Does rural local government spending depend on ethnic composition? Evidence from India

Ayush Lahiri[a,c,1], **Brian Holland**[b,1,2], **and Himangshu Kumar**[a]

[a]Georgetown University; [b]Georgetown University; [c]Georgetown University

This manuscript was compiled on December 7, 2022

Local public good levels are often dependent on the ethnic fragmentation of the area (Alesina et al, 2003). We use data on village level public goods in the Indian state of Odisha, and merge it using probabilistic string matching to a dataset of village expenditures. Our matching technique is able to overcome spelling differences, and increase matches by more than 40 percentage points compared to exact matching. Using this dataset, we find that public good expenditures by elected village councils have a negative relation with its fractionalization index. More heterogeneous areas have lower spending per capita. This extends prior work by Banerjee and Somanathan (2007).

ethnic fractionalization | decentralization | rural local funding | ...

Decentralisation, Public Goods and Social Division

A. Decentralisation in developing countries.

By the end of the 20th Century, the World Bank estimated that the majority of developing nations had enacted measures to decentralize governance (Crook and Manor 1998). India undertook such a large-scale reform in 1993, that required state governments to devolve ad-ministrative and fiscal responsibility for provision of local public goods and social services to the village level, to councils known as Gram Panchayats or GPs (Kochar et al, 2009). Ideally, such measures towards local self-government should result in fiscal expenditures better reflecting the needs or attributes of the people affected by them, especially marginalized groups (Johnson,2003).

Gram Panchayats (GPs) are village-level councils that predated colonial-era India. They were recognized as a legitimate level of government through a Constitutional Amendment in 1993, to be constituted through elections. A number of measures for political affirmative action were also built into the legislation. A third of seats on the council are reserved for women. As of 2020, 52 percent of all elected representatives in GPs were women in Odisha. Seats were similarly reserved for Scheduled Castes (SC) and Scheduled Tribes (ST) candidates, based on their demographic shares. (Bardhan, 2010). SCs and STs were both part of India's historically marginalized groups.

B. Social divisions and public goods in India.

However, STs were indigenous tribal communities, historically inhabiting remote places with dense forest cover. Different legal structures apply to areas with majority tribal populations, which are classified as Schedule 5 areas. The Central government can overrule State government decisions in these areas; preserved from colonial legislation that sought to balance the interests of tribal groups against the states. Seats in GPs in Schedule 5 areas are reserved for ST candidates only. The state chosen for this study, Odisha, is one of India's poorest states, ranked 32 of 36 states by Human Development Indicator in 2020. It has large mineral reserves of coal and steel, often in forested areas inhabited by ST populations, including several Schedule 5 areas. The most important source of funds for GPs is the Central government. The Central Finance Commission, a statutory body constituted every five years, specifies fund allocations between Central, State and local governments.The 14th Finance Commission recommended a grant of around 27 billion USD to GPs. There are also program-specific funds that GPs might get. These funds are called 'tied' funds – their receipt is tied to conditions on how the funds can be used. E.g. PMAY money can only be used for building certain types of houses. The period from 2015-2019 coincided with a nationwide central

Significance Statement

Fiscal decentralization is a powerful tool to bridge inequities between groups in developing countries. This paper uses public good and expenditure data from more than 5,800 village councils in an Indian state, to show that higher ethnic fragmentation is associated with lower expenditure.

push to spend on water and sanitation assets. (Centre for Policy Research, 2019). Loans and grants from state governments also account for a significant chunk of funds. Finally, income generation from community-owned assets and tourism accounts for a very small share of total expenditures. From 2015, GPs were mandated to prepare development plans (GPDPs), laying out the expenditure plan for the upcoming financial year in order to receive federal funds. These GPDPs were to be prepared in a participatory manner, with village-level assemblies to receive input on what the Panchayat should spend on. These plans were then to be uploaded to a government portal before December 31 of each year. This forms our key dataset of analysis.

Social heterogeneity influences availability of public goods. Especially in societies where resource allocation is not automatic or rule bound. (Alesina Et al, 200). Some groups have more bargaining power, and ability to get resources from the state.Prior evidence shows that the social fractionalization has mixed results on its association with public goods in Indian villages. However, ST settlements are more disadvantaged.(Banerjee and Somanathan, 2007). However, there is relatively lesser research on the effect of ethnic composition on local development expenditures.

Research Question. In this context, our study seeks to examine whether any relation exists between public good expenditures and the ethnic composition in Indian village, specifically in our case;the state of Odisha.

