

Merge Data with SHRUG Documentation

Nishant - Intern @ IIM Bangalore

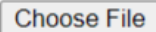
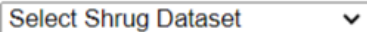

Contents

1	How to Use the Web Application	2
2	Internal Processes in the Application	3
2.1	Overview	3
2.2	SHRUG Identifier Naming Convention	3
2.3	Example Data	3
2.4	Creating SHRUG Identifiers for User Datasets	4
2.4.1	Two Possible Scenarios:	4
2.4.2	Challenge and Solution:	4
2.4.3	Steps to Map Villages in User Datasets to SHRIDs:	4
2.5	Complete steps explained through example	5
3	How Fuzzy Matching is performed	7
4	Appendix	8

1 How to Use the Web Application

Merge Data with SHRUG

This application helps you to merge your dataset with a variety of SHRUG datasets. Upload your CSV file and choose the SHRUG dataset to merge with. Refer to this following document to see the format of data to be uploaded and results expected.

 No file chosen  

Click here to choose your CSV file Select the Shrug Dataset you want your data to be merged with. Click on Upload to start the merging process

```
Mapping created for State Names
{'KARNATAKA': 'Karnataka', 'KERALA': 'Kerala'}

Step 2 started. Mapping District Names

Mapping created for Karnataka's District Names
{'SHIMOGA': 'Shimoga', 'TUMKUR': 'Tumkur', 'UDUPI ': 'Udupi', 'YADGIR': 'Yadgir'}

Mapping created for Karnataka's District Names
{'SHIMOGA': 'Shimoga', 'TUMKUR': 'Tumkur', 'UDUPI ': 'Udupi', 'YADGIR': 'Yadgir'}

Mapping created for Kerala's District Names
{'ALAPPUZHA': 'Alappuzha'}

Step 3 started. Mapping Block Names

Mapping created for Kerala's District Names
{'ALAPPUZHA': 'Alappuzha'}

Step 3 started. Mapping Block Names
```

This box keeps updating you about all internal processes running for the merging process

Nishant - Intern @ IIM Bangalore
Jal Jeevan Mission

Figure 1: Preview of the Web APP

2 Internal Processes in the Application

2.1 Overview

The Socioeconomic High-resolution Rural-Urban Geographic Platform for India (SHRUG) is a geographic platform designed to facilitate data sharing among researchers studying India. It is an open-access repository that currently includes dozens of datasets covering over 500,000 villages and 8,000 towns across India, using common geographic identifiers spanning 25 years.

The backbone of SHRUG is a set of keys that link all the datasets to have a unique identifier for each Village/Town. These keys are called “Shrid” i.e Shrug ID of villages. So once we map our dataset’s Villages to these Shrids, we can integrate all kinds of other data available on SHRUG datasets.

2.2 SHRUG Identifier Naming Convention

- For PC91: Shrid is - YY-SS-DD-ssss-TTTTT
- For PC01: Shrid is - YY-SS-DDD-ssss-TTTTTTTT
- For PC11: Shrid is - YY-SS-DDD-sssss-TTTTTT

In all SHRUG identifiers:

- PC is Population Census and 91 is 1991, 01 is short for 2001 and 11 is short for 2011
- YY indicates the most recent census year to which the identifier is matched. If observations are matched to 2011 census locations, we use ”11”. If matched to 2001 census locations but not 2011, we use ”01”, etc.
- SS indicates the state identifier corresponding to the census year YY.
- DDD indicates the district code.
- ss indicates the sub district code.
- VV/TT indicates the census code of the most populous town or village in the identifier, based on the census year YY.

