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 be proactive in shaping 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

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#

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

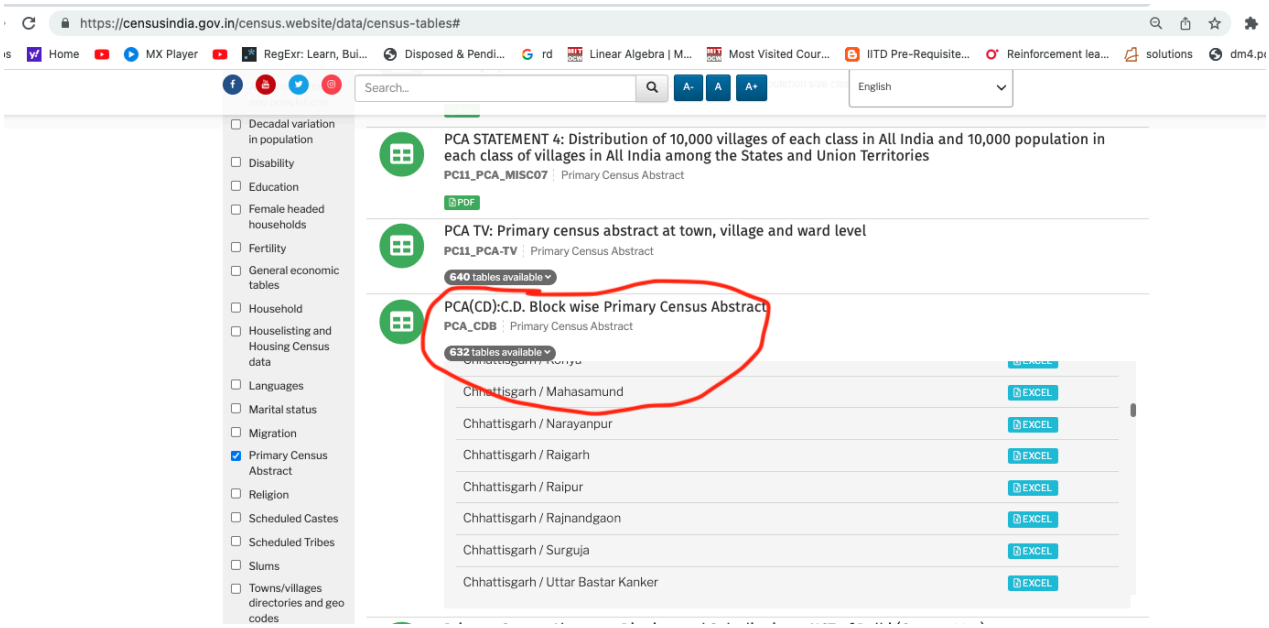
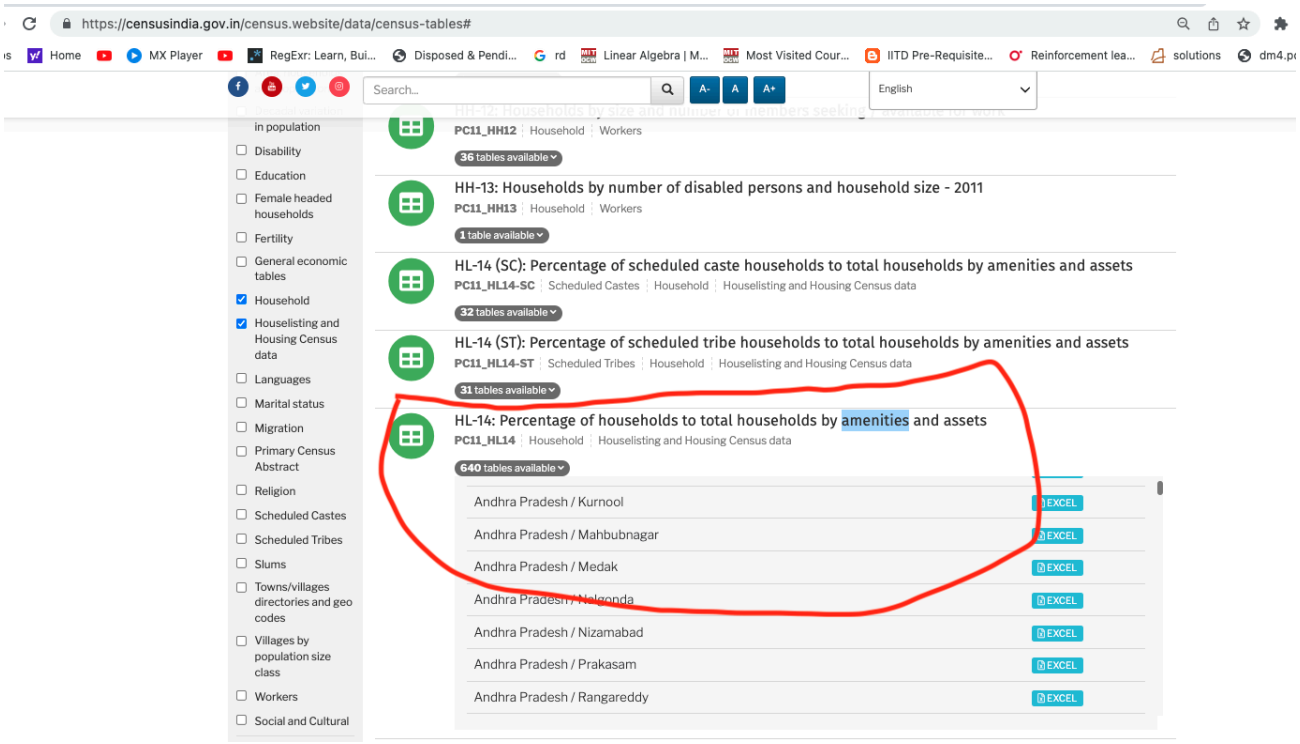


