**Documentation on the LGD Mapping of Processed Data Associated with NREGA Assets (State-wise)**

Process: *lgd\_mapping*

*Author: Ewan Nikhil Thomas*

External Files Used: Corrected\_St\_Dt\_SbDt\_Bk\_Vllg\_LGD\_Codes7.9.21 (in “data/external”) – the block level sheet in this file was moved to a new excel workbook namely, “block\_lgd\_index.csv” and called separately for LGD mapping.

This script conducts the LGD Mapping of States, Districts and Blocks for the NREGA data set. The script attempts to LGD Map the names of states, districts and blocks in three stages:

1. Exact String Match – the names of state, district and block for a n observation is concatenated together to make a single variable called “state\_dist\_block”. A similar variable by the same process has been created in the external LGD file. A simple pandas merge command attempts to merge both state processed data file and external LGD file by the variable “state\_dist\_block” (Key variable). The merge is successful only for cases where the string in the key variable matches exactly in both files.
2. Fuzzywuzzy String Match – After the exact string match, the unmatched blocks are identified and processed through a fuzzywuzzy string match (for the key variable) with block names in the external LGD file.

All matches with a fuzz ratio (or match percentage) of 90 percent or greater are corrected in the state data file by replacing the corresponding value with the key variable value in the external LGD file and merged again using the protocol for Exact String Match. In order to achieve this, the result from fuzzywuzzy string matching is created as Pandas data frame and a “mapper dictionary” for the key variable was created. In the mapper dictionary, the ‘key’ is the value in state file (for state\_dist\_block) and ‘value’ is the corresponding observation in the external lgd file which was identified by the fuzzywuzzy process.

Thus, the second stage also ends with another round of first stage exact string matching.

The name sand details of the unmatched blocks are exported as csv to the “data/interim” subdirectory bearing the name, “not\_lgd\_after\_fuzzy.csv”.

1. Manual Mapping – After fuzzywuzzy mapping, there are 314 blocks to be manually mapped. In order to map these 314 blocks, 5 separate functions were created for each specific reason of not being mapped dealt in 3 levels. The functions are:

i. Manual Mapping Level 1 – State level name changes

a. name\_cleaner: function to clean certain district and state level name changes and reformations created by administrative decisions.

ii. Manual Mapping Level 2 – District level name changes

a. dist\_name\_rectify: function to rectify errors associated with inputting data and spelling errors associated with names of districts

b. dist\_bifur: function to correct for district bifurcations and trifurcations undertaken by concerned administrative bodies and cases where blocks in a new district is still being associated to old district in the data.

iii. Manual Mapping Level 3 – Block level name changes

a. block\_mapper: function to correct differences in block names compared to lgd mapper file and these differences are strictly associated with spelling errors, input errors and administrative decisions for name changes.

b. block\_mapper2: function to identify blocks which split up or hand villages transferred to another block or has received villages or panchayats from other villages. These blocks have to treated by inspecting the village names or local body names nested within it and then rectified using the updated status of blocks and villages as present in the lgd mapper file.

After manual mapping exercise, all the state level processed files are updated with the names of state, district, and blocks as per the lgd mapper file. Also, the respective lgd codes are added to the existing data as 3 new columns at the beginning of the data set. These updated files are stored in a subdirectory called “lgd\_mapped” in the folder path “./data/processed/”.

All specific details about all manual mappings that have been done, which are exhaustingly detailed for the scope of this document has been explained with the note by its code in the concerned python script.

The script conducts the first two steps by primarily loading the external lgd file and then running each state processed data file in a functional loop. Special attention has been given to the state of Andaman and Nicobar, which uses the name “ANDAMAN AND NICOBAR ISLANDS” in the external lgd file. So, calling the file and Boolean conditioning using the “path.stem” function of Pathlib does not work correctly for this state.

PS: The LGD mapper file was corrected as part of this exercise. The blocks “Khirzabad” and “Mustafabad” in Yamunanagar district of Haryana was renamed as “Pratap Nagar” and “Saraswati Nagar” by the government. This change was made in the LGD mapper file.