Determining Politicians' Electorally-Relevant Caste Membership

William O'Brochta[*]
Texas Lutheran University

Sunita Parikh
Washington University in St. Louis

Caste identity is contextually dependent. We focus on electorally-relevant caste identity among politicians — individuals who greatly influence how caste is portrayed in political life. We describe how existing approaches to categorizing caste capture aspects other than its electoral relevance before combining these methods in a systematic way to code caste for this purpose. Our method utilizes government records, name classification, and archival sources to identify electorally-relevant caste. Using information about Delhi municipal corporators, we describe our approach and compare it to existing approaches. We conclude by discussing strategies to align caste coding objectives with methodological techniques and to expand this topic to include more general identity coding tasks.

(5,592 words + 4 inches of tables at 100 words each)

[*] Corresponding Author: 1000 W. Court Street, Seguin, Texas 78155. wobrochta@tlu.edu.

Caste remains an important individual and collective characteristic in India. Caste membership influences social interactions, marriage proposals, government benefits, and political campaigns (Corbridge, Harriss and Jeffrey, 2013). As a result of the myriad contexts in which caste is defined, caste identification can vary across time and locations. These variations occur because the concept and definition of caste have become disassociated from traditional markers, particularly occupational requirements (Beteille, 2012; O'Hanlon, 2017). Caste redefinition can be collectively organized: groups may petition the government for recognition in a reserved category, intentionally refuse work in an occupation traditionally associated with their caste, or become active in a political party most frequently associated with a different caste (Clark-Deces, 2007; Pushpendra, 1999). Caste identity may also be situational, e.g., in a small village where everyone is from the same varna or jati, individuals may identify by their jati or sub-jati respectively (Jodhka, 2004; Sahay, 2004). When researchers seek to identify and categorize caste, they must make intentional choices about how their methodology will produce a measure where comparisons across time and place are still meaningful (e.g., Samarendra, 2016). Yet, these projects have largely not considered how caste identity is used differently in different situations.

In this paper, we focus on what we call "electorally-relevant" caste membership — those caste categories and categorizations that specify how caste is used by politicians in their campaigns and during their time in office. Because socioeconomic and cultural caste categories can differ from political caste categories, we limit ourselves to the electoral context. Prior work seeking to code caste membership of politicians has used a variety of methods, including matrimonial websites, archival research, and expert name classification whose accuracy can be difficult to evaluate (e.g., Jaffrelot and Kumar, 2012). We present an approach to strategically combine these methods in a way that focuses on coding electorally-relevant caste identity among politicians. By combing existing methods, this approach can optimize classification cost, transparency, and ability to reflect electorally-relevant caste membership (Satyanarayana, 2014).

Treating caste as a political identity necessitates a coding approach that recognizes how politicians think about caste membership in an electorally-relevant setting. In this context, politicians are concerned about how constituents and potential voters *perceive* their caste identity. This differentiates electorally-relevant caste membership from self-identification on matrimonial websites and archival identification that seeks to find an individual's self or socially identified caste. Our approach clarifies the ways in which different aspects of caste identity best match onto methodological strategies to code caste. In presenting this approach to coding electorally-relevant caste, we highlight the need for greater clarity on how caste is coded and why specific coding methods are appropriate for a given situation. We proceed to apply our approach to identify the electorally-relevant caste of Delhi municipal corporators, comparing the results to other caste coding methods.

We contribute to ongoing research in caste coding by creating a typology of caste coding approaches and linking them to different ways to conceptualize caste. Our focus on electorally-relevant caste identity adds clarity about our coding objectives and allows us to carefully select caste categories and methods to match this context. Our suggestion for researchers is to define the context in which caste should be measured, to select caste categories to match the context, and to select and order different classification methods to best suit the caste identity of interest.

Defining Caste Context and Selecting Caste Categories

The first step in our proposed method is to select the context in which caste is being classified. Our interest is in classifying politicians in electorally-relevant contexts and making comparisons between politician caste across different Indian states. As such, the caste categories we choose should be electorally-relevant and comparable across different states. Other researchers may have different contexts of interest like social or self-identification and should state this context clearly.

Caste can be defined in different ways (Gupta, 2005; Sengupta, 2010). Thus, researchers need to select a group of caste categories based on the context that they previously identified. In this project, we seek to code electorally-relevant caste among politicians across India.

Caste may refer to the varna classification system, which allocates communities into one of four groupings — Brahmin, Kshatriya, Vaishya, and Shudra (Vaid, 2014, 393). Varnas are hierarchical, and Hindus are assigned to one of the four groups (Sundar, 2000; Waghmore, 2019). There is a fifth category into which Dalits and many tribal communities fall. Varnas are subdivided into jatis, which the term "caste" is also frequently used to describe (Beteille, 1996).[1] Jatis are the most commonly invoked categories of caste identity, especially in everyday social interactions (Jodhka, 2012).

We apply the "field-view" of caste by looking at how caste categories are practiced in electorally-relevant political life (Sahoo, 2017). We start with the national government constructed categories of Scheduled Caste (SC) and Other Backward Class (OBC) categories, alongside a Scheduled Tribe (ST) category (despite the fact that STs fall outside of the typical varna hierarchy) (Rathore, 2020). Because these categories are government constructed, they influence political life including reservations for access to government benefits and political seats (Vaid, 2014, 395). To this categorization, we add a distinction between Brahmin and other forward (OF). We base this distinction on the Indian Human Development Survey by Desai and Vanneman (2015) which demonstrates vast over-representation of Brahmins in political life (Desai and Dubey, 2012).