Table 1.2: Household Data (select year 2011):

https://censusindia.gov.in/census.website/data/census-tables#

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]

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]

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

Census data 2011 raw village level data : [LINK]

Census 2011 Metadata: [LINK]

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]

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 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]

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]

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_ADI_2001_2011_2019.csv

[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 4.3: Nightlight_Features.csv

[Link]

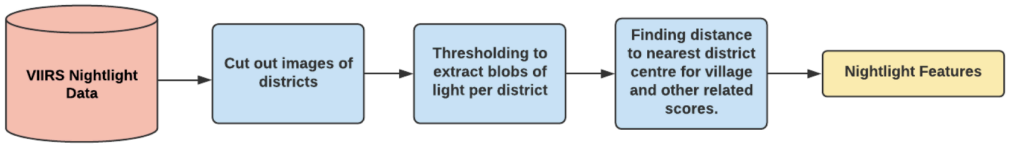
21 nightlight features were derived for the years 2001, 2011, and 2019.

Definition: Nightlight intensity obtained from satellite images refers to the amount of light emitted by human settlements and infrastructure during the night, as captured by specialized satellites. These satellites are equipped with sensors that can detect the intensity of light emitted in different parts of the electromagnetic spectrum.

Nighttime lights are used to understand various human activities and urbanization patterns, including population density, economic activity, and energy consumption. They are also used for environmental monitoring, disaster response, and security applications.

Nighttime light intensity is usually expressed in radiance or brightness values, which are measured in units of watts per square meter per steradian (W/m2/sr). The radiance values are then converted into digital numbers (DN) that can be processed and analyzed using specialized software.

Details are as follows:



- We extract the nightlight data from the DMSP-VIIRS dataset 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.

Sl No.	Selection Scheme	Description
1	$\frac{1}{D}$	Nearest development hub considered
2	$\frac{S}{D}$	Nearest and biggest development hub considered
3	$\frac{I}{D}$	Nearest and brightest development hub considered
4	$\frac{S*I}{D}$	Nearest, biggest and brightest development hub considered

- 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.

$$nW\,cm^{-2}\,sr^{-1}$$

#	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][Documentation].

Now, since the socio-economic groundtruth exists for the year 2011, we used Shrug data to map 2001 village census id to 2011 village census id. [2001 Data Link][2011 Data Link] (To read MetaData -> Access File -> Variable Metadata)

Table 5.1: Shrug_2001_2011_Village_Mapping

[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

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]

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

There are 20 csv files corresponding to the following states-

CSV File Name	State
AP.csv	Andhra Pradesh
BR.csv	Bihar
CG.csv	Chhattisgarh
GA.csv	Goa
GJ.csv	Gujrat
HR.csv	Haryana
HP.csv	Himachal Pradesh
JH.csv	Jharkhand
KA.csv	Karnataka
KL.csv	Kerala
MH.csv	Maharashtra
MP.csv	Madhya Pradesh
MZ.csv	Mizoram
OR.csv	Orissa
PB.csv	Punjab

RJ.csv	Rajasthan
SK.csv	Sikkim
TN.csv	Tamil Nadu
UP.csv	Uttar Pradesh
WB.csv	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

Section 6: Village_panchayat_mapping

[Link]

The district-wise village-panchayat mapping is available at https://lgdirectory.gov.in/rptConsolidateVillageGramPanchayat.do

This is a Local Government Directory therefore it is expected to be a reliable source.

We generated a unique ID for each panchayat by ourselves.

Column Name	Data type	Description
Village_2011_ID	INT	Village census id as per 2011
Village_Name	String	Village name as per census 2011
Panchayat_ID	INT	Unique panchayat id generated by us
Gram_panchayat_name	String	Name of Gram Panchayat
Subdistrict_2011_name	String	Subdistrict name as per census 2011
Subdistrict_2022_name	String	Subdistrict name in the year 2022
Subdistrict_2011_ID	INT	Subdistrict census id as per 2011
District_2011_name	String	District name as per census 2011
District_2022_name	String	District name in the year 2022
District_2011_ID	INT	District census id as per 2011
State_2011_name	String	State name as per census 2011
State_2022_name	String	State name in the year 2022
State_2011_ID	INT	State census id as per 2011