2.3 Example Data

Shrid	Elevation Mean	Elevation Median	Elevation Min	Elevation Max	Elevation Num Cells	Elevation Std
11-01-001-00001-000002	1758.410648	1745	1276	2406	8321	251.3905621
11-01-001-00001-000005	2399.254718	2395	1823	3081	6623	268.8287162
11-01-001-00001-000006	3176.797998	3206	2597	3651	6594	261.2333539
11-01-001-00001-000007	2846.648941	2875	2363	3255	4156	203.6043532
11-01-001-00001-000008	2825.052265	2830	2482	3123	3444	148.4662475
11-01-001-00001-000009	2285.574446	2290	1772	2796	6629	241.630916
11-01-001-00001-000010	1939.292801	1924	1720	2301	7418	132.4220257
11-01-001-00001-000011	1899.272352	1857	1727	2209	2662	118.246196
...
11-01-001-00001-000012	1834.946675	1791	1710	2207	6451	111.2142782
11-01-001-00001-000013	2051.902734	2060	1745	2332	3475	134.4856946
11-01-001-00001-000014	1709.933505	1691	1649	1873	5835	50.86910837
11-01-001-00001-000015	1710.769734	1706	1660	1787	5498	25.17922835
11-01-001-00001-000016	1746.466732	1720	1697	2081	11182	60.15782552
11-01-001-00001-000017	1776.560393	1748	1713	2074	6408	68.13228237
11-01-001-00001-000018	2362.450544	2439	1777	2749	6066	259.5195425

Table 1: Satellite-derived elevation and terrain ruggedness from SHRUG for example. This dataset has the shrid column and all the data.

It is important to note that SHRUG datasets only contain SHRUG IDs as the identifier for villages. Therefore, to merge the user’s data with SHRUG data, it is essential to create SHRIDS for the user’s data.

2.4 Creating SHRUG Identifiers for User Datasets

In order to create SHRUG identifiers (Shrids) for a user's dataset, it is essential to have the Village, Block, District, and State codes for each village in the dataset. Using the SHRUG naming convention, we can then generate the Shrid column and easily merge the user's data with SHRUG datasets.

2.4.1 Two Possible Scenarios:

1. **User's Data Contains Census 2011 Codes:**

If the user's dataset already includes Village, Block, District, and State codes from the 2011 census, we can directly use these to create the Shrid and merge with SHRUG data.

2. **User's Data Contains Only Names:**

If the user's dataset only has Village, Block, District, and State names without IDs, we need a reference dataset to map these names to the corresponding Census IDs. This reference file, created using open-source shapefiles for India available on SHRUG, provides the mapping of Census Village, District, Block, and State names to their Census IDs.

2.4.2 Challenge and Solution:

The primary challenge is ensuring that the user's dataset uses the same naming conventions as Census 2011. To address this, we perform fuzzy matching of the user's state, district, subdistrict, and village names to the reference files. Once matched, we can map these names to their corresponding Census IDs and create the SHRUG IDs.

2.4.3 Steps to Map Villages in User Datasets to SHRIDs:

1. **Fuzzy Match Names:**

- First, fuzzy match the State names.
- Then, fuzzy match the District names.
- Next, fuzzy match the Block names.
- Finally, fuzzy match the Village names from the user's dataset to the reference file.

2. **Merge with Reference Data:**

Merge the user's dataset with the reference dataset to obtain the IDs for each State, Village, District, and Block.

3. **Create SHRUG Identifiers:**

Using the SHRUG Identifier Naming Convention, create the Shrid column in the user's dataset.

4. **Merge with SHRUG Data:**

Finally, merge the user's dataset with the desired SHRUG dataset.