Data Sources

GP-level expenditure data. Filename: 'exp_t4.csv'

- **Source**: Public expenditure plans uploaded by GPs to government portal, compiled by researchers (Bhatia and Leighton, 2022). A plan consists of multiple activities, each classified into a sector.
- **Time Window**: Financial year 2015-2016 to FY 2022-2023. We examine the entire time range.
- **Original unit of analysis**: GP-Year-activity
- **Used unit of analysis**: Aggregated to GP-Year-Sector level.
- **Limitations of the data source**: The time gap presents a considerable issue. In the period from 2011-2022, new districts, blocks, GPs were created and older ones renamed.
- **Definitions for key variables**:
  - **Identifying variables**: District, block and GP names
  - **Activity**: Text description of an activity to be carried out in the GP's boundaries. It has been approved and funds have been allocated by the State government.
  - **Sector**: The activities are classified into broad sectors like 'education', 'health', 'GP office infrastructure'.
  - **Estimated cost**: The approved cost estimate, or fund allocation for the activity.
  - **Plan Year**: The financial year (March - April) in which the eactivity will be executed.

Census of India, 2011.
Filename: 'census_infra_od.csv'
We download the village-level amenities file of the Indian census of 2011. This contains data on the levels of local public goods and amenities for villages. A note about administrative boundaries in India. India is divided into states, which are further divided into districts, subdivided into blocks. Each block contains Gram Panchayats (GPs), which consists of multiple villages which share the same elected council. (DIAGRAM HERE)

- **Source**: Census of India 2011, official website.
- **Time Window**: 2011
- **Original Unit of Analysis**: Village (there are multiple villages within each GP)
- **Used unit of analysis**: Aggregate to GP level
- **Limitations of the data source**: The only issue is that this is the most recent, official complete enumeration of village-level public goods and amenities. In the period from 2011 - 2022, new villages were created and names were changed.

Data Processing and Merging

Our project is split into two notebooks, one for cleaning ('01_cleaning.ipynb') and one for analysis ('02_analysis.ipynb').

C. Data processing.

C.1. Selection of constituent variables to create amenity level variables: Census Data . The census data captures multiple indicators of various amenities/public goods in a given village. We focus on 4 main indicators for a village's level of amenities the data-set contains, i.e. Health, Roads, Education infrastructure and availability of electricity measured in terms of hours of power a village has. To find aggregated effects of these variables of interest, we create measures for each of these indicators by summing across their constituent variables (Cleaning notebook section: "Creating groups of variables"). The only way to group these variables was to select them manually, though we tried exploring loops and programmatic ways to capture them. This is because there is no single unique word that identifies all variables in a group. For example, both 'Dispensary doctors total strength' and 'community health centre (numbers)' refers to health related variables.

Examples of the constituent and number of constituent variables for each amenity can be observed in Table 1.

Table 1. Groups of infrastructure variables

Group	Examples of constituent variables	Total No. of columns
Health	Community Health Centre (Numbers), Community Health Centre Doctors Total Strength (Numbers)	**64**
Education	'Govt Middle School (Numbers), Private Middle School (Numbers), Govt Secondary School (Numbers)	**28**
Roads	National Highway (Status A(1)/NA(2)), State Highway (Status A(1)/NA(2))	**7**
Finances	ATM (Status A(1)/NA(2)),Commercial Bank (Status A(1)/NA(2)), Cooperative Bank (Status A(1)/NA(2))	**5**
Power	Power Supply For All Users Summer (April-Sept.) per day (in Hours), Power Supply For All Users Winter (Oct.-March) per day (in Hours)	**2**
TOTAL		**106**

C.2. Summing individual amenity variables into broader groups: Census Data. Constituent variables are of two types:

- Binary (for eg, whether a state highway exists), indicated by "(Status A(1)/NA(2))" in parenthesis next to the original variable name:
  - Originaly in the dataset these variables take value 1 if the amenity exists and 2 otherwise.
  - We recode these variables such that they take value 1 if the amenity exists and 0 otherwise(Cleaning notebook section "Recoding binary") . Re-coded binary variables are then aggregated through addition. *(Cleaning notebook section "Sum Numeric Columns")*
- Numerical (for eg, the number of hospitals/number of schools), indicated by "numbers" in parenthesis next to the original variable name:
  - Numerical variables are aggregated through addition of all constituent variables. (Cleaning notebook section "Sum Numeric Columns")

C.3. Creation of unique IDs: Census Data and Expenditure Data.