Finally, we add an other religion category to capture non-Hindu politicians. Other religious groups can and do identify with the caste categories listed above (Sahoo, 2017). This identification is typically state-specific. For example, Muslims have been added to the OBC rolls in some states, but there are no national reservations for Muslims (Alam, 2014). Further, other religions are still portrayed as politically distinct voting blocs (Farooqui, 2020). For example, Ahmed (2022) notes that Muslims have been increasingly politically galva-

[1] Jatis do not fit neatly into varnas and can span multiple varnas. As Beteille (1996, 22) notes, "varna refers primarily to order, the primary reference of jati is to birth and the social identity of birth."

nized by Hindutva policies like the Citizenship Amendment Act.[2] This practice warrants constructing an other religion category in this caste coding context.

The six category system we use — Brahmin, OF, SC, ST, OBC, and other religion — best reflects electoral competition in a way that can be compared across Indian political divisions. Were our interest in one state or municipality, we would need to adopt a categorization system that fits that context. For example, Susewind (2015, 2017) develops a method of categorizing religion from names that can be applied to members of the public in cases where religious identification is of primary interest.

Existing Caste Categorization Approaches

Self-identification, archival research, and expert classification are three popular and sometimes overlapping methods used to code caste. In this section, we describe these methods and discuss how they have been used to code electorally-relevant caste.

Self-Identification

Self-identification involves people describing their own caste membership, usually in a survey. Self-identification is frequently used to categorize caste among non-politicians. Surveys like the Indian Human Development Survey (Desai and Vanneman, 2015) ask respondents to self-identify their caste and the process for doing so has a long history in colonial-era censuses (Gill, 2007; Walby and Haan, 2012). Various studies have used administrative data like these to examine caste-influenced phenomena like residential segregation (Bharathi et al., 2022; Adukia et al., 2019). To our knowledge, scholars have not used self-identification as a measure of the caste composition of politicians to measure electorally-relevant caste, likely due in part to the fact that caste is a particularly controversial census question, despite its appearance in the Socio Economic and Caste Census in 2011 (Omvedt, 2010; Sundar, 2000).

[2] See the Supplemental Information SI.1.

At scale, self-identification data can be used as training data to create algorithmic methods for caste coding. Susewind (2015, 2017) implements such a procedure successfully for religious classification. Obtaining such data for caste categories of interest is often challenging. Apart from Fisman, Paravisini and Vig (2017) using financial records, Indian matrimonial websites provide access to a large database of individuals who have self-identified their caste. Matrimonial websites, which evolved from long-established newspaper classified ads, are a popular way to look for marriage partners in India. Like dating websites, users have profiles where personal information is listed to help potential partners (or their families) determine if an individual is a suitable match. A question about caste identity is usually included. A person making a profile can answer this question however they wish, and it is likely that individuals self-identify in socially desirable ways.

Matrimonial data on caste is useful in that it can be aggregated to determine the relative frequency of caste identification for given surnames. Since many surnames have been historically linked to particular caste categories, a given name will trigger an association with a caste category (Banerjee et al., 2009; Jayaraman, 2005). Vissa (2011) obtained 2.1 million matrimonial profiles from the two largest matrimonial websites and aggregated these profiles to determine the relative frequency of caste identification for given surnames. This dataset was subsequently used in Chen, Chittoor and Vissa (2015) and expanded to 6 million names from three matrimonial websites in Bhagavatula et al. (2022). Damaraju and Makhija (2018) and Rajadesingan, Mahalingam and Jurgens (2019) have independently created their own matrimonial website datasets, each using only one website and fewer profiles than the Vissa (2011) approach. Using matrimonial data is an example of a name-based classification method, wherein an individual's name provides information about their caste.

Matching politicians' surnames to a database of self-identified caste from matrimonial websites assumes that individuals on matrimonial websites are reporting electorally-relevant caste identities. Finding a marriage partner is a social activity, so individuals have incentives to self-identify as a socially desirable caste. As Ahuja and Ostermann (2016) show, inter-caste marriage is a delicate social phenomenon that does not always align with electorally-relevant caste. Therefore, though much work has gone into using self-identification to measure caste, these methods do not align well with our goal of measuring electorally-relevant caste membership among politicians.

Archival Research

Archival research entails trying to find caste information about specific individuals, not just those who happen to share a person's name (Narain and Sharma, 1972). For political actors, this means finding their electorally-relevant caste identity. As such, if we are trying to classify Indian Prime Minister Narendra Modi we would need to find information stating Modi's caste category; we would not rely on any signal that the surname Modi provided or our knowledge of the caste category of other people named Narendra Modi. The need to find caste information about the specific individuals one seeks to classify means that archival research is a time consuming process. Scholars typically hire local experts who know the archival resources available and can efficiently sort through archival material. Newspaper articles, interviews, and government records rarely directly mention the caste of a given politician. Further, even though some politicians hold relatively prominent roles in Indian government, few receive much media attention. This makes finding archival information quite difficult, especially for politicians who served a long time ago, ministers from less influential states, and politicians who held less influential posts (see Lee, 2022).

The challenge with archival research is that different archival sources may identify different aspects of caste identity. Newspaper articles about politics, published political interviews, and government candidate records likely record electorally-relevant caste, since this information is broadcast to the public. Government education records, conversations with neighbors, and newspaper articles prior to political candidacy could record electorally-relevant caste, socially-relevant caste, or some combination of these. Therefore, unless carefully documented and justified, archival research can mix different aspects of caste identity. Ethnographic research methods are an appropriate tool to sort through and process archival material.

Expert Classification

Experts can play a role in studies using self-identification or archival research. However, existing work has also employed experts to operate as their own method of caste categorization. The main idea behind this method is that experts are familiar with caste categorization in a given area. Thus, when given a list of politicians to categorize, experts use their knowledge of caste names (sometimes with additional information provided) to complete the categorization. There is no definition of how one becomes an expert caste coder, but familiarity with demography and naming patterns throughout India is a reasonable prerequisite (Mateos, Webber and Longley, 2007). Researchers have control over the caste categories expert coders are asked to use. Previous work that has employed expert review has revealed little about the process experts used to code caste, making it difficult to assess the accuracy of the method and making replication essentially impossible (Aggarwal, Dreze and Gupta, 2015; Ajit, Donker and Saxena, 2012; Jaffrelot, 1996; Jaffrelot and Kumar, 2012; Krishna, 1966).