2.5 Complete steps explained through example

	StateName	DistrictName	BlockName	VillageName	Contaminents
0	KARNATAKA	SHIMOGA	BHADRAVATI	AGARADAHALLI	Fluoride : 1.000, Iron : 0.400, Chloride : 250...
1	KARNATAKA	SHIMOGA	BHADRAVATI	AGARADAHALLI	Fluoride : 1.000, Iron : 0.300, Chloride : 400...
2	KARNATAKA	SHIMOGA	BHADRAVATI	ANAVERI	Fluoride : 1.000, Iron : 0.400, Chloride : 300...
3	KARNATAKA	SHIMOGA	BHADRAVATI	ANAVERI	Fluoride : 0.500, Iron : 0.500, Chloride : 450...
4	KARNATAKA	SHIMOGA	BHADRAVATI	ITTIGEHALLY	Fluoride : 0.500, Iron : 0.500, Chloride : 450...
...
148	KERALA	ALAPPUZHA	ARYAD	PATHIRAPPALLY	Fluoride : 1.490, Iron : 0.050, Chloride : 195...
149	KERALA	ALAPPUZHA	ARYAD	PATHIRAPPALLY	Fluoride : 2.280, Chloride : 100.000, TDS : 48...
150	KERALA	ALAPPUZHA	ARYAD	PATHIRAPPALLY	Fluoride : 2.200, Iron : 0.050, Chloride : 560...
151	KERALA	ALAPPUZHA	ARYAD	PATHIRAPPALLY	Fluoride : 2.180, Iron : 0.060, Chloride : 110...
152	KERALA	ALAPPUZHA	ARYAD	PATHIRAPPALLY	Fluoride : 2.240, Chloride : 540.000, TDS : 14...

1344 rows × 5 columns

(a) Example Dataset with which we'll merge Shrug Dataset

	pc11_s_id	pc11_d_id	pc11_sd_id	pc11_tv_id	tv_name	State Name	sd_name	d_name
0	35	638	5916	0	Batti Malv Island	Andaman & Nicobar Islands	Car Nicobar	Nicobars
1	35	638	5916	645012	Mus	Andaman & Nicobar Islands	Car Nicobar	Nicobars
2	35	638	5916	645013	Teetop	Andaman & Nicobar Islands	Car Nicobar	Nicobars
3	35	638	5916	645014	Sawai	Andaman & Nicobar Islands	Car Nicobar	Nicobars
4	35	638	5916	645015	Arong	Andaman & Nicobar Islands	Car Nicobar	Nicobars
Population Census 2011 Le PC 11 IDs for State, District, Subdistrict and Town/Village					Names of State, District, Subdistrict and Village			
654301	19	330	2196	309796	Posaltair	West Bengal	Raiganj	Uttar Dinajpur
654302	19	330	2196	309797	Pardha	West Bengal	Raiganj	Uttar Dinajpur
654303	19	330	2196	309798	Nachhratpur Katabari	West Bengal	Raiganj	Uttar Dinajpur
654304	19	330	2196	309799	Kasba	West Bengal	Raiganj	Uttar Dinajpur
654305	19	330	2196	801651	Raiganj	West Bengal	Raiganj	Uttar Dinajpur

654306 rows × 8 columns

(b) Reference file, provides the mapping of Census Village, District, Block, and State names to their Census IDs.

	StateName	DistrictName	BlockName	VillageName	Contaminents	Matched State Name	Matched District name	Matched SubDistrict name	Matched Village name
0	KARNATAKA	SHIMOGA	BHADRAVATI	AGARADAHALLI	Fluoride : 1.000, Iron : 0.400, Chloride : 250...	Karnataka	Shimoga	Bhadravati	Agaradahalli
1	KARNATAKA	SHIMOGA	BHADRAVATI	AGARADAHALLI	Fluoride : 1.000, Iron : 0.300, Chloride : 400...	Karnataka	Shimoga	Bhadravati	Agaradahalli
2	KARNATAKA	SHIMOGA	BHADRAVATI	ANAVERI	Fluoride : 1.000, Iron : 0.400, Chloride : 300...	Karnataka	Shimoga	Bhadravati	Anaveri
3	KARNATAKA	SHIMOGA	BHADRAVATI	ANAVERI	Fluoride : 0.500, Iron : 0.500, Chloride : 450...	Karnataka	Shimoga	Bhadravati	Anaveri
4	KARNATAKA	SHIMOGA	BHADRAVATI	ITTIGEHALLY	Fluoride : 0.500, Iron : 0.500, Chloride : 450...	Karnataka	Shimoga	Bhadravati	Ittigehalli
...
148	KERALA	ALAPPUZHA	ARYAD	PATHIRAPPALLY	Fluoride : 1.490, Iron : 0.050, Chloride : 195...	Kerala	Alappuzha	Kuttanad	Pathirappally
149	KERALA	ALAPPUZHA	ARYAD	PATHIRAPPALLY	Fluoride : 2.280, Chloride : 100.000, TDS : 48...	Kerala	Alappuzha	Kuttanad	Pathirappally
150	KERALA	ALAPPUZHA	ARYAD	PATHIRAPPALLY	Fluoride : 2.200, Iron : 0.050, Chloride : 560...	Kerala	Alappuzha	Kuttanad	Pathirappally
151	KERALA	ALAPPUZHA	ARYAD	PATHIRAPPALLY	Fluoride : 2.180, Iron : 0.060, Chloride : 110...	Kerala	Alappuzha	Kuttanad	Pathirappally
152	KERALA	ALAPPUZHA	ARYAD	PATHIRAPPALLY	Fluoride : 2.240, Chloride : 540.000, TDS : 14...	Kerala	Alappuzha	Kuttanad	Pathirappally