- The primary problem our data poses is that villages are not identified by a formalized unique ID like the FIPS code.
- Before creating IDs, we manually rename districts for 4 districts so that no disparities exist in district names between the two IDs. Doing so helps us improve our fuzzy matching results(Cleaning notebook section "Changing district names")
- We create unique IDs by concatenating the district name, block name and the GP name. (Cleaning notebook section: "Create unique IDs in census" + Code Cleaning notebook section "Create unique IDs for expenditure data")

We create IDs with GP names instead of village names, as the expenditure data is captured at the GP level and our analysis is at the GP level. Therefore, in the census data all villages which fall under the jurisdiction of a particular GP, will have the same ID.

C.4. Grouping Data to the GP level: Census and Expenditure.

- We group the **census data by unique ID**, summing all variables, yielding census data at the GP level. (Cleaning notebook section "groupby census villages to GP level")
- We group the **expenditure data by unique ID, sector and plan year**, yielding the total expenditures for a particular GP in a given sector for a particular year. (Cleaning notebook section "Group to sector level")

C.5. Final list of census variables and generating demographics as proportions: Census Data.

The final list of variables we decided to keep from the census data are as follows:

- Unique ID
- Total Geographical Area
- Forest Area
- Total Households
- Total Population of Village
- Total Male Population of Village
- Total Female Population of Village
- Total Scheduled Castes Population of Village
- Total Scheduled Tribes Population of Village
- health_sum (aggregated variable for health)
- educ_sum (aggregated variable for education)
- fin_sum (aggregated variable for banks/financial institutions)
- roads_sum (aggregated variable for roads)
- power_hrs_sum (aggregated variable for hours of electrification)
- n_vill (number of villages for a given ID)

We create amenity variables based on per-capita and percentage of population to ensure greater interpretability (Cleaning notebook section "Generating variables"). These include generation percentage of SC and ST in villages, female to male ratio in a village and forest area as percentage of total area.*

C.6. Miscellanous processing steps.

- We convert all district , block , GP and villages names into lowercase and removing any whitespace in these names, for creating an ID(Cleaning notebook section "Sum numeric columns" + Cleaning notebook section "Creating unique IDs for expenditure data")
- We create village counter variables in census data, in order to have number of villages for a given ID, when grouping by ID to the GP level. (Cleaning notebook section "Create unique IDs in census")
- We also create an activity counter in expenditure data, in order to have the number of activities undertaken in a GP for a given sector in a given year, when grouping by ID,sector and plan year. (Cleaning notebook section "Group to sector level")
- Similar to the previous step we create a variable to track the amount of a GPs own funds used in a given sector for a given plan year, when grouping by ID,setor and plan year. (Cleaning notebook section "Group to sector level")

Fuzzy merging and matching

D. Fuzzy Merging.

Despite creation of unique IDs, the issue we now face is in merging the census and expenditure datasets. The unique IDs generated for the two datasets do not match one to one, due to the following reasons:

- The expenditure data is recorded between 2015 and 2022. During this time new villages or blocks have been created. This cannot be captured in the census data since the most recent census we have access to was conducted in 2011. As there exists no methodical way of solving this other than incorporating the currently unavailable Census 2021 data, our aim are to discard data corresponding to the IDs of such variables.
- The second problem we face and focus on is the issue of variation in spelling for the same blocks and/or GP names. Due to a lack of uniformity in norms in recording village and/or block names, the same village names have been spelt differently in the two datasets.

Therefore even when the IDs in the census are referring to the same GP, they are not "equal" and hence do not match. Since for the second issue, we are essentially dealing with "spelling differences", we can tackle this problem using approximate string matching. The method to do so in our case is textbffuzzy matching.

*For our analysis, we further create measures such as ranks, which we discuss in later sections.

E. Fuzzy matching: Overview.

In approximate string matching, "the objective is to find matches for short strings in many longer texts, in situations where a small number of differences is to be expected." Also known as fuzzy matching, it calculates how different two strings are, by assigning an edit distance to calculate which, the Levenshtein distance is a widely used method. To match between the expenditure and Census datasets, we use the fuzzywuzzy package in python. The package relies on two constituent packages- the Levenshtein distance and difflib.

Levenshtein distance: Also known as the minimum edit distance between two strings. It is the minimum number of editing operations needed to convert one string into another. The editing operations at the simplest level can be inserting, deleting and/or substituting a single symbol. If two strings are more different, more operations are required to convert one to the other, and a higher Levenshtein distance is assigned.

Difflib: This module provides classes and functions for comparing sequences. The basic algorithm is similar to a more widely known one for gestalt pattern matching, by Ratcliff and Obershelp (1983).

Utilizing Levenshtein distances, Fuzzywuzzy has four primary methods which are dependant on the nature of strings being matched. We briefly discuss these methods.

E.1. Fuzzy matching: Four primary methods.

