total marginal worker population of age 0-3 |
| 78 | MARGWORK\_0\_3\_M | integer | Total marginal male population of age 0-3 |
| 79 | MARGWORK\_0\_3\_F | integer | Total marginal female population of age 0-3 |
| 80 | MARG\_CL\_0\_3\_P | integer | Total marginal cultivator population of age 0-3 |
| 81 | MARG\_CL\_0\_3\_M | integer | Total marginal cultivator male of age 0-3 |
| 82 | MARG\_CL\_0\_3\_F | integer | Total marginal cultivator female of age 0-3 |
| 83 | MARG\_AL\_0\_3\_P | integer | Total marginal agricultural labourer population of age 0-3 |
| 84 | MARG\_AL\_0\_3\_M | integer | Total marginal agricultural labourer male of age 0-3 |
| 85 | MARG\_AL\_0\_3\_F | integer | Total marginal agricultural labourer female of age 0-3 |
| 86 | MARG\_HH\_0\_3\_P | integer | Total marginal household industry population of age 0-3 |
| 87 | MARG\_HH\_0\_3\_M | integer | Total marginal household industry male of age 0-3 |
| 88 | MARG\_HH\_0\_3\_F | integer | Total marginal household industry female of age 0-3 |
| 89 | MARG\_OT\_0\_3\_P | integer | Total marginal other work population of age 0-3 |
| 90 | MARG\_OT\_0\_3\_M | integer | Total marginal other work male of age 0-3 |
| 91 | MARG\_OT\_0\_3\_F | integer | Total marginal other work female of age 0-3 |
| 92 | NON\_WORK\_P | integer | Total non-working population |
| 93 | NON\_WORK\_M | integer | Total non-working males |
| 94 | NON\_WORK\_F | integer | Total non-working females |

**Table 1.4: Village\_level\_data\_2001\_raw.csv**

**[**[**Link**](https://drive.google.com/file/d/1G-8mTXUrHWoA79rHeOEeiYPThcR8cm1o/view?usp=drive_link)**]**

**This data was acquired from Priyamvada Trivedi and includes different socio-economic indicators at village level for the year 2001.**