Expert classifiers frequently rely on archival research as part of their classification process. Karekurve-Ramachandra and Lee (2020) are a good example of this. While they first hired experts to identify caste using their knowledge of caste-surname patterns, for those politicians with ambiguous names they then asked elected officials, party members, and other prominent individuals for help them by either telling the researchers the caste of politicians or by finding out this information.

Jaffrelot and Kumar (2012) are another prominent example of researchers using a combination of expert name classification and archival methods. In this edited volume, the researchers responsible for each chapter produced caste categorizations either based on their own knowledge of a particular Indian state's elected officials or the knowledge of individuals with whom they consulted. They used this strategy because names can encompass multiple sub-castes and can vary geographically, so while a name-caste category link may be clear in a particular geographic context, it often does not generalize. These methods are described in different ways in different chapters of their book including attributing the coding to an expert (e.g., 306), author preparation (e.g., 34), or fieldwork (e.g., 37). The edited volume is focused on the political implications of caste identity so it is reasonable to assume that experts attempted to code electorally-relevant caste categories, though this is not explicitly stated. Since existing expert-based methods do not provide clear details on how experts were trained to code caste, electorally-relevant caste identity may be being captured alongside other forms of caste identity. These issues arise from problems of opacity — we do not know how the coding method was implemented.

Measuring Electorally-Relevant Caste

To effectively measure electorally-relevant caste identity, researchers need to select classification methods that align, or have the potential to align, with electorally-relevant caste divisions. This is not possible using self-identification because respondents answer how they feel is most appropriate even if researchers attempt to direct them to think about certain aspects of caste identity. Measuring electorally-relevant caste is possible using archival data and expert classification if these procedures are conducted in a way that clearly focuses on finding archival sources or coding caste from a political perspective. Otherwise, knowing that an archival source or expert was used to code caste does not inform a researcher about whether electoral relevance was considered.

We develop an approach designed to measure electorally-relevant caste categories. This approach uses government records, name classification, surname lists, archival research, and machine learning algorithms to perform the categorization. These approaches are ordered by decreasing linkage to electoral relevance: government records are exactly linked to the ways in which politicians compete for electoral seats whereas machine learning algorithms use self-identified matrimonial data. By utilizing approaches more closely linked with electorally-relevant caste first, we attempt to minimize coding caste in a non-electorally relevant context. We introduce our approach and walk through an application classifying electorally-relevant caste using Delhi municipal corporators as our exemplary case.

An Electorally-Relevant Approach

Based on the context in which we want to code caste and our selected categories, we order different caste coding methods to best determine electorally-relevant caste. We have used our caste coding method to calculate electorally-relevant caste for several datasets, including state cabinet minister data from seventeen Indian states from 1977 to 2018 (4,737 ministers) and municipal corporators from twenty-five corporations in five Indian states (406 corporators). For this example, we wish to compare our method with existing data from expert classification. We find such a possibility with municipal corporators in Delhi. Municipal corporators are local elected officials who are responsible for providing local public goods and services (Shah and Bakore, 2006). We focus on the 272 corporators present between 2018 and 2019. Karekurve-Ramachandra and Lee (2020, 767) rely on expert coding and "interviews with various party members and elected officials" to categorize these municipal corporators into twelve caste categories. We will, therefore, compare our classification approach to Karekurve-Ramachandra and Lee (2020) to better discern the significance of focusing on electorally-relevant caste. SI.2 discusses municipal corporators and the Delhi case in more detail.

Government Records

Our approach begins by examining official government records. Not all government records show electorally-relevant caste identities (e.g., education records), but since we are interested in elected politicians, these politicians can qualify to run in caste-reserved constituencies. Forty-six of the 272 corporators did so, meaning that we know the caste category that an individual identifies with politically based on the constituency that they won. It makes little political sense for a politician to run in a caste-reserved constituency while also stating that they identify as another caste.

Table 1 shows how the coding proceeded, starting with government records.

Table 1: Caste Coding Method

Coding Method	Number	Percent	Success Rate
Government Records	46	16.91	16.91
Name Classification	185	66.01	81.86
Surname List	13	4.78	31.71
Archival	18	6.62	64.29
Matrimonial	7	2.57	70.00
Educated Guess	3	1.10	
Total	272		

Government records refers to individuals who were elected in caste reserved constituencies. Name classification was conducted by two specially trained coders. Surname list refers to the Delhi Central List of OBCs. Archival research included searching for basic biographical information about corporators sometimes required searching through family history. Matrimonial data is from Bhagavatula et al. (2022). Percent refers to the percentage of the sample coded using a given method. Success rate refers to the effectiveness of the method, as in the number of successfully coded names out of the total number of names left to code.

Name Classification

After coding the electorally-relevant caste of individuals who appear in government records, we conducted name classification. We choose to implement name classification next because we can ask coders to specifically consider political competition and electoral relevance when completing the coding. As mentioned earlier, name classification is a popular way of coding caste in India. Name classification is inherently linked to electorally-relevant contexts. If a corporator's name is associated with a particular caste category, then it will be difficult for them to disassociate themselves from this category when they run for office (unless they change their name). Therefore, our strategy is to start name classification by asking coders to identify electorally-relevant caste based on names, flagging any names without clear and obvious classifications for further review.

Shah and Davis (2017) use a crowd sourced method that involves hiring online workers to identify racial categories based on names in the United States. We choose to apply this method of name classification instead of the more traditional "expert" classification because prior work demonstrates their comparability to expert methods when carefully implemented (e.g., Benoit et al., 2016).

