Development from Representation?
A Study of Quotas for the Scheduled Castes in India

Online Appendix

Francesca Refsum Jensenius

December 1, 2014

1 Description of Data

The constituency-level estimates of variables from the Indian Census of 1971 and 2001 used in this paper were developed in collaboration with Rikhil Bhavnani. In the following I will briefly describe how the data were created. Further details about the dataset are provided in Bhavnani and Jensenius (forthcoming).

1.1 The creation of constituency-level estimates of 1971 census data

The perfect way of creating constituency-level estimates of the India census data is to aggregate the data from the village level. We first tried to create estimates of data from the 1971 census this way, but this proved difficult—both because of the enormous task of making the village-level data from 1971 electronic and because the delimitation refers to groups of villages that are not listed in census publications. The delimitation report from 1976 specifies the extent of constituencies by mentioning blocks, and sub-block units such as *patwar circles*. The census of India had prepared district-wise booklets for the Delimitation Commission in the 1970s, specifying which villages fell into these sub-blocks. We found a couple of these booklets in the Election Commission record room in New Delhi, but were not been able to locate more of them.

Instead, constituency-level estimates of the census variables were created based on block-level data. The Primary Census Abstract (PCA) books containing block-level data were scanned and then made electronic with the use of text recognition software (OCR). This was done for data from India's 15 largest states: Andhra Pradesh, Bihar, Gujarat, Haryana, Himachal Pradesh, Karnataka, Kerala, Madhya Pradesh, Maharashtra, Orissa, Punjab, Rajasthan, Tamil Nadu, Uttar Pradesh and West Bengal. The OCR were not able to perfectly identify the numbers in the publications, and remaining mistakes were manually cleaned with the help of some 14 research apprentices at UC Berkeley and another 7 research assistants in India. There were many mistakes in the data, and the cleaning work consisted of setting up logical tests for all the data, such as checking that the values for 'male' and 'female' add up to the 'total' values, and that 'rural' and 'urban' add up to the 'total', that the sum of all the values for blocks within the same district add up to the values listed for the whole district, and so on. Where these tests were negative, the numbers were checked up against the original census publications and corrected.

The result was a soft-copy of the data included in the PCA for each of the 3,261 blocks in the 15 largest states in India. The variables in the dataset include population size, literacy rates, employment rates, and occupations. These variables are all available for the whole population, for male and female, for SC/ST and for the rural and urban population in each block.

Once we had block-level data, the next task was to merge it to the constituency level. We started out with block-level census data and the Delimitation Report of 1976, which specifies which blocks fell into each state assembly constituency. There were, on average, approximately 1.5 blocks within the area of a constituency. We had acquired the total population of the constituencies from the internal communication of the Delimitation Commission of India from the 1970s from the Election Commission record room in New Delhi, and we had the population of each block in the census data. The merging files were created by calculating the proportion of the population of a block that fell into a constituency. For example, if the delimitation report listed that Constituency X (population 150.000) consisted of all of block A and part of Block B (each having a population of 100.000), then the census values for Constituency X was calculated as:

$$\text{Est(X)} = 1 * (\text{values Block A}) + 0.5 * (\text{values Block B}) \tag{1}$$

In some cases, two or more constituencies consisted of parts of the same two blocks. In these cases the exact proportions could not be calculated on the basis of the population proportion. We solved these cases in two different ways. In most cases, we used the information about the exact population proportions mentioned in the records of the Delimitation Commission from the 1970s. In the few cases where we could not find written sources among the records of the Delimitation Commission, we made estimates of the population based on the average population size of villages in that region and the number of villages in the constituency.

The implicit assumption behind this aggregation process is that the population has a similar distribution on the various census variables across the split blocks (e.g. that the literacy rate was the same in the part of Block B that fell under constituency X and the remaining part of Block B). This will not always be the case, and our estimates of the various census variables should therefore be expected to deviate somewhat from the true values at the constituency level. However, we have no reason to believe that these inaccuracies are systematic nor that they they correlate with any specific variables, and they should therefore be unbiased in aggregate analyses.