| S.No. | FIELD NAME | DATA TYPE | WIDTH | REMARKS |
| --- | --- | --- | --- | --- |
| 1 | ST\_CODE | Character | 2 | STATE CODE |
| 2 | DIST\_CODE | Character | 2 | DISTRICT CODE |
| 3 | THSIL\_CODE | Character | 4 | TAHSIL/TALUK CODE |
| 4 | BLOCK\_CODE | Character | 4 | C.D. BLOCK CODE |
| 5 | V\_CT\_CODE | Character | 8 | VILLAGE CODE |
| 6 | VILL\_NAME | Character | 40 | VILLAGE NAME |
| 7 | AREA | Numeric | 5 | Area of Village (in hectares) |
| Population data based on 2001 Census | | | | |
| 8 | T\_HH | Numeric | 6 | Number of Households |
| 9 | T\_P | Numeric | 8 | Total population- Persons |
| 10 | T\_M | Numeric | 7 | Total population- Males |
| 11 | T\_F | Numeric | 7 | Total population- Females |
| 12 | SC\_P | Numeric | 7 | Scheduled Castes population- Persons |
| 13 | SC\_M | Numeric | 7 | Scheduled Castes population- Males |
| 14 | SC\_F | Numeric | 7 | Scheduled Castes population- Females |
| 15 | ST\_P | Numeric | 7 | Scheduled Tribes population- Persons |
| 16 | ST\_M | Numeric | 7 | Scheduled Tribes population- Males |
| 17 | ST\_F | Numeric | 7 | Scheduled Tribes population- Females |
| Amenities data | | | | |
| 18 | EDU\_FAC | Numeric | 1 | Educational facilities (A/NA) |
| 19 | P\_SCH | Numeric | 2 | Number of Primary School |
| 20 | RANG\_P\_SCH | Numeric | 1 | If not available, Provide the Range Code |
| 21 | M\_SCH | Numeric | 1 | Number of Middle School |
| 22 | RANG\_M\_SCH | Numeric | 1 | If not available, Provide the Range Code |
| 23 | S\_SCH | Numeric | 1 | Number of Secondary School |
| 24 | S\_S\_SCH | Numeric | 1 | Number of Senior Secondary School |
| 25 | COLLEGE | Numeric | 1 | Number of Collage |
| 26 | RANG\_COLL | Numeric | 1 | If not available, Provide the Range Code |
| 27 | IND\_SCH | Numeric | 1 | Number of Industrial School |
| 28 | TR\_SCH | Numeric | 1 | Number of Taining School |
| 29 | ADLT\_LT\_CT | Numeric | 1 | Number of Adult literacy Class/Centre |
| 30 | OTH\_SCH | Numeric | 1 | Number of Other educational facilities |
| 31 | MEDI\_FAC | Numeric | 1 | Medical facilities (A/NA) |
|  |  |  |  | If not available, Range Code is to be provided for Allopathic hospital, Maternity and Child Welfare Centre and Primary Health Centre |
| 32 | ALL\_HOSP | Numeric | 1 | Number of Allopathic Hospital |
| 33 | RANG\_ALL | Numeric | 1 | If not available, Provide the Range Code |
| 34 | AYU\_HOSP | Numeric | 1 | Number of Ayurvedic Hospital |
| 35 | UN\_HOSP | Numeric | 1 | Number of Unani Hospital |
| 36 | HOM\_HOSP | Numeric | 1 | Number of Homeopathic Hospital |
| 37 | ALL\_DISP | Numeric | 1 | Number of Allopathic Dispensary |
| 38 | AYU\_DISP | Numeric | 1 | Number of Ayurvedic Dispensary |
| 39 | UN\_DISP | Numeric | 1 | Number of Unani Dispensary |
| 40 | HOM\_DISP | Numeric | 1 | Number of Homeopathic Dispensary |
| 41 | MCW\_CNTR | Numeric | 1 | Number of Maternity and Child Welfare Centre |
| 42 | RANG\_MCW | Numeric | 1 | If not available, Provide the Range Code |
| 43 | M\_HOME | Numeric | 1 | Number of Maternity Home |
| 44 | CWC | Numeric | 1 | Number of Child Welfare Centre |
| 45 | H\_CNTR | Numeric | 1 | Number of Health Centre |
| 46 | PH\_CNTR | Numeric | 1 | Number of Primary Health Centre |
| 47 | RANG\_PHC | Numeric | 1 | If not available, Provide the Range Code |
| 48 | PHS\_CNT | Numeric | 1 | Number of Primary Health Sub Centre |
| 49 | FWC\_CNTR | Numeric | 1 | Number of Family Welfare Centre |
| 50 | TB\_CLN | Numeric | 1 | Number of T.B. Clinic |
| 51 | N\_HOME | Numeric | 1 | Number of Nursing Home |
| 52 | RMP | Numeric | 2 | Number of Registered Private Medical Practiotioners |
| 53 | SMP | Numeric | 2 | Number of Subsidised Medical Practitioners |
| 54 | CHW | Numeric | 2 | Number of Community Health workers |
| 55 | OTH\_CNTR | Numeric | 1 | Number of Other medical facilities |
| 56 | DRNK\_WAT\_F | Numeric | 1 | Drinking Water facility (A/NA) |
| 57 | RANG\_WAT\_F | Numeric | 1 | If not available, Provide the Range Code |
| 58 | TAP | Numeric | 1 | Tap Water (T) |
| 59 | WELL | Numeric | 1 | Well Water (W) |
| 60 | TANK | Numeric | 1 | Tank Water (TK) |
| 61 | TUBEWELL | Numeric | 1 | Tubewell Water (TW) |
| 62 | HANDPUMP | Numeric | 1 | Handpumb (HP) |
| 63 | RIVER | Numeric | 1 | River Water(R) |
| 64 | CANAL | Numeric | 1 | Canal (C) |
| 65 | LAKE | Numeric | 1 | Lake (L) |
| 66 | SPRING | Numeric | 1 | Spring (S) |
| 67 | OTHER | Numeric | 1 | Other drinking water sources (O) |
| 68 | SOU\_SUMM | Character | 2 | Source of Drinking Water during Summer (indicate code from above) |
| 69 | RANG\_SS | Numeric | 1 | If not available, Provide the Range Code |
| 70 | SS\_CODE | Character | 2 | Source code from above as applicable |
| 71 | P\_T\_FAC | Numeric | 1 | Post, Telegraph and Telephone facilities (A/NA), if available |
| 72 | POST\_OFF | Numeric | 1 | Number of Post Office |