To find coders, we used a popular freelance website where we hired two coders who worked independently. This allowed us to compare the accuracy of the coders to one another. Coder 1 was a teacher in Bangalore who had broad prior experience living and teaching students throughout India. Coder 2 had experience in translation and market research in both Northern and Southern India. Local knowledge is key to successful name-caste coding, and experiences in different parts of India meant that our coders were quite familiar with electorally-relevant conceptualizations of caste in different Indian states — an important feature in general, but also when coding Delhi corporators, not all of whom have deep historical ties to Delhi.

Our approach differs from most existing name-based coding methods which tend to involve employing academic researchers across Indian states, labeling those coders "experts" (e.g., Jaffrelot and Kumar, 2012). Our approach is also directed specifically toward electorally-relevant caste. Since we chose to train coders ourselves, it was possible to ensure that they clearly understood the coding task and the type of caste coding we were looking for.

We started the name classification process by providing a coder with a list of all unique surnames that remained to be coded in the dataset. Coders were instructed to only classify surnames where the surname clearly indicated electorally-relevant caste affiliation in a way that was unlikely to change over time. For example, Muslim names are often quite distinctive and can be easily coded with only information about individuals' surnames. During this process, the coders flagged all surnames where more information was needed to accurately classify an individual's electorally-relevant caste. Each name coding was accompanied by a confidence level of high (90%+ confident), medium (75%-90% confident), or low (less than 75% confident).[3] Coders were asked to provide remarks describing their rationale for selecting a particular caste category.

After completing surname classification, we took all remaining names and provided coders with the full name of the person to be coded. This information enabled them to perform basic Internet research on the history of certain surnames similar to Damaraju and Makhija (2018)'s approach. Again, the coders provided confidence levels for their coding, and all names not classified with at least a medium level of confidence were left for more intensive review. Cohen's Kappa is 0.695, which is quite high, indicating very good to excellent interrater reliability. We discuss reliability in SI.3.

Archival Research

Remaining individuals to be coded were subjected to a more intensive evaluation of both their surname and of archival records. In particular, we first searched through the Delhi Central List of OBCs to determine whether uncategorized surnames were on this caste list. The list measures electorally-relevant caste.

We then conducted archival research wherein we attempted to find biographical information on unclassified individuals. This task was difficult because of the lack of public visibility of many corporators and an overall unwillingness to print electorally-relevant caste membership in official documents or newspaper articles. The second coder was particularly helpful here in that they had a network of locally-based scholars and journalists whom they contacted for help understanding how corporators had portrayed their caste in the media. Media portrayal is one component of electorally-relevant caste classification.

The coders did not agree on the electorally-relevant classification for ten corporators for whom archival research and the caste list was also unhelpful. These remaining corporators could not be coded using electorally-relevant caste methods. Thus, they were subjected to coding from matrimonial data (Bhagavatula et al., 2022). At best, this is an imperfect solu-

[3] See SI.3.

tion because matrimonial data measures self-identified caste in a socially desirable setting, whereas our objective is to measure electorally-relevant caste. We triangulated coders' responses to provide educated guesses for the final three corporators. Therefore, our method produces successful classification in about 96% of cases.

Correspondence with Expert Classification

After developing a correspondence between our method of measuring electorally-relevant caste and the Karekurve-Ramachandra and Lee (2020) expert method, we find that they agree 85.4% of the time with Cohen's Kappa at 0.78, indicating excellent reliability.

The 85.4% agreement means that 39 corporators were coded differently by the two approaches. The most common discrepancy was that our approach coded ten corporators as OBC that the Karekurve-Ramachandra and Lee (2020) approach coded as OF and six that our approach coded as OF that Karekurve-Ramachandra and Lee (2020) coded as OBC. Our approach resulted in nine Brahmin corporators that Karekurve-Ramachandra and Lee (2020) coded as OF. These discrepancies are not surprising, given that shifts between OBC and OF caste categories are common as groups try to gain or lose recognition on OBC lists. Recall further that our approach is specifically focused on electorally-relevant caste categorization. Therefore, these discrepancies also illustrate how electorally-relevant caste categorization can differ from generally asking experts to classify caste.

Discussion and Conclusion

We demonstrated a method to code electorally-relevant caste categories that are comparable across Indian states. In doing so, we establish three steps that we believe will help to improve caste coding. In implementing these steps, researchers should clearly explain their choices:

1. Context: Clearly identify the context in which caste is being coded and justify how this context relates to the proposed application of the coding (e.g., electoral, social,

economic, self-identification).

2. Categories: Select caste categories that reflect the context and the scope of the comparison of interest (e.g., comparing caste within a village, within a city, within a state, across states, or cross-nationally).

3. Select and Order Methods: Adopt caste coding methods that fit the context and desired categories. Consider utilizing multiple methods that are ordered to balance trade-offs between cost, transparency, and ability to reflect the quantity of interest.

The process of selecting and ordering caste coding methods emphasizes the fact that utilizing a single method is unlikely to code the caste of all individuals at the researcher's desired mix of cost, transparency, and the particular quantity of interest. We chose to start with government records because they exactly match electorally-relevant caste. We end with matrimonial records resulting from caste self-identification and calculated using machine learning algorithms. Using matrimonial records is inexpensive (if one has already constructed the algorithm) and the process is conducted in a transparent fashion, but it does not reflect electorally-relevant caste and, therefore, is substantially less applicable for our purpose than other methods.

Table 2 displays subjective rankings of different methods based on these three criteria. We define cost as the combination of speed to complete the coding and financial expense. Here expert name classification and archival research both require hiring highly-trained people to conduct time consuming work. Using pre-existing data like government records or surname lists or algorithmic methods (if training data has already been collected) have no financial cost and take little time. Transparency is the extent to which the coding procedure is written and interpreted with little ambiguity — a key component in producing replicable research. Automated methods are highly transparent because the training data can be fully specified. Crowd sourced name classification typically has more training and procedures for coders than does expert classification. Finally, different methods are more amenable to coding caste in different contexts.