1.2 The creation of constituency-level estimates of 2001 census data

The same procedure of manually matching blocks to constituencies could not be followed for the 2001 census data. In this case, the constituencies had remained unchanged since 1974 and so many blocks and districts had changed borders and names during this time that it was difficult to identify the blocks listed in the delimitation in 2001 census data. Also, there are geocoded (GIS) maps available for the 2001 census blocks that makes it much easier to identify overlapping blocks and constituencies. Such maps were not available for the 1971 census data.

For the 2001 data we started with GIS maps of the 4,208 blocks across the 15 major states in our dataset, and overlaid them with maps of the 3,341 state assembly constituencies within these states. Using ArcGIS's intersect tool, we used these overlapping maps to identify the blocks that fell within each state assembly constituency, along with area-weights for each block. These weights were calculated as the proportion of the land area of the block that fell within a constituency. We used these weights to aggregate the block-level data from the 2001 census to the state assembly constituency level.

A number of checks were run to ensure the quality of the merges. First, the delimitation process ensured that all assembly constituencies fell entirely within administrative districts. We checked to ensure that this was the case in the data as well. Second, since all of India's state assembly constituencies are mutually exclusive and collectively exhaustive, we checked that all block weights across constituencies summed to one. Third, and due to the slightly imperfect overlap of map boundaries, we ignored instances where less than 5% of a block fell within a constituency. The constituency-level estimates created this way relies on the approximate matching of regions. The major implicit assumption behind this aggregation process is that the population is evenly distributed across the land mass of the split blocks.

1.3 Aggregation of village-level data

The Village Directory (VD) of India for 2001 includes village-level information about amities such as electricity, roads, schools, and hospitals. To aggregate these data, I used a merge file created by Asher and Novosad (2012) linking villages to constituencies across India. This merge file was created by overlaying maps with villages coded as points and GIS shape files of state assembly constituencies.

To aggregate the VD information to the constituency level, I summed up the number of people who lived in a village with access to each amenity. For example, if constituency X had six villages, five villages with 100 people in each and no primary school, and one village with 500 people and a primary school, the constituency level estimate would be that 50% of the population in the constituency lived in a village with a primary school. The aggregation was done the same way for the SCs population and for the non-SC population in order to capture the gap in the services available to SCs and others within each constituency. An alternative way of aggregating the data would have been to count the percentage of villages with access to each amenity within a constituency, but this seemed like less valid measure given the great variation in the population across different villages.

The census of India also includes information about amenities in towns, in the Town Directory—a dataset separate from the VD. I chose to only use the VD and not the TD in this analysis for three reasons. First, while villages are small and fall within one constituency, towns and cities are often large and span several constituencies. The GIS maps of towns code them as a point and therefore assigns the whole town or city to the constituency that overlaps with that point. This can create severely biased estimates at the constituency level. Second, the variables recorded in the VD and TD are not the same, thus making it a non-trivial task to merge the data. Third, all cities are generally listed to have electricity, schools, health centers, and similar amenities. Adding the TD data to the VD data would therefore simply add a 100% coverage of these services among the urban population. By using only the VD, I therefore capture the proportion of people within the *rural part* of a constituency that had access to these services, while it can be assumed that all of the urban population in the constituencies lived in towns that had the same services.

2 Illustration of matched pairs of constituencies

Figure 1: Matched pairs of constituencies (448 pairs from 15 Indian states).