1344 rows × 9 columns

Figure 3: After applying Fuzzy Matching we get the Matches for State, District, Subdistrict and Village in the user's input dataset

	pc11_s_id	pc11_d_id	pc11_sd_id	pc11_tv_id	tv_name	State Name	sd_name	d_name	StateName	DistrictName	BlockName	VillageName	Contaminents
0	29	568	5521	608518	Diggenahalli	Karnataka	Bhadravati	Shimoga	KARNATAKA	SHIMOGA	BHADRAVATI	DIGGENAHALLI	Iron : 0.500, Chloride : 450.000, Nitrate : 55...
1	29	568	5521	608522	Anaveri	Karnataka	Bhadravati	Shimoga	KARNATAKA	SHIMOGA	BHADRAVATI	ANAVERI	Fluoride : 1.000, Iron : 0.400, Chloride : 300...
2	29	568	5521	608522	Anaveri	Karnataka	Bhadravati	Shimoga	KARNATAKA	SHIMOGA	BHADRAVATI	ANAVERI	Fluoride : 0.500, Iron : 0.500, Chloride : 450...
3	29	568	5521	608523	Ittigehalli	Karnataka	Bhadravati	Shimoga	KARNATAKA	SHIMOGA	BHADRAVATI	ITTIGEHALLY	Fluoride : 0.500, Iron : 0.500, Chloride : 450...
4	29	568	5521	608526	Gudumagatta	Karnataka	Bhadravati	Shimoga	KARNATAKA	SHIMOGA	BHADRAVATI	GUDAMANGATTA	Fluoride : 1.000, Iron : 0.400, Chloride : 400...
...
1206	32	598	5674	628230	Karumady	Kerala	Ambalappuzha	Alappuzha	KERALA	ALAPPUZHA	AMBALAPUZHA	KARUMADY	Fluoride : 0.990, Iron : 0.100, Chloride : 25...
1207	32	598	5674	628230	Karumady	Kerala	Ambalappuzha	Alappuzha	KERALA	ALAPPUZHA	AMBALAPUZHA	KARUMADY	Fluoride : 0.200, Iron : 0.050, Chloride : 15...
1208	32	598	5674	628230	Karumady	Kerala	Ambalappuzha	Alappuzha	KERALA	ALAPPUZHA	AMBALAPUZHA	KARUMADY	Fluoride : 0.200, Iron : 0.150, Chloride : 25...
1209	32	598	5674	628230	Karumady	Kerala	Ambalappuzha	Alappuzha	KERALA	ALAPPUZHA	AMBALAPUZHA	KARUMADY	Fluoride : 0.230, Chloride : 25.000, TDS : 175...
1210	32	598	5674	628230	Karumady	Kerala	Ambalappuzha	Alappuzha	KERALA	ALAPPUZHA	AMBALAPUZHA	KARUMADY	Fluoride : 0.690, Iron : 0.050, Chloride : 210...