| 73 | RANG\_PO | Numeric | 1 | If not available, Provide the Range Code |
| 74 | TELE\_OFF | Numeric | 1 | Number of Telegraph Office |
| 75 | POST\_TELE | Numeric | 1 | Number of Post and Telegraph Office |
| 76 | PHONE | Numeric | 3 | Number of Telephone connections |
| 77 | RANG\_PHONE | Numeric | 1 | If not available, Provide the Range Code |
| 78 | COMM\_FAC | Numeric | 1 | Communication (Y/N) |
| 79 | BS\_FAC | Numeric | 1 | Bus services |
| 80 | RANG\_BS | Numeric | 1 | If not available, Provide the Range Code |
| 81 | RS\_FAC | Numeric | 1 | Railways services |
| 82 | RANG\_RS | Numeric | 1 | If not available, Provide the Range Code |
| 83 | NW\_FAC | Numeric | 1 | Navigable water way including River, Canal etc. |
| 84 | RANG\_NW | Numeric | 1 | If not available, Provide the Range Code |
| 85 | BANK\_FAC | Numeric | 1 | Banking facility (Y/N) |
| 86 | COMM\_BANK | Numeric | 1 | Number of Commercial Bank |
| 87 | RANG\_COMM | Numeric | 1 | If not available, Provide the Range Code |
| 88 | COOP\_BANK | Numeric | 1 | Number of Co-operative Commercial Bank |
| 89 | RANG\_COOP | Numeric | 1 | If not available, Provide the Range Code |
| 90 | CRSOC\_FAC | Numeric | 1 | Credit Societies (Y/N) |
| 91 | AC\_SOC | Numeric | 1 | Number of Agricultural Credit Societies |
| 92 | RANG\_ACS | Numeric | 1 | If not available, Provide the Range Code |
| 93 | NAC\_SOC | Numeric | 1 | Number of Non Agricultural Credit Societies |
| 94 | RANG\_NAC | Numeric | 1 | If not available, Provide the Range Code |
| 95 | OTHER\_SOC | Numeric | 1 | Number of Other Credit Societies |
| 96 | RANG\_OTH | Numeric | 1 | If not available, Provide the Range Code |
| 97 | RC\_FAC | Numeric | 1 | Recreational and Cultural facilities (Y/N) |
| 98 | C\_V\_HALL | Numeric | 1 | Number of Cinema/Video-hall |
| 99 | RANG\_CV | Numeric | 1 | If not available, Provide the Range Code |
| 100 | SP\_CL\_FAC | Numeric | 1 | Number of Sports Club |
| 101 | RANG\_SPCL | Numeric | 1 | If not available, Provide the Range Code |
| 102 | ST\_AU\_FAC | Numeric | 1 | Number of Stadium/Auditorium |
| 103 | RANG\_STAU | Numeric | 1 | If not available, Provide the Range Code |
|  |  |  |  | Approach to Village (Y/N) |
| 104 | APP\_PR | Numeric | 1 | Approach - Paved Road |
| 105 | APP\_MR | Numeric | 1 | Approach - Mud Road |
| 106 | APP\_FP | Numeric | 1 | Approach - Foot Path |
| 107 | APP\_NAVRIV | Numeric | 1 | Approach - Navigable River |
| 108 | APP\_NAVCAN | Numeric | 1 | Approach - Navigable Canal |
| 109 | APP\_NW | Numeric | 1 | Approach - Navigable water-way other than river or canal |
| 110 | NEAR\_TOWN | Character | 15 | Nearest Town |
| 111 | DIST\_TOWN | Numeric | 3 | Distance from the nearest Town (in Kilometer(s)) |
| 112 | POWER\_SUPL | Numeric | 1 | Power supply (A/NA) |
| 113 | POWER\_DOM | Numeric | 1 | Electricity for Domestic use |
| 114 | POWER\_AGR | Numeric | 1 | Electricity of Agricultural use |
| 115 | POWER\_OTH | Numeric | 1 | Electricity of other purposes |
| 116 | POWER\_ALL | Numeric | 1 | Electricity for all purposes |
| 117 | PAP\_MAG | Numeric | 1 | Newspaper/Magazine (Y/N) |
| 118 | NEWS\_PAP | Character | 1 | News Paper (Indicate N, if arrived) |
| 119 | MAGAZINE | Character | 1 | Magazine (indicate M, if arrived) |
|  |  |  |  | Income and Expenditure of the village ( in Rs.' 00) |
| 120 | A\_INCEXP | Numeric | 1 | Separate figures available (Y/N), if Yes: |
| 121 | TOT\_INC | Numeric | 10 | Total Income |
| 122 | TOT\_EXP | Numeric | 10 | Total Expenditure |
|  |  |  |  | Most Important Commodities manufactured |
| 123 | MAN\_COMM1 | Character | 15 | Manufactured Item No. 1 |
| 124 | MAN\_COMM2 | Character | 15 | Manufactured Item No. 2 |
| 125 | MAN\_COMM3 | Character | 15 | Manufactured Item No. 3 |
|  |  |  |  | Land use i.e. area under different types (rounded upto two decimal places ) in hectares |
| 126 | LAND\_FORES | Numeric | 9.2 | Forest |
|  |  |  |  | Irrigated (by source) |
| 127 | CANAL\_GOVT | Numeric | 9.2 | Government Canal |
| 128 | CANAL\_PVT | Numeric | 9.2 | Private Canal |
| 129 | WELL\_WO\_EL | Numeric | 9.2 | Well (without electricity) |
| 130 | WELL\_W\_EL | Numeric | 9.2 | Well (with electricity) |
| 131 | TW\_WO\_EL | Numeric | 9.2 | Tube-well (without electricity) |
| 132 | TW\_W\_EL | Numeric | 9.2 | Tube-well (with electricity) |
| 133 | TANK\_IRR | Numeric | 9.2 | Tank |
| 134 | RIVER\_IRR | Numeric | 9.2 | River |
| 135 | LAKE\_IRR | Numeric | 9.2 | Lake |
| 136 | W\_FALL | Numeric | 9.2 | Waterfall |
| 137 | OTH\_IRR | Numeric | 9.2 | Others |
| 138 | TOT\_IRR | Numeric | 9.2 | Total Irrigated Area |
| 139 | UN\_IRR | Numeric | 9.2 | Unirrigated Area |
| 140 | CULT\_WASTE | Numeric | 9.2 | Culturable waste (including gauchar and groves) |
| 141 | AREA\_NA\_CU | Numeric | 9.2 | Area not available for cultivation |