Table 2: Comparison of Caste Coding Methods

Method	Cost	Transparency	Contexts
Government Records	Low	High	Electoral, Social
Name Classification (Expert)	High	Medium-Low	Electoral, Social
Name Classification (Crowd Sourced)	Medium	Medium	Electoral, Social
Surname Lists	Low	High	Electoral
Archival Research	High	Medium-Low	Social
Matrimonial Data	Low	High	Self
Other Algorithmic Methods	Low	High	Depends on Training Data

Table 2 is not exhaustive, and additional evaluation criteria may also be appropriate. The table does, however, illustrate how a researcher might evaluate trade-offs between different methods and select a process that orders the methods in an appropriate manner.[4]

Future research would do well to implement additional systematic approaches to coding caste like this one in different contexts and using different caste categories. Such work would require selecting a potentially different set of coding methods and re-ordering them to best match the coding context and caste categories. Additional research on caste coding procedures and methods, including developing and assessing techniques to more effectively measure self-identified caste or to establish and measure other caste categories, is welcome and needed. Our study is limited to electorally-relevant caste because we focus on caste among politicians, but electorally-relevant caste may also be measurable in members of the public. Beyond India, politically relevant identities are a subject of much research and discussion (Cederman, Wimmer and Min, 2010). This work has focused on identifying whether an identity is politically-relevant or not. Future work could emphasize processes for measuring electorally-relevant identities in politicians and members of the public.

[4] SI.4 discusses algorithmic methods in more detail.

References

Adukia, Anjali, Sam Asher, Paul Novosad and Brandon Tan. 2019. “Residential Segregation in Urban India.”.

Aggarwal, Ankita, Jean Dreze and Aashish Gupta. 2015. “Caste and the Power Elite in Allahabad.” *Economic and Political Weekly* 50(6):45–51.

Ahmed, Hilal. 2022. “New India, Hindutva Constitutionalism, and Muslim Political Attitudes.” *Studies in Indian Politics* 10(1):62–78.

Ahuja, Amit and Susan L. Ostermann. 2016. “Crossing Caste Boundaries in the Modern Indian Marriage Market.” *Studies in Comparative International Development* 51(3):365–387.

Ajit, D., Han Donker and Ravi Saxena. 2012. “Corporate Boards in India: Blocked by Caste?” *Economic and Political Weekly* 47(32):39–43.

Alam, Mohd. Sanjeer. 2014. “Affirmative Action for Muslims? Arguments, Contentions and Alternatives.” *Studies in Indian Politics* 2(2):215–229.

Banerjee, Abhijit, Marianne Bertrand, Saugato Datta and Sendhil Mullainathan. 2009. “Labor Market Discrimination in Delhi: Evidence from a Field Experiment.” *Journal of Comparative Economics* 37(1):14–27.

Benoit, Kenneth, Drew Conway, Benjamin E. Lauderdale, Michael Laver and Slava Mikhaylov. 2016. “Crowd-Sourced Text Analysis: Reproducible and Agile Production of Political Data.” *American Political Science Review* 110(2):278–295.

Beteille, Andre. 1996. “Varna and Jati.” *Sociological Bulletin* 45(1):15–27.

Beteille, Andre. 2012. “The Peculiar Tenacity of Caste.” *Economic and Political Weekly* 47(13):41–48.

Bhagavatula, Suresh, Manaswini Bhalla, Manisha Goel and Balagopal Vissa. 2022. "Diversity in Corporate Boards and Firm Outcomes.".

Bharathi, Naveen, Deepak Malghan, Sumit Mishra and Andaleeb Rahman. 2022. "Residential Segregation and Public Services in Urban India." *Urban Studies* 59(14):2912–2932.

Cederman, Lars-Erik, Andreas Wimmer and Brian Min. 2010. "Why Do Ethnic Groups Rebel? New Data and Analysis." *World Politics* 62(1):87–119.

Chen, Guoli, Raveendra Chittoor and Balagopal Vissa. 2015. "Modernizing without Westernizing: Social Structure and Economic Action in the Indian Financial Sector." *Academy of Management Journal* 58(2):511–527.

Clark-Deces, Isabelle. 2007. "How Dalits Have Changed the Mood at Hindu Funerals: A View from South India." *International Journal of Hindu Studies* 10(3):257–269.

Corbridge, Stuart, John Harriss and Craig Jeffrey. 2013. *India Today: Economy, Politcs, and Society.* Cambridge: Polity Press.

Damaraju, Naga Lakshmi and Anil K. Makhija. 2018. "The Role of Social Proximity in Professional CEO Appointments: Evidence from Caste/Religion-Based Hiring of CEOs in India." *Strategic Management Journal* 39(7):2051–2074.

Desai, Sonalde and Amaresh Dubey. 2012. "Caste in 21st Century India: Competing Narratives." *Economic and Political Weekly* 46(11):40–49.

Desai, Sonalde and Reeve Vanneman. 2015. "India Human Development Survey-II.".

Farooqui, Adnan. 2020. "Political Representation of a Minority: Muslim Representation in Contemporary India." *India Review* 19(2):153–175.

Fisman, Raymond, Daniel Paravisini and Vikrant Vig. 2017. "Cultural Proximity and Loan Outcomes." *American Economic Review* 107(2):457–492.

Gill, Mehar Singh. 2007. "Politics of Population Census Data in India." *Economic and Political Weekly* 42(3):241–249.

Gupta, Dipankar. 2005. "Caste and Politics: Identity Over System." *Annual Review of Anthropology* 34(1):409–427.

Jaffrelot, Christophe. 1996. *The Hindu Nationalist Movement and Indian Politics: 1925 to the 1990s.* London: Hurst & Company.

Jaffrelot, Christophe and Sanjay Kumar, eds. 2012. *Rise of the Plebeians?: The Changing Face of the Indian Legislative Assemblies.* New Delhi: Routledge.