1211 rows × 13 columns

Figure 4: Now using the Matched column, The Reference Csv and User's csv are merged

	pc11_s_id	pc11_d_id	pc11_sd_id	pc11_tv_id	tv_name	State Name	sd_name	d_name	StateName	DistrictName	BlockName	VillageName	Contaminents	shrid2
0	29	568	5521	608518	Diggenahalli	Karnataka	Bhadravati	Shimoga	KARNATAKA	SHIMOGA	BHADRAVATI	DIGGENAHALLI	Iron : 0.500, Chloride : 450.000, Nitrate : 55...	11-29-568-05521-608518
1	29	568	5521	608522	Anaveri	Karnataka	Bhadravati	Shimoga	KARNATAKA	SHIMOGA	BHADRAVATI	ANAVERI	Fluoride : 1.000, Iron : 0.400, Chloride : 300...	11-29-568-05521-608522
2	29	568	5521	608522	Anaveri	Karnataka	Bhadravati	Shimoga	KARNATAKA	SHIMOGA	BHADRAVATI	ANAVERI	Fluoride : 0.500, Iron : 0.500, Chloride : 450...	11-29-568-05521-608522
3	29	568	5521	608523	Ittigehalli	Karnataka	Bhadravati	Shimoga	KARNATAKA	SHIMOGA	BHADRAVATI	ITTIGEHALLY	Fluoride : 0.500, Iron : 0.500, Chloride : 450...	11-29-568-05521-608523
4	29	568	5521	608526	Gudumagatta	Karnataka	Bhadravati	Shimoga	KARNATAKA	SHIMOGA	BHADRAVATI	GUDAMANGATTA	Fluoride : 1.000, Iron : 0.400, Chloride : 400...	11-29-568-05521-608526
...
1206	32	598	5674	628230	Karumady	Kerala	Ambalappuzha	Alappuzha	KERALA	ALAPPUZHA	AMBALAPUZHA	KARUMADY	Fluoride : 0.990, Iron : 0.100, Chloride : 25...	11-32-598-05674-628230
1207	32	598	5674	628230	Karumady	Kerala	Ambalappuzha	Alappuzha	KERALA	ALAPPUZHA	AMBALAPUZHA	KARUMADY	Fluoride : 0.200, Iron : 0.050, Chloride : 15...	11-32-598-05674-628230
1208	32	598	5674	628230	Karumady	Kerala	Ambalappuzha	Alappuzha	KERALA	ALAPPUZHA	AMBALAPUZHA	KARUMADY	Fluoride : 0.200, Iron : 0.150, Chloride : 25...	11-32-598-05674-628230
1209	32	598	5674	628230	Karumady	Kerala	Ambalappuzha	Alappuzha	KERALA	ALAPPUZHA	AMBALAPUZHA	KARUMADY	Fluoride : 0.230, Chloride : 25.000, TDS : 175...	11-32-598-05674-628230
1210	32	598	5674	628230	Karumady	Kerala	Ambalappuzha	Alappuzha	KERALA	ALAPPUZHA	AMBALAPUZHA	KARUMADY	Fluoride : 0.690, Iron : 0.050, Chloride : 210...	11-32-598-05674-628230

1211 rows × 14 columns

Figure 5: Finally using the Census Ids the SHRID column is formed

	shrid2	elevation_mean	elevation_median	elevation_percentile_5	elevation_percentile_25	elevation_min	elevation_max	elevation_num_cells	elevation_std
0	11-01-001-00001-000002	1758.410648	1745.0	1339.00	1585.00	1276.0	2406.0	8321	251.390562
1	11-01-001-00001-000005	2399.254718	2395.0	1944.10	2213.00	1823.0	3081.0	6623	268.828716
2	11-01-001-00001-000006	3176.797998	3206.0	2742.00	2953.00	2597.0	3651.0	6594	261.233354
3	11-01-001-00001-000007	2846.648941	2875.0	2484.75	2690.75	2363.0	3255.0	4156	203.604353
4	11-01-001-00001-000008	2825.052265	2830.0	2575.00	2714.00	2482.0	3123.0	3444	148.466248
...
576450	11-35-640-05924-645566	27.163448	25.0	9.00	19.00	0.0	69.0	7715	12.523217
576451	11-35-640-05924-645567	126.060891	134.0	57.00	116.00	28.0	175.0	28658	27.959493
576452	11-35-640-05924-645568	43.306925	32.0	8.00	21.00	0.0	120.0	10022	30.359780
576453	11-35-640-05924-645569	21.714302	18.0	6.00	9.00	0.0	89.0	8544	17.364967
576454	11-35-640-05924-645570	115.598222	118.0	73.00	101.00	20.0	164.0	25086	25.045453