| Codes used for --(Y/N) or (A/NA) are-- | (Y/N) = YES/NO (A/NA) = AVAILABLE/NOT AVAILALBLE |
| --- | --- |
| VALUE IS: 1- FOR YES OR AVAILABLE, AND |
| 2- FOR NO OR NOT AVAILABLE |
| 0- For Nil information/Not available |

| Codes used for giving distance by ranges where amenities are not available are-- | RANGE CODE 1 FOR < 5 Kms |
| --- | --- |
| RANGE CODE 2 FOR 5 TO 10 Kms |
| RANGE CODE 3 FOR > 10 Kms |

Note: In case the village is having electricity for all purposes, code '1' is given in column number 116 and code '0' is used in column number 113, 114 & 115 i.e.-- for 'Power domestic', 'Power agricultural' and 'Power for other purposes' instead of code '1'.

**Section 2: Preprocessing Census Data For 2011**

To preprocess the census data, we first divide each socio-economic indicator into 3 categories- Rudimentary, Intermediate, and Advance. The mapping from raw census to these categories is presented in Table 2.1

For each indicator, the villages are clustered on the percentage of rudimentary/intermediate/advance households they have (Table 2.2). This helps us to rank each village at level-1 (low), level-2 (medium), or level-3 (advance) against these indicators. These discretized labels are presented as groundtruth in the next section.

**Table 2.1: Census Raw data to indicator mapping for the year 2011**

This table gives the mapping between our continuous Indicators and raw census data available from above census data sources. If there are multiple columns for an indicator, then it is obtained by summing over all these columns. The output of this task is Table-1 described after this mapping table.

| **Main Indicator** | **Sub-Indicator** | **Census File** | **Column Name/ Number** |
| --- | --- | --- | --- |
| Bathroom Facility | BF\_RUD | Household | Number of households not having latrine facility within the premises  Column Number - 100 |
| BF\_INT | Household | Pit latrine    With slab/ ventilated improved pit  Column Number - 95    Without slab/ open pit  Column Number - 96 |
| BF\_ADV | Household | Flush/pour flush latrine connected to piped sewer system  Column Number - 92    Septic tank  Column Number - 93    Other system  Column Number - 94 |
| Fuel Of Cooking | FC\_RUD | Household | Firewood  Column Number - 109 |
| FC\_INT | Household | Crop Residue  Column Number - 110  Cow Dung Cake  Column Number - 111  Coal, Lignite, Charcoal  Column Number - 112  Kerosene  Column Number - 113 |
| FC\_ADV | Household | LPG/PNG  Column Number - 114    Electricity  Column Number - 115    Biogas  Column Number - 116 |
| Main Source of Water | MSW\_RUD | Household | Covered Well  Column Number - 74  Un-covered Well  Column Number - 75    Spring  Column Number - 78    River/Canal  Column Number - 79    Tank/Pond/Lake  Column Number - 80    Other Sources -  Column Number - 81 |
| MSW\_INT | Household | Handpump  Column Number - 76    Tube well/ Borehole  Column Number - 77 |
| MSW\_ADV | Household | Tap water from treated Source  Column Number - 72    Tap water from untreated source  Column Number - 73 |
| ASSET | TV | Household | Television  Column Number - 129 |
| Tele | Household | Landline Only  Column Number - 132  Mobile Only  Column Number - 133    Both  Column Number - 134 |
| 2w | Household | Scooter/ Motorcycle/Moped  Column Number - 136 |
| 4w | Household | Car/Jeep/Van  Column Number - 137 |
| Literacy Population |  | PCA | P\_LIT |
| Population 2011 |  | PCA | TOT\_P |
| Number of Households 2011 |  | PCA | NO\_HH |

**Table 2.2: Village\_Household\_Percentages\_RudIntAdv2011.csv**

**[**[**Link**](https://drive.google.com/file/d/1GrEz4lCE_S0gByfVHbwao1R3T8OcCuWA/view?usp=drive_link)**]**

This table contains the computed values of percentage of households under Rudimentary/ Intermediate/ and Advance level of selected socio-economic indicators for the year 2011.

| **Column** | **Type** | **Description** |
| --- | --- | --- |
| Village\_id\_2011 | INT | 2011 Census Village ID used as index to this table |
| BF\_RUD\_2011 | DOUBLE | Percentage of households in a village with rudimentary bathroom facilities. |
| BF\_INT\_2011 | DOUBLE | Percentage of households in a village with intermediate bathroom facilities. |
| BF\_ADV\_2011 | DOUBLE | Percentage of households in a village with advanced bathroom facilities. |
| FC\_RUD\_2011 | DOUBLE | Percentage of households in a village with rudimentary fuel for cooking. |
| FC\_INT\_2011 | DOUBLE | Percentage of households in a village with intermediate fuel for cooking. |
| FC\_ADV\_2011 | DOUBLE | Percentage of households in a village with advanced fuel for cooking. |
| MSW\_RUD\_2011 | DOUBLE | Percentage of households in a village with rudimentary main source of drinking water. |
| MSW\_INT\_2011 | DOUBLE | Percentage of households in a village with intermediate main source of drinking water. |
| MSW\_ADV\_2011 | DOUBLE | Percentage of households in a village with advanced main source of drinking water. |
| TV\_2011 | DOUBLE | Percentage of households in a village that own a television. |
| Tele\_2011 | DOUBLE | Percentage of households in a village that own a telephone. |
| 2w\_2011 | DOUBLE | Percentage of households in a village that own a two-wheeler. |
| 4w\_2011 | DOUBLE | Percentage of households in a village that own a four-wheeler. |
| MSL\_RUD\_2011 | DOUBLE | Percentage of households in a village with a rudimentary main source of light. |
| MSL\_INT\_2011 | DOUBLE | Percentage of households in a village with intermediate main source of light. |
| MSL\_ADV\_2011 | DOUBLE | Percentage of households in a village with advanced main source of light. |
| LIT\_2011 | DOUBLE | Percentage of literate households in a village. |