Each of the methods provides a score based on string similarity. This score however, is different from edit distances, in that, it is inversely related to the edit distance. The lower/higher the edit distance, the higher/lower the fuzzywuzzy score. **Essentially this implies that, the higher the fuzzywuzzy score, the higher is the string similarity**.

Each of the four primary methods under fuzzywuzzy serves a different purpose.

- **Fuzz.ratio** : the simplest measure to calculate string similarity. It calculates the edit distance between the strings as they are provided.
  For eg: "New York mets" and "New York meats", would receive a high score due to the difference of only one letter.

- **Fuzz.partial_ratio**: when two strings are of different lengths and the shorter string is length m, and the longer string is length n, it finds the score of the best matching length-m substring.
  For eg: "Yankees" and "New York Yankees" would receive a score of 100 under partial ratio but a much lower score with fuzz.ratio, even though they do refer to the same entity.

- **Fuzz.token_sort_ratio**: when two strings being matched are out of order, then the above two methods both return incorrect low scores. Token sort, tokenizes the strings in question, sorts the tokens alphabetically, and joins them back into a string.
  For eg: "New York Mets vs Atlanta Braves" and "Atlanta Braves vs New York Mets" would correctly receive a high score under token sort ratio, but not with the previous methods.

- **Fuzz.token_set_ratio**: Token set ratio, allows for more flexibility to the token set ratio. It tokenizes and sorts the strings and then finds the intersection of the string. It then checks for string similarity between this intersection, with the remainder of string 1 and similarly for the intersection and remainder of string 2. It finally returns the highest score between the two.
  For eg: In "mariners vs angels" and "los angeles angels of anaheim at seattle mariners"

  Step 1 - Finding intersection of tokenized and sorted strings: intersection = "angels mariners".
  Step 2 - Create first match pair: Match 1 = intersection + remainder of string 1 = "angels mariners vs"
  Step 3 - Create second match pair: Match 2 = intersection + remainder of string 2 = "angels mariners anaheim angeles at los of seattle"
  Step 4: Score computation: The similarity scores between intersection and match 1 and then intersection and match 2 ar computed and we are returned the higher of the two scores.

For our purposes we use the first metric, i.e. fuzz ratio. We do not face issues of partial matches or unordered stings. We further tested both with partial ratio and fuzz ratio and found no differences in results. Due to fuzz ratio's better interpretability, we proceed with the same.

Fuzzy matching: Process.

1. We first find the unique IDs that perfectly match between the census and expenditure and find that there are 2997 such IDs. Since the fuzz ratio would automatically score these as "100", we remove them for better processing time. We then proceed with the resultant Census ID array expenditure ID arrays after removing the perfect matches. (Cleaning notebook section "Processing for fuzzy matching").

2. We set the threshold score for fuzzy matching at >80, discarding all matches below 80. We found through trial and error that these scores had only spurious matches and the IDs receiving these scores were for villages that were created after the 2011 Census was done and hence exist only in the expenditure data (Cleaning notebook section: “Fuzzy merging”).

3. Through parallel processing for faster process time, we use a list comprehension which does the following: for each unique ID in the expenditure dataset, gives a corresponding match score with each unique ID in the Census dataset (and discarding all below 80 as noted previously). We store all resulting match pairs and corresponding scores in a dataframe(Cleaning notebook section: “Fuzzy merging”). Note: Run time for fuzzy matching was on average <2 minutes.

4. In the scores dataframe attained in the previous step, for each unique expenditure ID, we have multiple matches from census ID and each pair with a different score above 80. Hereon,for each expenditure ID we only keep the match from the Census ID with the highest score. (Cleaning notebook section: “Cleaning Fuzzy scores tables”).

5. We then append the perfect matches to the scores table (in appending the perfect matches the values will be equal in both columns for expenditure ID and census ID and we set the score to 100). (Cleaning notebook section: “Cleaning Fuzzy scores tables”)

6. We find that 3714 unique IDs have been matched. After appending the perfect matches, we find that 6711 IDs in total have matched between the two dataframes. 192 IDs did not get matched between the two dataframes, indicating that they received a score of <80, which implies that these were new villages created after the Census.

7. However, we find through sampling, that for the expenditure IDs whose highest score match pair had a score <93, the matches are spurious. We therefore, drop these ID pairs. (Cleaning notebook section: “Cleaning Fuzzy scores tables”)

8. We merge the two datasets using the score table. First we match the score table to the expenditure data using expenditure ID as the key. Then we merge this new expenditure dataset with the Census data using census ID as the key. (Cleaning notebook section: “Merge Census + Expenditure data”)

Fuzzy Merging: Results.