Jayaraman, Raja. 2005. "Personal Identity in a Globalized World: Cultural Roots of Hindu Personal Names and Surnames." *The Journal of Popular Culture* 38(3):476–490.

Jodhka, Surinder S. 2004. "Sikhism and the Caste Question: Dalits and Their Politics in Contemporary Punjab." *Contributions to Indian Sociology* 38(1-2):165–192.

Jodhka, Surinder S. 2012. *Caste.* New Delhi: Oxford University Press.

Karekurve-Ramachandra, Varun and Alexander Lee. 2020. "Do Gender Quotas Hurt Less Privileged Groups? Evidence from India." *American Journal of Political Science* 64(4):757–772.

Krishna, Gopal. 1966. "The Development of the Indian National Congress as a Mass Organization, 1918-1923." *The Journal of Asian Studies* 25(3):413–431.

Lee, Alexander. 2022. "The Library of Babel: How (and How Not) to Use Archival Sources in Political Science." *Journal of Historical Political Economy* 2(3):499–526.

Mateos, Pablo, Richard Webber and P. A. Longley. 2007. The Cultural, Ethnic and Linguistic Classification of Populations and Neighbourhoods Using Personal Names. Technical Report 116 University College London London: .

Narain, Iqbal and Mohan Lal Sharma. 1972. "Election Politics, Secularization and Political Development: The 5th Lok Sabha Elections in Rajasthan." *Asian Survey* 12(4):294–309.

O'Hanlon, Rosalind. 2017. "Caste and Its Histories in Colonial India: A Reappraisal." *Modern Asian Studies* 51(2):432–461.

Omvedt, Gail. 2010. "Caste in the Census." *Social Change* 40(4):405–414.

Pushpendra. 1999. "Dalit Assertion through Electoral Politics." *Economic and Political Weekly* 34(36):2069–2618.

Rajadesingan, Ashwin, Ramaswami Mahalingam and David Jurgens. 2019. Smart, Responsible, and Upper Caste Only: Measuring Caste Attitudes through Large-Scale Analysis of Matrimonial Profiles. In *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 13 pp. 393–404.

Rathore, Aakash Singh. 2020. "Force-Fitting Ethnicity onto Caste." *Economic and Political Weekly* 55(47):27–32.

Sahay, Gaurang R. 2004. "Hierarchy, Difference and the Caste System: A Study of Rural Bihar." *Contributions to Indian Sociology* 38(1-2):113–136.

Sahoo, Sarbeswar. 2017. Caste. In *The Wiley-Blackwell Encyclopedia of Social Theory*, ed. Bryan S Turner. 1 ed. Wiley pp. 1–7.

Samarendra, Padmanabh. 2016. "Local "jatis" and Pan-Indian Caste: The Unresolved Dilemma of M.N. Srinivas." *Contributions to Indian Sociology* 50(2):214–239.

Satyanarayana, K. 2014. "Dalit Reconfiguration of Caste: Representation, Identity and Politics: Dalit Reconfiguration of Caste: Representation, Identity and Politics." *Critical Quarterly* 56(3):46–61.

Sengupta, Anasuya. 2010. "Concept, Category and Claim: Insights on Caste and Ethnicity from the Police in India." *Ethnic and Racial Studies* 33(4):717–736.

Shah, Parth J. and Makarand Bakore. 2006. *Ward Power: Decentralized Urban Governance.* New Delhi: Centre for Civil Society.

Shah, Paru R. and Nicholas R. Davis. 2017. "Comparing Three Methods of Measuring Race/Ethnicity." *The Journal of Race, Ethnicity, and Politics* 2(1):124–139.

Sundar, Nandini. 2000. "Caste as Census Category: Implications for Sociology." *Current Sociology* 48(3):111–126.

Susewind, Raphael. 2015. "What's in a Name? Probabilistic Inference of Religious Community from South Asian Names." *Field Methods* 27(4):319–332.

Susewind, Raphael. 2017. "Muslims in Indian Cities: Degrees of Segregation and the Elusive Ghetto." *Environment and Planning A: Economy and Space* 49(6):1286–1307.

Vaid, Divya. 2014. "Caste in Contemporary India: Flexibility and Persistence." *Annual Review of Sociology* 40(1):391–410.

Vissa, Balagopal. 2011. "A Matching Theory of Entrepreneurs' Tie Formation Intentions and Initiation of Economic Exchange." *Academy of Management Journal* 54(1):137=158.

Waghmore, Suryakant. 2019. *Hierarchy without System? Why Civility Matters in the Study of Caste.* London: SAGE Publications, Inc. pp. 182–194.

Walby, Kevin and Michael Haan. 2012. "Caste Confusion and Census Enumeration in Colonial India, 1871–1921." *Histoire sociale/Social history* 45(90):301–318.

Supplemental Information: Determining Politicians’ Electorally-Relevant Caste Membership

SI.1: Caste and Religious Categories

The main text describes our decision to create an “other religion” category instead of assessing caste identity among non-Hindus. There are, of course, caste-like divisions among Muslims (T. Ahmad 2023). Here we are interested in how these divisions map onto our already established caste categories that reflect electorally-relevant caste. We will reiterate two reasons for this choice. First, assigning electorally-relevant caste identity to non-Hindus differs by geographic area within India because political competition emphasizes different socio-political cleavages in different regions. Muslims, for example, have political reservations in some states as part of the OBC rolls. This is not true in all states (Ali 2012). Additionally, the extent to which Muslims have integrated caste identity into their religious background differs by geographic area (e.g., Ahmad 1962; Nazir 1993). Since our interest is in developing a coding method that can work consistently across states, we need to adopt more general caste categories.