**Section 3: Pre-processed Groundtruth data used in Village Development Model**

**Table 3.1: Village\_Level\_GroundTruth\_14\_States.csv**

**[**[**Link**](https://drive.google.com/file/d/1AAn4tVsm5DGY3aa-xiLGYIgZvNxFbv_w/view?usp=drive_link)**]**

This table contains a list of socio-economic indicators at village level for the years 2001 and 2011.

Indicators like (BF, FC, ASSET, LIT, MSW) are computed only for the year 2011. Each village is ranked either 1, 2, or 3 against these indicators based on their clustering output.

Other indicators related to schools, hospitals, cinemas, etc. are the common set of indicators between 2001 and 2011 census data.

| # | Columns | Non-Null | Dtype |
| --- | --- | --- | --- |
| 0 | Village\_2011\_ID | 529099 | float64 |
| 1 | Village\_2001\_ID | 529099 | object |
| 2 | Village\_Name | 529099 | object |
| 3 | Govt\_Primary\_School\_(Numbers)\_2001 | 315740 | float64 |
| 4 | Govt\_Middle\_School\_ (Numbers)\_2001 | 315740 | float64 |
| 5 | Govt\_Arts\_and\_Science\_Degree\_College\_(Numbers)\_2001 | 315740 | float64 |
| 6 | Hospital\_Allopathic\_(Numbers)\_2001 | 315740 | float64 |
| 7 | Maternity\_And\_Child\_Welfare\_Centre\_(Numbers)\_2001 | 315740 | float64 |
| 8 | Primary\_Health\_Centre\_(Numbers)\_2001 | 315740 | float64 |
| 9 | Post\_Office\_(Status\_A(1)/NA(2))\_2001 | 315740 | float64 |
| 10 | Telephone\_( landlines)\_(Status\_A(1)/NA( 2 ))\_2001 | 315740 | float64 |
| 11 | Public\_Bus\_Service\_(Status\_A(1)/NA (2))\_2001 | 315740 | float64 |
| 12 | Railway\_Station\_(Status\_A(1)/NA(2))\_2001 | 315740 | float64 |
| 13 | Navigable\_Waterways\_(River/Canal)\_(Status\_A(1)/NA(2))\_2001 | 315740 | float64 |
| 14 | Sports\_Club/Recreation\_Centre\_(Status\_A(1)/NA(2))\_2001 | 315740 | float64 |
| 15 | Cinema/Video\_Hall\_ (Status\_A(1)/NA(2))\_2001 | 315740 | float64 |
| 16 | Govt\_Primary\_School\_(Numbers)\_2011 | 345018 | float64 |
| 17 | Govt\_Middle\_School\_ (Numbers)\_2011 | 345018 | float64 |
| 18 | Govt\_Arts\_and\_Science\_Degree\_College\_(Numbers)\_2011 | 345018 | float64 |
| 19 | Hospital\_Allopathic\_(Numbers)\_2011 | 345018 | float64 |
| 20 | Primary\_Health\_Centre\_(Numbers)\_2011 | 345018 | float64 |
| 21 | Maternity\_And\_Child\_Welfare\_centre\_(Numbers)\_2011 | 345018 | float64 |
| 22 | Post\_Office\_(Status\_A(1)/NA(2))\_2011 | 345018 | float64 |
| 23 | Telephone\_(landlines)\_(Status\_A(1)/NA(2))\_2011 | 345018 | float64 |
| 24 | Public\_Bus\_service\_ (Status\_A(1)/NA(2))\_2011 | 345018 | float64 |
| 25 | Railway\_Station\_(Status\_A(1)/NA(2))\_2011 | 345018 | float64 |
| 26 | Navigable\_Waterways\_(River/Canal)\_(Status\_A(1)/NA(2))\_2011 | 345018 | float64 |
| 27 | Sports\_Club/Recreation\_Centre\_(Status\_A(1)/NA(2))\_2011 | 345018 | float64 |
| 28 | Cinema/Video\_Hall\_(Status\_A(1)/NA(2))\_2011 | 345018 | float64 |
| 29 | ADI\_2011 | 345018 | float64 |
| 30 | BF\_2011 | 345018 | float64 |
| 31 | FC\_2011 | 345018 | float64 |
| 32 | ASSET\_2011 | 345018 | float64 |
| 33 | LIT\_2011 | 345018 | float64 |
| 34 | MSW\_2011 | 345018 | float64 |