Using exact matching we were able to match 43% of unique IDs. Fuzzy matching increases our unique ID matches to 84%. In terms of data points, we find that merging with approximate string matching almost doubles the number of rows in the merged dataframe as compared to exact string matching (149k with exact merging to 293k with fuzzy merging, of the total 332k originally in the expenditure data). The results of our fuzzy matching are illustrated in Figure 1.

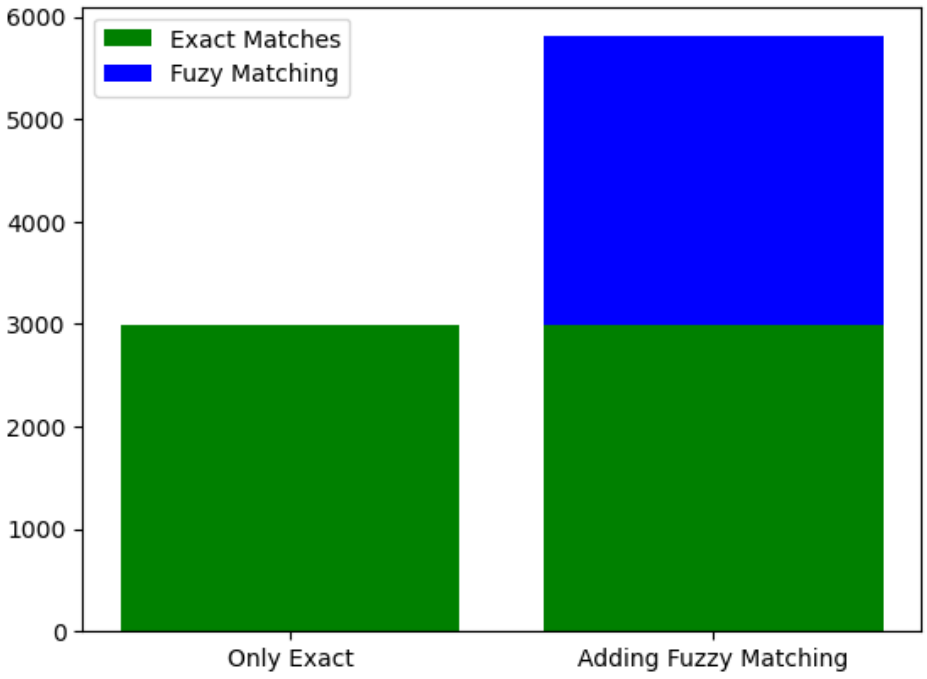

Fig. 1. Fuzzy matching results: How many unique GPs were matched

Analysis variables.

For the purpose of our analysis we generate new variables for greater interpretability and comparison across villages (analysis notebook section 'Variable creation'). The following processing occurs on the data obtained after fuzzy merging the Census and expenditure data. We firstly drop rows where village populations are zero to avoid issues in creating per capita or percentage of population variables. We also replace missing values with their mean values and drop rows

E.2. Amenities.

We generate per capita variables per village population only for amenities which are aggregated based on continuous variables i.e. health and education, and not for variables aggregated based on binaries, i.e. roads and banks/financial institutions. We rank the values for all amenities, for better interpretability of results. Since the amenities are aggregated values, a one unit increase in the amenities is ambiguous. For eg, health is generated based on the sum of hospitals, doctors, para-medics, community health centers and many more. A ranked approach allows us to interpret the "aggregated status" of the village for each amenity.

E.3. Spending.

We also generate per capita spending variables for the pertinent variables on spending, i.e. total spending by GPs across all sectors in a financial year.

E.4. SC/ST status of villages.

We generate population shares of Scheduled Castes and Tribes individuals in each GP. Then,We generate "majority demographic" based variables, i.e. we classify villages as majority SC, majority ST and majority non SC/ST based on whether more than 50% of the total village population is composed of SCs, STs or non-reserved categories.

E.5. GP level fractionalization.

We find that the population of SCs and STs are strongly correlated (we discuss correlations in later sections). As a result quantitative analysis where SC and ST populations will be poor predictors of village level expenditure due to multi-collinearity. We therefore compute the fractionalization of villages based on reservation status utilizing the ethnolinguistic fractionalization measure (ELF) which is computed based on the Herfindahl index of ethnolinguistic group shares.(Alesina Et. al, 2003).Specifically the frationalization of a village is calculated as

$$FRACT_j = 1 - \sum_{n=1}^{N} S_{ij}^2$$

where we compute fractionalization for share of group $i(i$ = SC,ST, NON SC/ST) in village j. The higher the value the more a village is fractionalized. If $FRACT_J = 1$ it implies that the village is not fractionalized and only contains non SC/ST population and when $FRACT_J = 0$, it implies the village is highly fractionalized.