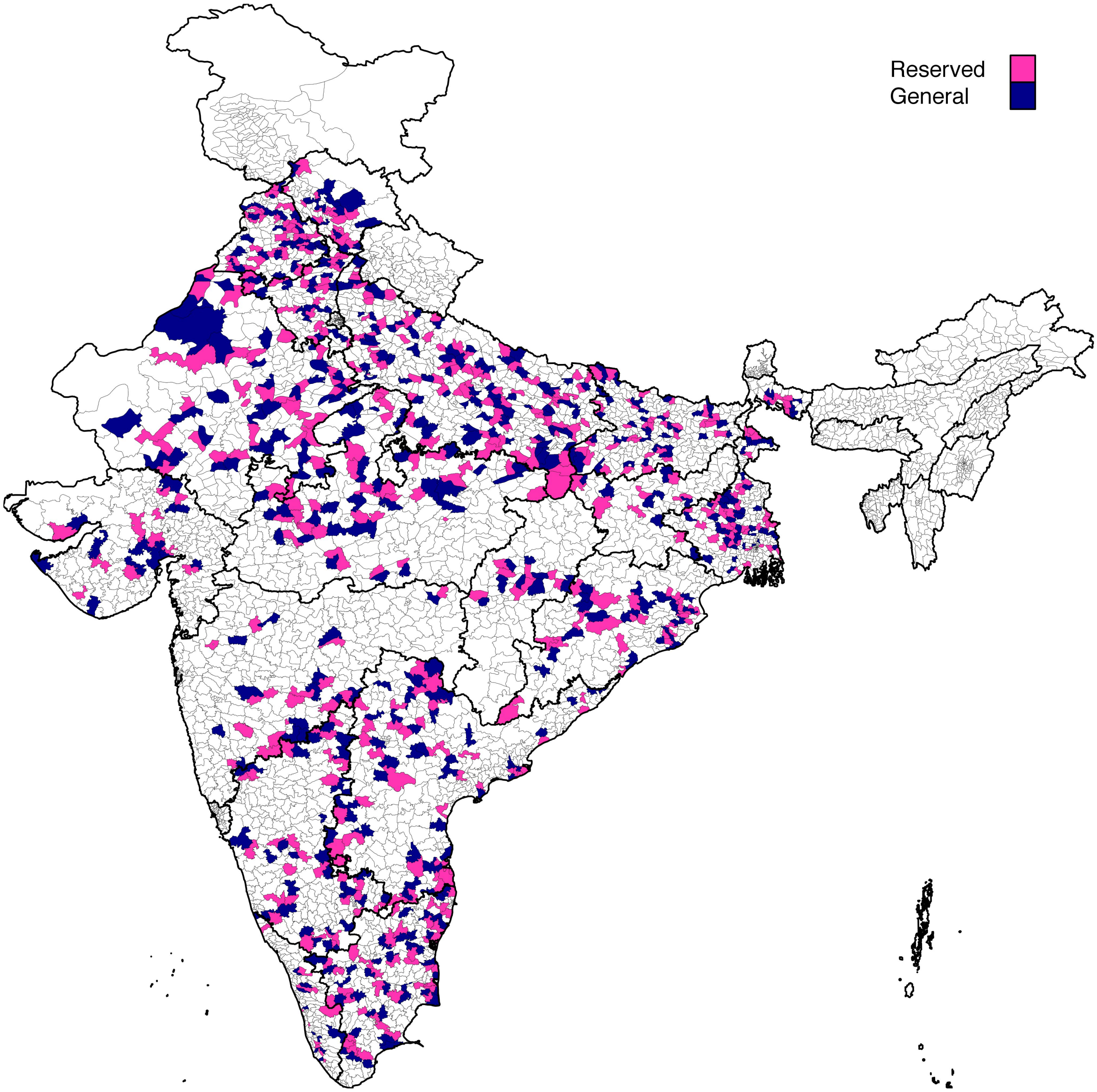
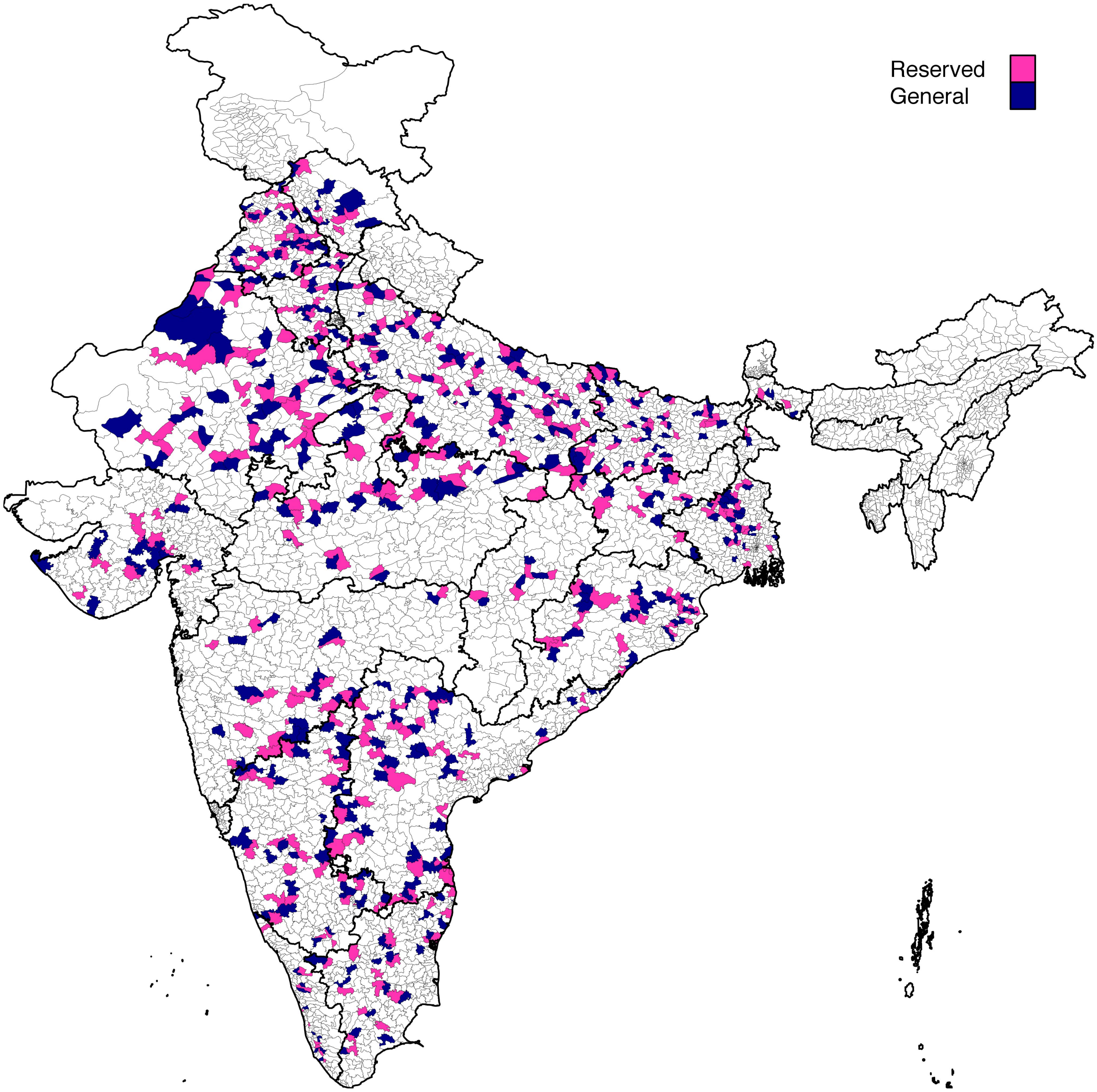


Figure 2: Matched pairs of constituencies (324 pairs from 15 Indian states).



References

Asher, Sam, and Paul Novosad. 2012. “Political Favoritism and Economic Growth: Evidence from India.” Working paper.

Bhavnani, Rikhil, and Francesca Refsum Jensenius. forthcoming. “Socio-economic profiles for India’s old constituencies.” In *Fixing Political Boundaries in India and Its Implications for Political Representation.* , ed. Sanjeer Alam. Oxford University Press.