576455 rows × 9 columns

Figure 6: For demonstration, we here use this Elevation SHRUG Data

pc11_s_id	pc11_d_id	pc11_sd_id	pc11_tv_id	tv_name	State Name	sd_name	d_name	StateName	DistrictName	...	Contaminants	shrid2	elevation_mean	elevation_median	elevation_percentile_5	elevation_percentile_25	
0	29	568	5521	Diggenahalli	Karnataka	Bhadravati	Shimoga	KARNATAKA	SHIMOGA	...	Iron : 0.500, Chloride : 450.000, Nitrate : 55...	11-29-568-05521-608518	614.478064	611.0	585.0	595.0	
1	29	568	5521	Anaveri	Karnataka	Bhadravati	Shimoga	KARNATAKA	SHIMOGA	...	Fluoride : 1.000, Iron : 0.400, Chloride : 300...	11-29-568-05521-608522	577.343785	572.0	563.0	567.0	
2	29	568	5521	Anaveri	Karnataka	Bhadravati	Shimoga	KARNATAKA	SHIMOGA	...	Fluoride : 0.500, Iron : 0.500, Chloride : 450...	11-29-568-05521-608522	577.343785	572.0	563.0	567.0	
3	29	568	5521	Ittigehalli	Karnataka	Bhadravati	Shimoga	KARNATAKA	SHIMOGA	...	Fluoride : 0.500, Iron : 0.500, Chloride : 450...	11-29-568-05521-608523	613.079022	607.0	586.0	596.0	
4	29	568	5521	Gudumagatta	Karnataka	Bhadravati	Shimoga	KARNATAKA	SHIMOGA	...	Fluoride : 1.000, Iron : 0.400, Chloride : 400...	11-29-568-05521-608526	600.657229	597.0	584.0	592.0	
...	
1204	32	598	5674	628230	Karumady	Kerala	Ambalappuzha	Alappuzha	KERALA	ALAPPUZHA	...	Fluoride : 0.990, Iron : 0.100, Chloride : 25...	11-32-598-05674-628230	1.085698	1.0	-4.0	-1.0
1205	32	598	5674	628230	Karumady	Kerala	Ambalappuzha	Alappuzha	KERALA	ALAPPUZHA	...	Fluoride : 0.200, Iron : 0.050, Chloride : 15...	11-32-598-05674-628230	1.085698	1.0	-4.0	-1.0
1206	32	598	5674	628230	Karumady	Kerala	Ambalappuzha	Alappuzha	KERALA	ALAPPUZHA	...	Fluoride : 0.200, Iron : 0.150, Chloride : 25...	11-32-598-05674-628230	1.085698	1.0	-4.0	-1.0
1207	32	598	5674	628230	Karumady	Kerala	Ambalappuzha	Alappuzha	KERALA	ALAPPUZHA	...	Fluoride : 0.230, Chloride : 25.000, TDS : 175...	11-32-598-05674-628230	1.085698	1.0	-4.0	-1.0
1208	32	598	5674	628230	Karumady	Kerala	Ambalappuzha	Alappuzha	KERALA	ALAPPUZHA	...	Fluoride : 0.690, Iron : 0.050, Chloride : 210...	11-32-598-05674-628230	1.085698	1.0	-4.0	-1.0