Second, and often relatedly, prevailing political competition sees Muslims as a political bloc, whether this is accurate or not (Heath, Verniers, and Kumar 2015; Verma and Gupta 2016). Prime Minister Narendra Modi strongly condemned reservations for Muslims in the 2024 election (HT 2024). The National Commission for Backward Castes has ruled that reservations cannot occur based on religion alone (PTI 2024). Of course, Muslim reservations need not necessarily occur based on religion and could occur based on class using equivalencies between the Muslim class system and Hindu castes (Patnaik 2020). In places where such a reservation does exist like Rajasthan, that reservation is being reviewed as part of a broader BJP campaign seemingly against Muslim reservation, even in the form of a class-based reservation (Dutta 2024; FP 2024; The Hindu 2024).

Therefore, while national government policy may change in the future, current policy and current political discourse suggest that it is most appropriate to create an "other religion" category instead of integrating other religions into Hindu caste categories. A different approach is certainly warranted if a researcher is interested in caste categories in one state and that state has Muslims on the OBC lists. This discussion illustrates the need to align the context in which caste is being coded with the caste categories that are selected.

SI.2: Municipal Corporators in Delhi

As discussed in the main text, we implement our method using the 272 Delhi Municipal Corporators serving from 2018 to 2019. Municipal government in the National Capital Territory of Delhi consists of multiple, separate government structures. The National Capital Territory has its own government, which functions like a state government. Municipal governance is split between three bodies: the New Delhi Municipal Council, which governs central Delhi; the Delhi Cantonment Board, which governs military areas; and the Municipal Corporation of Delhi (MCD). The MCD was split into three bodies during the period under investigation --- the North, South, and East Delhi Municipal Corporations, but it was reunified into a single body in 2022. The new, unified MCD performs the same functions as the three separate bodies and has a very similar set-up, essentially creating a super-structure to encapsulate the three corporations. Corporations are responsible for making local-level decisions mostly on quality-of-life issues within the corporation.

Municipal corporations are comprised of corporators from single member districts (called constituencies) elected every five years. Constituencies are grouped into wards, with multiple corporators representing adjacent constituencies serving on a ward committee. Ward committees are responsible for managing public service requests within the ward (Shah and Bakore 2006). Corporators can also serve on corporation-level committees including a standing committee, the highest form of elected governance in the corporation. Corporators vote to select the members of these committees. Party leaders play an important role in this process, as well as in the operation of committee meetings.

Several key features make Delhi Municipal Corporators appropriate for use in this caste coding exercise. First, municipal corporators are elected officials. They run public campaigns that are reported on in the media. Second, they have a public presence. Corporators' names are listed online, and corporators run political campaigns to win office. Third, caste categorization in Delhi follows nationally relevant patterns. As the capital of India, officials elected in Delhi municipal corporations have occasion to politically clash with national leaders. This means that political competition involving caste falls along the six caste categories that we describe in the main text.

SI.3: Additional Coding Details

As mentioned in the main text, two coders engaged in caste coding seeking to identify electorally-relevant caste categories. We described electoral-relevance to these coders as the categorization that they would get if they asked a politician’s constituents how that politician portrayed their caste politically. Since caste coding is not absolute, we asked both coders to provide a confidence level of high (90%+ confident), medium (75%-90% confident), or low (less than 75% confident) in their classification. In all instances where a coder lacked medium confidence, we proceeded to archival review.

Most studies using coders evaluate their performance using reliability measures like those described below instead of asking coders to engage in confidence exercises. Hak and Bernts (1996) describe how socializing coders to the coding process through training exercises and opportunities to evaluate their performance can produce more reliable coding. Our confidence measure is adapted from that in the Varieties of Democracy Project (Coppedge et al. 2024). In this measure, the authors ask coders to provide a confidence percentage for each of their ratings on a 0 to 100 scale in 5% intervals, with a description next to six values (18). They use this confidence measure as a weighting scheme for their coded data. Importantly, Marquardt et al. (2019) find that self-reported confidence is an important predictor of coder reliability.

Given our iterative approach to coding caste, we use confidence to establish coding decisions that should be subjected to further review. Our description provided to coders was to code things as low confidence if there was any more than a small amount of doubt in the coder’s mind about the coding. As such, we adopted a high bar for “medium” confidence (75%+). Of course, though we provided a description, a label, and a percentage, coders may treat “medium” confidence differently. Again, this emphasizes the importance of coder training, which we used extensively in our procedure. To assist with benchmarking, we asked coders to complete a sample dataset first and provided them with feedback on their coding and confidence levels before they proceeded to the full dataset. Additionally, we always asked coders to provide brief remarks describing the rationale behind their coding. This enabled us to see how well the confidence measure aligned with the coding description and to correct any discrepancies.

Since we had two coders, we subjected all names where the coders disagreed to coding using other methods. Coders disagreed on the caste categories 38 times or in 17.04% of the 223 names that they coded. Using the coders' explanations for their choices, we coded the most likely reason behind the coding errors (Table SI.3.1) The both possible type means that a corporator could plausibly belong to either of the caste categories mentioned by the coders, and each category is roughly equally likely to occur. The both errors type means that neither coder listed a likely caste category. If coder 1 clearly made an error, that is listed as coder 1 error type and similarly for coder 2. We can see that the coders had between a 5% and 7% error rate.

Table SI.3.1: Coding Discrepancies

Discrepancy Type	Number	Percent
Both Possible	11	28.95
Coder 1 Error	15	29.47
Coder 2 Error	11	28.95
Both Errors	1	2.63
Total	38	

Note: Assessment of discrepancy reason with number and percentage out of all discrepancies in coding.

We calculated Cohen's Kappa to measure interrater agreement. For this test, we count any disagreement in categorization equally. One could make the argument that a disagreement over Brahmin versus OF is less significant than Brahmin versus ST, but we do not down-weight for more “minor” discrepancies. The Kappa value is 0.695, which is quite high, indicating very good to excellent interrater reliability. Archival research was conducted to resolve discrepancies.