**Section 4: Predicted Labels at Village Level**

**Table 4.1: Village\_Level\_Predictions\_14\_States.csv**

**[**[**Link**](https://drive.google.com/file/d/1jNXn3QeQBF_lOQMCSeQUngzhuFEll1Sp/view?usp=drive_link)**]**

This table contains the list of predicted socio-economic indicators at village level for the years 2001, 2011, and 2019. The classifier was trained on the data of year 2011. For the villages over which this prediction was not done, their row contains null values.

| # | Column | Non-Null Count | Dtype |
| --- | --- | --- | --- |
| 0 | Village\_2011\_ID | 529099 | float64 |
| 1 | Village\_2001\_ID | 529099 | object |
| 2 | Village\_Name | 529099 | object |
| 3 | BF\_2001 | 315744 | float64 |
| 4 | FC\_2001 | 315744 | float64 |
| 5 | MSW\_2001 | 315744 | float64 |
| 6 | ASSET\_2001 | 315744 | float64 |
| 7 | LIT\_2001 | 315744 | float64 |
| 8 | ADI\_2001 | 315744 | float64 |
| 9 | BF\_2011 | 345018 | float64 |
| 10 | FC\_2011 | 345018 | float64 |
| 11 | MSW\_2011 | 345018 | float64 |
| 12 | ASSET\_2011 | 345018 | float64 |
| 13 | LIT\_2011 | 345018 | float64 |
| 14 | ADI\_2011 | 345018 | float64 |
| 15 | BF\_2019 | 313476 | float64 |
| 16 | FC\_2019 | 313476 | float64 |
| 17 | MSW\_2019 | 313476 | float64 |
| 18 | ASSET\_2019 | 313476 | float64 |
| 19 | LIT\_2019 | 313476 | float64 |
| 20 | ADI 2019 | 313476 | float64 |

**Table 4.2: Gini\_RWI\_2001\_2011\_2019.csv**

**[**[**Link**](https://drive.google.com/file/d/1THBjqHexYPGrfBOQAnw3AxAWVM_LhSJ1/view?usp=drive_link)**]**

This table contains the list of predicted socio-economic indicators (Relative Wealth Index) at village level for the years 2001, 2011, and 2019.

| # | Column | Non-Null Count | Dtype |
| --- | --- | --- | --- |
| 0 | Village\_2011\_ID | 529099 | float64 |
| 1 | RWI 2001 | 315910 | object |
| 2 | RWI 2011 | 315910 | object |
| 3 | RWI 2019 | 315910 | float64 |

**Table 4.3: Gini\_ADI\_2001\_2011\_2019.csv**

**[**[**Link**](https://drive.google.com/file/d/1bXi_-chgAhIRdLlVUdfMGHIc4vp4H-7U/view?usp=drive_link)**]**

This table contains different coefficients computed at the district level in India like the Gini Coefficient.

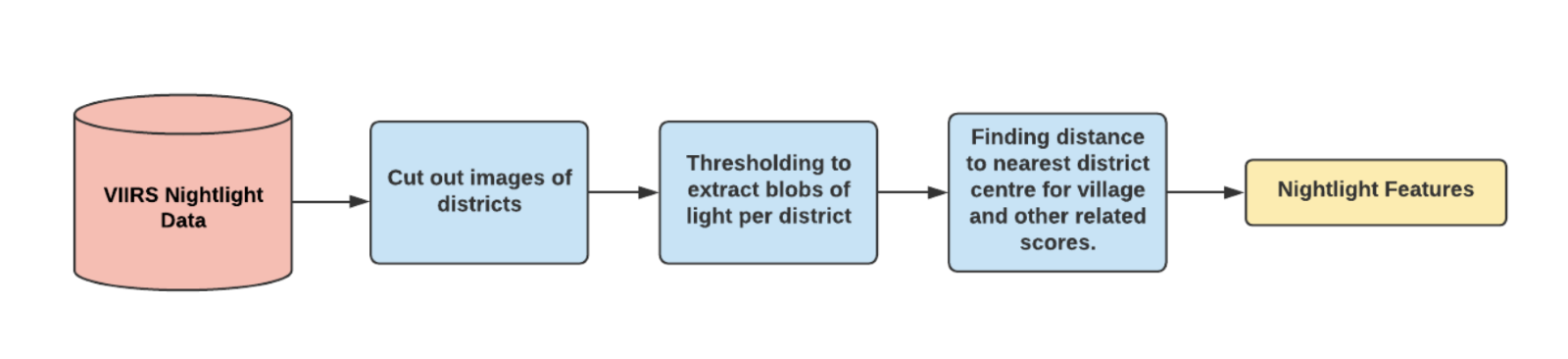
| **Column** | **Type** | **Description** |
| --- | --- | --- |
| STATE\_CODE\_2011 | INT | 2011 Census State Code |
| DIST\_CODE\_2011 | INT | 2011 Census District Code |
| Num\_villages\_in\_District\_2001 | INT | Number of villages in a district in 2001 |
| Num\_villages\_in\_District\_2011 | INT | Number of villages in a district in 2011 |
| Num\_villages\_in\_District\_2019 | INT | Number of villages in a district in 2019 (identical to 2011) |
| Gini\_2001 | DOUBLE | Gini Index of a District for 2001 |
| Gini\_2011 | DOUBLE | Gini Index of a District for 2011 |
| Gini\_2019 | DOUBLE | Gini index of a district for 2019 |
| Population\_sum\_2001 | INT | Population Sum for a District in 2001 |
| Population\_sum\_2011 | INT | Population Sum for a District in 2011 |
| Population\_sum\_2019 | INT | Population Sum for a District in 2019 |
| Adi\_product\_population\_sum\_2001 | INT | (ADI X Population) sum for a District in 2001 |
| Adi\_product\_population\_sum\_2011 | INT | (ADI X Population) sum for a District in 2011 |
| Adi\_product\_population\_sum\_2019 | INT | (ADI X Population) sum for a District in 2019 |
| Num\_villages\_in\_District\_2011 | INT | Number of villages in a District |
| Average\_ADI\_per\_population\_2001 | DOUBLE | Average Rural ADI in a district in 2001 |
| Average\_ADI\_per\_population\_2011 | DOUBLE | Average Rural ADI in a district in 2011 |
| Average\_ADI\_per\_population\_2019 | DOUBLE | Average Rural ADI in a district in 2019 |