1209 rows x 22 columns

1209 rows × 22 columns

Figure 7: Using the Shrid2 we combine the user data finally with shrug data

3 How Fuzzy Matching is performed

Fuzzy Matching involves 4 steps

1. **Tokenization** The strings are split into smaller units (tokens), such as words or characters. This helps in comparing the components of the strings more effectively.
2. **Normalization** The strings are normalized by converting them to a common case (e.g., all lowercase) and removing any non-alphanumeric characters.
3. **Scoring** The similarity between strings is quantified using a scoring mechanism. `fuzzywuzzy` uses the Levenshtein distance to calculate the differences between sequences. The Levenshtein distance is the minimum number of single-character edits (insertions, deletions, or substitutions) required to change one word into the other.
4. **Ratio Calculation** The similarity score is then converted into a ratio, usually expressed as a percentage. This ratio reflects the degree of similarity between the two strings

Example

Tokenization and Normalization

- **Original Strings:**
 - string1: "Apple Inc."
 - string2: "Apple Incorporated"
- **After Normalization:**
 - string1: "apple inc"
 - string2: "apple incorporated"

Levenshtein Distance Calculation

The Levenshtein distance between "apple inc" and "apple incorporated" involves calculating the minimum number of edits needed to transform one string into the other. This can be done via dynamic programming, where we build a matrix to track the minimum edits.

Ratio Calculation

The ratio is calculated based on the Levenshtein distance and the lengths of the strings. For example, if the distance is 12 and the length of the longer string is 18, the similarity ratio might be calculated as:

$$\text{ratio} = \left(1 - \frac{\text{distance}}{\max(\text{len}(\text{string1}), \text{len}(\text{string2}))}\right) \times 100 \quad (1)$$

Finding the Best Match

The `process.extractOne` function tokenizes and normalizes the input string and the choices. It calculates the similarity ratio for each choice. The choice with the highest similarity ratio is returned as the best match.

4 Appendix

Making of the Reference File

```
## Village shapefile acts as base. To this csv we keep adding State,  
## district and Subdistrict names as codes are already present.
```

```
import pandas as pd  
import geopandas as gpd  
shapefile_path = 'shrug-pc11-village-poly-shp/village.shp'  
tv_name_gdf = gpd.read_file(shapefile_path)
```

```
tv_name_gdf = tv_name_gdf.drop('geometry', axis=1)  
tv_name_gdf
```

	pc11_s_id	pc11_d_id	pc11_sd_id	pc11_tv_id	tv_name
0	01	001	00001	000001	Bore
1	01	001	00001	000002	Keran
2	01	001	00001	000003	Bugna
3	01	001	00001	000004	Bichwal
4	01	001	00001	000005	Mindiyan
...
649613	35	640	05924	645567	Butler Bay Forest Camp 4-IV (FDCA)
649614	35	640	05924	645568	Red Oil Palm (Nursery Camp)
649615	35	640	05924	645569	Butler Bay Forest Camp 4-II (FDCA)
649616	35	640	05924	645570	Butler Bay Forest Camp 4-I (FDCA)
649617	01	000	00000	000000	None

649618 rows × 5 columns

(a)

District name with code

```
import pandas as pd  
import geopandas as gpd  
shapefile_path = 'shrug-pc11dist-poly-shp/district.shp'  
d_name_gdf = gpd.read_file(shapefile_path)
```

```
d_name_gdf = d_name_gdf.drop(['geometry', 'pc11_s_id'], axis=1)  
d_name_gdf
```

	pc11_d_id	d_name
0	468	Kachchh
1	469	Banas Kantha
2	470	Patan
3	471	Mahesana
4	472	Sabar Kantha
...
636	587	Lakshadweep
637	638	Nicobars
638	639	North & Middle Andaman
639	640	South Andaman
640	000	None

641 rows × 2 columns

(c)