1. Analysis and Results

Demographic settlement patterns.

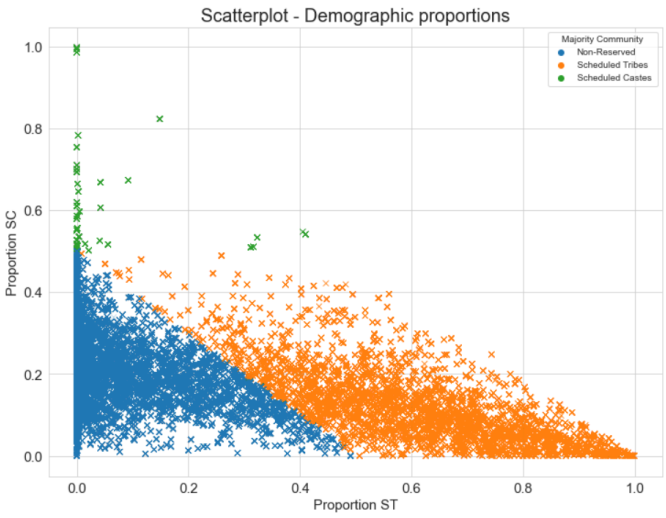

Fig. 2. Percentage of GPs with total number of roads

The distribution of GPs according to majority community shows some distinct patterns. The scatterplot in figure 2 plots the percentage SC on the Y-axis, and percentage ST on the X-axis (analysis notebook section 'Scatterplot - Demographic percentages'). The points are colored by the majority community. The largest share, 61.9% of the GPs have a majority of non-SC and non-ST population. There are 37.2% GPs with with a majority ST population. However, majority SC GPs only constituted 0.8% of the total of 5806 in our final analysis sample. Despite constituting 17.6% of Odisha's total population in 2011, SCs generally do not reside as homogenous majorities in villages.

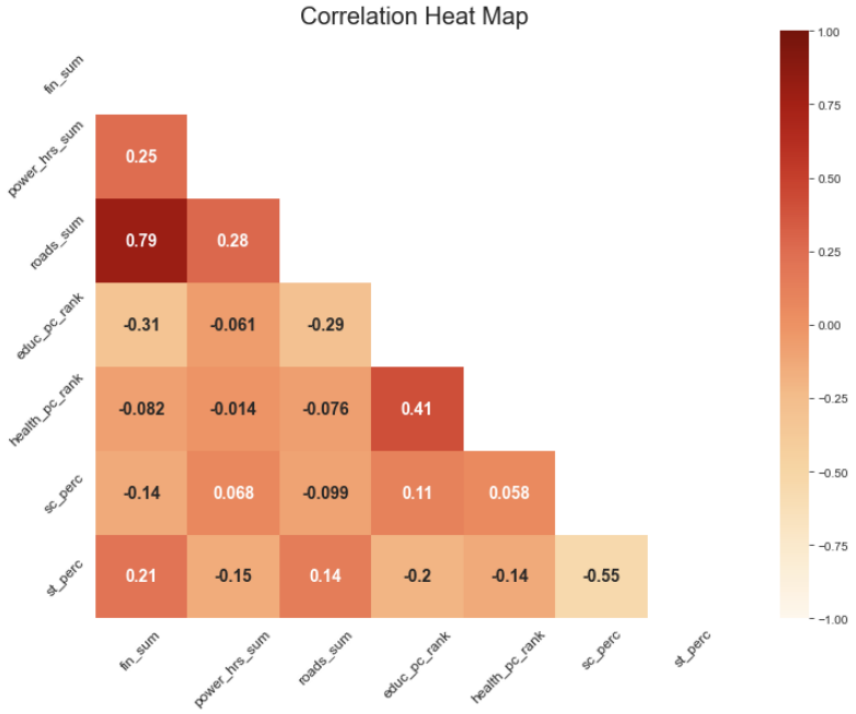

Fig. 3. Correlation between variables of interest

Figure 3 plots the correlation between the amenity levels, as a heatmap colored by the correlation coefficients. We observe high correlation among GP's amenity levels. The number of roads in a GP is highly positively correlated with the number of financial institutions, with a coefficient of 0.79. Importantly, we see significant negative correlation between the percentages of SC and ST populations. This may imply fractionalized settlement patterns among the two groups. It further lends evidence to the issue of multi-collinearity we may see if we attempt to isolate the effect of the groups' specific population percentages on village level expenditures.

Expenditure and Demographics.