**Table 4.4: Nightlight\_Features.csv**

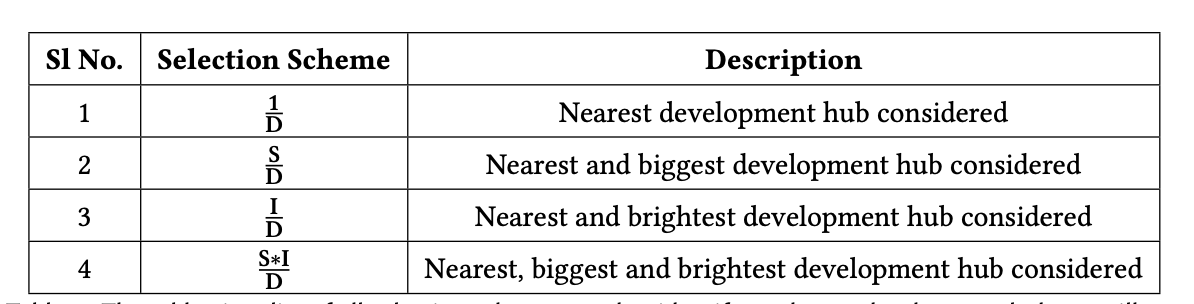
**[**[**Link**](https://drive.google.com/file/d/1URhhdjFtk5DkSywKzFNGtJFL9h3DDdDF/view?usp=drive_link)**]**

21 nightlight features were derived for the years 2001, 2011, and 2019.

Details are as follows:

****

* We extract the nightlight data from the [DMSP-VIIRS dataset](https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/YGIVCD) and then identify all the light blobs in districts of India. We call these light blobs Development Hubs, which indicate regions of socio-economic activity and signs of development in that district. In order to identify development hubs (light blobs) within a district we use Otsu thresholding
* Once all the blobs(development hub) are identified in all the districts, we can use various blob selection schemes based on development hub physical features like distance from development hub (D), the size of the light blob or development hub (S), the intensity of the development hub (I), i.e most likely to affect a village’s development.

****

* For all the nightlight features of a village mathematical wrapping functions were applied. These include square root, log or no feature at all.
* A particular nightlight feature is represented as *dc\_score\_wrapper\_function\_feature\_computation\_scheme*, where *feature\_computation\_scheme* and *wrapper\_function* are of multiple types.
* Dc score indicates average nightlight luminosity over the village. The first argument indicates the wrapper function and it is one of {none, sqrt, log}. The second argument indicates the features that were considered while mapping nightlight intensity of villages to blobs, it is a combination of {none, size, intensity} The overall unit of measurement of dc\_score is the following.



| **#** | **Column** | **Non-null Count** | **Dtype** |
| --- | --- | --- | --- |
| 0 | Village\_ 2011\_ID | 537449 | float64 |
| 1 | Distance from Development Hub | 497692 | float64 |
| 2 | dc\_ score\_ none\_ none\_ 2001 | 497692 | float64 |
| 3 | dc\_score\_none\_size\_2001 | 497692 | float64 |
| 4 | dc\_score\_none\_size\_intensity\_2001 | 497692 | float64 |
| 5 | dc\_ score\_ none\_ intensity\_2001 | 497692 | float64 |
| 6 | dc\_score\_log\_none\_2001 | 497692 | float64 |
| 7 | dc\_ score\_ log\_ size\_ 2001 | 497692 | float64 |
| 8 | dc\_ score\_ log\_ size\_ intensity\_ 2001 | 497692 | float64 |
| *9* | dc\_score\_ log\_intensity\_2001 | 497692 | float64 |
| 10 | dc\_ score\_sqrt\_ none\_ 2001 | 497692 | float64 |
| 11 | dc\_ score\_ sqrt\_ size\_ 2001 | 497692 | float64 |
| 12 | dc\_score\_sqrt\_size\_intensity\_2001 | 497692 | float64 |
| 13 | dc\_ score\_ sqrt\_ intensity\_2001 | 497692 | float64 |
| 14 | dc\_ score\_ none\_ size\_only\_2001 | 497692 | float64 |
| 15 | dc\_ score\_ none\_ size\_ intensity\_only\_ 2001 | 497692 | float64 |
| 16 | dc\_ score\_ none\_ intensity\_only\_ 2001 | 497692 | float64 |
| 17 | dc\_score\_ log\_size\_only\_2001 | 497692 | float64 |
| 18 | dc\_ score\_ log\_ size\_ intensity\_ only\_ 2001 | 497692 | float64 |
| 19 | dc\_ score\_ log\_ intensity\_only\_ 2001 | 497692 | float64 |
| 20 | dc\_score\_sqrt\_size\_only\_2001 | 497692 | float64 |
| 21 | dc\_ score\_ sqrt\_ size\_ intcnsity\_only\_ 2001 | 497692 | float64 |
| 22 | dc\_ score\_ sqrt\_ intensity\_only\_ 2001 | 497692 | float64 |
| *23* | dc\_ score\_ none\_none\_2011 | 497692 | float64 |
| 24 | dc\_ score\_ none\_ size\_ 2011 | 497692 | float64 |
| 25 | dc\_score\_none\_size\_intensity\_2011 | 497692 | float64 |