Adding the State Name

1. We prepare a mapping and then use it to make a state name column in Base csv
To get codes for states we refer census 2011 code from govt website as Everyother
shapefiles names and codes are from 2011 shapefiles

```
state_names = [  
    "Jammu and Kashmir", "Himachal Pradesh", "Punjab", "Chandigarh", "Uttarakhand",  
    "Haryana", "NCT of Delhi", "Rajasthan", "Uttar Pradesh", "Bihar",  
    "Sikkim", "Arunachal Pradesh", "Nagaland", "Manipur", "Mizoram",  
    "Tripura", "Meghalaya", "Assam", "West Bengal", "Jharkhand",  
    "Odisha", "Chhattisgarh", "Madhya Pradesh", "Gujarat", "Daman & Diu",  
    "Dadra & Nagar Haveli", "Maharashtra", "Andhra Pradesh", "Karnataka", "Goa",  
    "Lakshadweep", "Kerala", "Tamil Nadu", "Puducherry", "Andaman & Nicobar Islands"  
]  
pc11_s_id = [  
    "01", "02", "03", "04", "05",  
    "06", "07", "08", "09", "10",  
    "11", "12", "13", "14", "15",  
    "16", "17", "18", "19", "20",  
    "21", "22", "23", "24", "25",  
    "26", "27", "28", "29", "30",  
    "31", "32", "33", "34", "35"  
]
```

```
state_code_df = pd.DataFrame({  
    'State Name': state_names,  
    'pc11_s_id': pc11_s_id  
})
```

state_code_df

(b)

Subdistrict names with code

```
import pandas as pd  
import geopandas as gpd  
shapefile_path = 'shrug-pc11subdist-poly-shp/subdistrict.shp'  
sd_name_gdf = gpd.read_file(shapefile_path)
```

```
sd_name_gdf = sd_name_gdf.drop(['geometry', 'pc11_s_id', 'pc11_d_id'], axis=1)  
sd_name_gdf
```

	pc11_sd_id	sd_name
0	00000	Rann Of Kachchh
1	03722	Lakhpat
2	03723	Rapar
3	03724	Bhachau
4	03725	Anjar
...
5964	05921	Rangat
5965	05923	Port Blair
5966	05924	Little Andaman
5967	05922	Ferrargunj
5968	00000	None

5969 rows × 2 columns

(d)

Merging all with base csv

```
tv_name_gdf = tv_name_gdf.merge(state_code_df, on='pc11_s_id', how='outer')
tv_name_gdf = tv_name_gdf.merge(sd_name_gdf, on='pc11_sd_id', how='outer')
tv_name_gdf = tv_name_gdf.merge(d_name_gdf, on='pc11_d_id', how='outer')
tv_name_gdf
```

	pc11_s_id	pc11_d_id	pc11_sd_id	pc11_tv_id	tv_name	State Name	sd_name	d_name
0	01	001	00001	000001	Bore	Jammu and Kashmir	Kupwara	Kupwara
1	01	001	00001	000002	Keran	Jammu and Kashmir	Kupwara	Kupwara
2	01	001	00001	000003	Bugna	Jammu and Kashmir	Kupwara	Kupwara
3	01	001	00001	000004	Bichwal	Jammu and Kashmir	Kupwara	Kupwara
4	01	001	00001	000005	Mindiyan	Jammu and Kashmir	Kupwara	Kupwara
...
654815	35	640	05924	645567	Butler Bay Forest Camp 4-IV (FDCA)	Andaman & Nicobar Islands	Little Andaman	South Andaman
654816	35	640	05924	645568	Red Oil Palm (Nursery Camp)	Andaman & Nicobar Islands	Little Andaman	South Andaman
654817	35	640	05924	645569	Butler Bay Forest Camp 4-II (FDCA)	Andaman & Nicobar Islands	Little Andaman	South Andaman
654818	35	640	05924	645570	Butler Bay Forest Camp 4-I (FDCA)	Andaman & Nicobar Islands	Little Andaman	South Andaman
654819	NaN	NaN	03265	NaN	NaN	NaN	Gharghoda	NaN

654820 rows × 8 columns

Figure 9: Reference CSV prepared