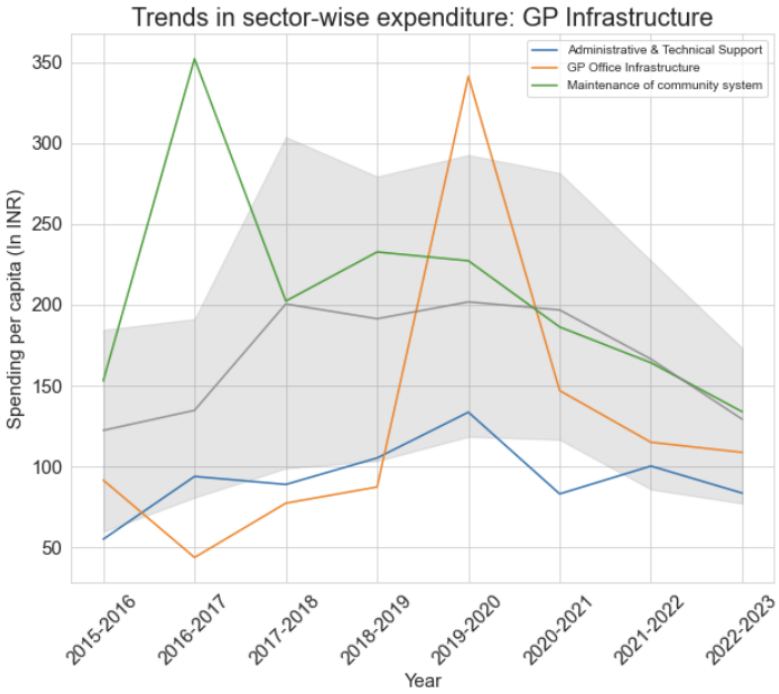

Fig. 4. Trends in spending on a GPs own infrastructure

Figure 4 plot trends in spending for sectors like 'GP office infrastructure' and 'Administrative and technical support', which relate to enhancing the governance capacities of the GP, e.g. by buying personal computers for the Panchayat office (analysis notebook section: 'Spending on public goods vs govt. capabilities'). The grey region in the plot highlights the overall trends in spending on public goods such as drinking water, social welfare, education and health. We see that in some years spending on the GPs own infrastructure is much higher than that on public goods for certain years. The long time frame of the data may indicate that some GPs focus much more on improving GP infrastructure after a period of low investment in these areas. We also see a sharp decline in these trends after COVID.

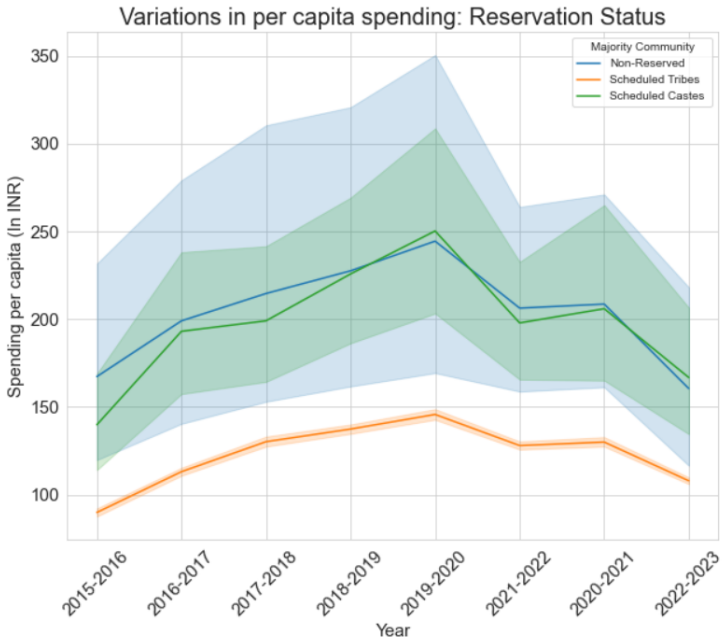

Fig. 5. Trends in spending by SC/ST demographics

Figure 5 plots the average per capita spending in GPDPs. As an instance, the green line represents the mean spending

per-capita received by majority SC villages (analysis notebook section 'Lineplot: spending trends by demographics'). The light green band around it represents the 95% confidence interval for majority SC villages. Put simply, it is 95% of the 'range' of variation in spending seen for majority SC villages.
Two interesting patterns are observed. Majority ST villages (yellow line) made the lowest per-capita GP development expenditures through the entire period with much lesser variation. There was little variation among expenditures of GPs. Majority SC and majority non-reserved GPs do not seem to be too different when compared based on their mean values. In this case as well, we see a sharp decline in per-capita spending after COVID.