NOTE: There are more columns in this table which are made from combinations of selection functions which aren’t shown here.

**Section 5: Village Shapefile Data**

Village-level shapefiles were downloaded for the year 2001 from Nasa Sedac website [[Link](https://sedac.ciesin.columbia.edu/data/set/india-india-village-level-geospatial-socio-econ-1991-2001/data-download)] [[Documentation1](https://drive.google.com/file/d/1yJ43xgocOFKh2N8EqaczmZSYuH9ukBRP/view?usp=drive_link)][[Documentation](https://sedac.ciesin.columbia.edu/downloads/docs/india/india-india-village-level-geospatial-socio-econ-1991-2001-documentation.pdf)2].

**Table 5.1: Shrug\_2001\_2011\_Village\_Mapping**

**[**[**Link**](https://drive.google.com/file/d/1yJ43xgocOFKh2N8EqaczmZSYuH9ukBRP/view?usp=drive_link)**]**

Note on 2001-2011 Village Mapping:

**-** SHRUG identified the mapping of villages and towns (District) as a very challenging task as each decennial Census opted to create new location identifiers since 1991

- The Census provides digital keys to link villages and towns to prior censuses, but they were highly incomplete. The Census district handbooks contain detailed descriptions of boundary changes in narrative format only. All these sources had errors and inconsistencies.

**-** SHRUG used both the digital linking keys and the district handbooks to create the best possible correspondence between villages and towns across the 1991, 2001, and 2011 censuses. They also supplemented this with a custom fuzzy string-matching program to match village and town names over time.

Further Details can be found [here](http://shrug-assets-ddl.s3.amazonaws.com/static/main/assets/other/almn-shrug.pdf)

| **Column** | **Type** | **Description** |
| --- | --- | --- |
| pc01\_state\_id | INT | State id as per census 2001 |
| pc01\_district\_id | INT | District id as per census 2001 |
| pc01\_subdistrict\_id | INT | Subdistrict id as per census 2001 |
| pc01\_village\_id | INT | Village id as per census 2001 |
| shrid |  | Shrug id |
| pc11\_state\_id | INT | State id as per census 2011 |
| pc11\_district\_id | INT | District id as per census 2011 |
| pc11\_subdistrict\_id | INT | Subdistrict id as per census 2011 |
| pc11\_village\_id | INT | Village id as per census 2011 |

**Table 5.2: Village\_Shapefiles**

**[**[**Link**](https://drive.google.com/drive/folders/1NU60SUWkcOn_yCTjAe176zwGjVoXmghm?usp=drive_link)**]**

This folder contains village shapefiles divided into state-wise geojson files.

There are 20 geojson files corresponding to the following states-

| **File Name** | **State** |
| --- | --- |
| AP.geojson | Andhra Pradesh |
| BR.geojson | Bihar |
| CG.geojson | Chhattisgarh |
| GA.geojson | Goa |
| GJ.geojson | Gujrat |
| HR.geojson | Haryana |
| HP.geojson | Himachal Pradesh |
| JH.geojson | Jharkhand |
| KA.geojson | Karnataka |
| KL.geojson | Kerala |
| MH.geojson | Maharashtra |
| MP.geojson | Madhya Pradesh |
| MZ.geojson | Mizoram |
| OR.geojson | Orissa |
| PB.geojson | Punjab |
| RJ.geojson | Rajasthan |
| SK.geojson | Sikkim |
| TN.geojson | Tamil Nadu |
| UP.geojson | Uttar Pradesh |
| WB.geojson | West Bengal |

Each entry in the csv file has the following schema-

| **Column Name** | **Data Type** | **Description** |
| --- | --- | --- |
| Village\_2011\_ID | INT | Village 2011 census id mapped as per SHRUG |
| Village\_2001\_ID | INT | Village 2001 census id |
| Village\_Name | String | Village name |
| Vill\_Geometry | Json object | Village shapefile from NASA SEDAC |
| **Contains Meta Data as well** | | |