A. Regression Analysis.

We aim to see, whether there exists a significant relationship between fractionalization and village level per-capita expenditure (analysis notebook section 'Which sectors saw the highest expenditure?'). We hence estimate the following OLS regression equation:

$$Y = \beta_0 + \beta_1 Fract + beta_2 educ\_pc\_rank + \beta_3 fin\_sum\_rank + \beta_4 health\_pc\_rank + \beta_5 roads\_sum\_rank + \beta_6 power\_hrs\_sum$$

where our independent variable is the log of total per capita spending by a village.Our primary variable of interest is Fract, which takes values between 0 and 1, based on how fractionalized our village is. We further control for ranks of educational amenities, financial institutions, health facilities and roads with coefficients β_1-β_6 respectively and control for the number of power hours a village receives. We also control for sector based fixed effects, since some sectors may be prone to always receiving higher priority or higher funding.

	Dependent variable:
	(1)
const	5.698***
	(0.025)
Fract	-0.152***
	(0.029)
educ_pc_rank	-0.0001***
	(0.000)
fin_sum_rank	0.0001***
	(0.000)
health_pc_rank	-0.00001***
	(0.000)
power_hrs_sum	-0.0005***
	(0.000)
roads_sum_rank	0.00003548***
	(0.000)
Observations	74,808
R^2	0.462
Adjusted R^2	0.461
Residual Std. Error	1.112(df = 74767)
F Statistic	1602.952*** (df = 40.0; 74767.0)
Note:	*p<0.1; **p<0.05; ***p<0.01

We see that for a 1 percentage-point increase in fractionalization there a 16.4% decrease in spending per capita in villages.The results are consistent with the literature and our hypothesis that higher fractionalization does lead to lower public good outcomes, in turn reflected by how much a GP is willing to spend on public goods. This may be due to the village's inability to form coalitions in demanding for better public goods (which in turn is reflected in spending) due to the demographic differences and disagreement/hostility there may exist between communities. Another possible mechanism is through elite capture, in fractionalized villages by the non SC/ST group. Further villages with greater SC/ST populations, may also have the added incentive to demand in coalitions due to their political reservation status, i.e. their GP must necessarily have members from the SC/ST community.

Discussion and limitations:

The primary thrust of this project was to accomplish the fuzzy merging based on village names. We would need to compare the performance of other algorithms, like phonetic matching or Soundex. Our subsequent empirical analysis is preliminary, and needs much more rigorous analysis to offer conclusive evidence. In particular, the high multicollinearity and high dimensionality in our public goods variables requires a statistical solution like Principal Components Analysis (PCA). The dataset used has some limitations. First, we cannot be sure if the villages which could not merge were systematically different from the ones

in our sample. If they are, it would bias estimates. Secondly, we did not distinguish between fund sources. Some funds are mandated at the Central or state levels, and would not depend on ethnic composition of the villages. However, some special funds are specifically earmarked for Scheduled Tribe GPs, especially in Schedule 5 areas.
In support of our findings, even with some exclusive allotted expenditures, Scheduled Tribes get much lower spending per capita from the nationwide GPDP scheme. Some observers argue that the requirement imposed by the Central government, of uploading the GPDP to receive funds is itself against the idea of fiscal decentralization. GP officials might have to travel to the nearest Block administration office to upload them. (Centre for Policy Research, 2020).

ACKNOWLEDGMENTS. The authors would like to thank Dr. Rebecca Johnson, Yifan Liu and Sonali S.R for their patient guidance in shaping this project. They would also like to thank Dr. Kartika Bhatia and Dr. Margaret Leighton for making the data available.

References

Johnson, Craig (2003). Decentralisation in India: Poverty, Politics and Panchayati Raj.Working Paper 199, Overseas Development Institute

Kochar, A., Singh, K., Singh, S. (2009). Targeting public goods to the poor in a segregated economy: An empirical analysis of central mandates in rural India. Journal of Public Economics, 93(7-8), 917-930.

Crook, Richard; Manor, James. Democratic decentralization (English). Operations Evaluation Department (OED) working paper series ; no. 11 Washington, D.C. : World Bank Group.

Alesina, A., Devleeschauwer, A., Easterly, W., Kurlat, S., Wacziarg, R. (2003). Fractionalization. Journal of Economic growth, 8(2), 155-194.

Banerjee, A., Somanathan, R. (2007). The political economy of public goods: Some evidence from India. Journal of development Economics, 82(2), 287-314.

Center for Policy Research, Analysis of Fund Flow to rural local bodies, https://fincomindia.nic.in/writereaddata/html_en_files/fincom15/StudyReports/Analysis%20of%20Fund%20flows%20to%20Rural%20local%20bodies.pdf