SI.4: Cost and Transparency of Algorithmic Methods

We might be interested in how the cost and transparency of algorithmic methods can vary and whether cost is always lower, and transparency is always higher using algorithmic methods instead of other approaches. In terms of transparency, procedures for algorithmic approaches can usually be described in detail. This is especially the case if the underlying training data lists people's names and caste identification. If the underlying training data consists of names that experts, crowd workers, or others then classify, such an approach has the same transparency benefits and challenges as does directly utilizing a name classification approach.

Assuming that such training data is available, the next question is whether enough information about the training data is known such that another researcher could reasonably understand how the process of converting the training data into the algorithm works, if they were so inclined. Highly transparent algorithmic methods utilize open-source data or publish the underlying data such that it is open source. Other training datasets are highly classified or that are proprietary are transparent in the sense that a researcher theoretically understands the training data, but there is no possibility of that researcher ethically collecting these data. Collecting matrimonial data, for example, will violate matrimonial website terms-of-service. In short, algorithmic methods are generally transparent, but transparency is not guaranteed. A fully specified expert or crowd sourced name classification method might employ more transparent procedures compared to an algorithmic method using proprietary data and no details on how the training data was processed.

Cost --- both time to conduct the process and financial cost --- is generally low when applying algorithmic methods, if the researcher has access to an already existing method. As mentioned before, while researchers have developed algorithmic methods, many do not publish them publicly. Further, finding sufficient training data to develop certain kinds of algorithmic methods for specific contexts and categories can be time consuming and expensive. This is likely to be the case if relatively few individuals' castes are to be coded, the context is not self-identification, and there are granular categories. Table SI.4.1 describes some different contexts and caste categories and their likely cost to illustrate this potential variation.

Table SI.4.1: Cost for Algorithmic Methods in Different Contexts and Categories

Context	Categories	Cost	Description
Self-Identification	General, SC, ST, OBC, Other religion	Low	Most training data will contain these categories. Training data will reflect self-identification.
Electoral	Hindu, Muslim, Other religion	Low	Self-identification will largely match electoral identification in this context. Training data is publicly available.
Social	Jati	High	Jati is very granular and location specific. Large amount of training data needed at a level of specificity not usually available.

As described in Table SI.4.1, the challenges with finding or generating training data of sufficient size compound when examining more granular caste categories because caste identification there is location and time dependent. In other words, even if there was an algorithmic method developed for Bihar using matrimonial data, such a method would not fully align with coding caste at the jati level for people from 1950, as caste categorization changes over time and across locations. Given the highly constrained amount of historical caste data, this problem gets worse as a researcher needs to classify caste further into the past. Like transparency, generally the costs of algorithmic methods are lower than other coding methods.

References

Ahmad, Tausif. 2023. “Politics of Recognition and Caste among Muslims: A Study of Shekhra Biradari of Bihar, India.” *CASTE / A Global Journal on Social Exclusion* 4(1): 92–108.

Ahmad, Zarina. 1962. “Muslim Caste in Uttar Pradesh.” *The Economic Weekly*: 325–36.

Ali, Manjur. 2012. “Indian Muslim OBCs: Backwardness and Demand for Reservation.” *Economic and Political Weekly* 47(36): 74–79.

Coppedge, Michael et al. 2024. “V-Dem Methodology V14.”

Dutta, Prabhash. 2024. “Why Lalu Is Both Right and Wrong about Muslim Reservation.” *Times of India*. https://timesofindia.indiatimes.com/india/why-lalu-is-both-right-and-wrong-about-muslim-reservation/articleshow/110020598.cms.

FP. 2024. “No Reservation Based on Religion, Asserts PM: What Does the Constitution Say?” *Firstpost*. https://www.firstpost.com/explainers/lok-sabha-polls-religion-based-reservation-pm-narendra-modi-constitution-13768047.html.

Hak, Tony, and Ton Bernts. 1996. “Coder Training: Theoretical Training or Practical Socialization?” *Qualitative Sociology* 19(2): 235–57.

Heath, Oliver, Gilles Verniers, and Sanjay Kumar. 2015. “Do Muslim Voters Prefer Muslim Candidates? Co-Religiosity and Voting Behaviour in India.” *Electoral Studies* 38: 10–18.

HT. 2024. “No Reservation Based on Religion to Muslims as Long as I Am Alive: PM Narendra Modi.” *Hindustan Times*. https://www.hindustantimes.com/india-news/no-reservation-based-on-religion-to-muslims-as-long-as-i-am-alive-pm-modi-101714488053718.html.

Marquardt, Kyle L., Daniel Pemstein, Brigitte Seim, and Yi-ting Wang. 2019. "What Makes Experts Reliable? Expert Reliability and the Estimation of Latent Traits." *Research & Politics* 6(4): 205316801987956.

Nazir, Pervaiz. 1993. "Social Structure, Ideology and Language: Caste among Muslims." *Economic and Political Weekly* 28(52): 2897–2900.

Patnaik, Pratik. 2020. "Caste Among Indian Muslims Is a Real Issue. So Why Deny Them Reservation?" *The Wire*. https://thewire.in/caste/caste-among-indian-muslims-real-why-deny-reservation.

PTI. 2024. "NCBC Slams Blanket Categorisation of Muslims as Backward Caste in Karnataka." *Economic Times*. https://economictimes.indiatimes.com/news/india/ncbc-slams-blanket-categorisation-of-muslims-as-backward-caste-in-karnataka/articleshow/109520328.cms?from=mdr.

Shah, Parth J., and Makarand Bakore. 2006. *Ward Power: Decentralized Urban Governance*. New Delhi: Centre for Civil Society.

The Hindu. 2024. "Rajasthan to Review Reservation Granted to Muslims under OBC Category." *The Hindu*. https://www.thehindu.com/news/national/rajasthan-to-review-reservation-granted-to-muslims-under-obc-category/article68215780.ece.

Verma, Rahul, and Pranav Gupta. 2016. "Facts and Fiction about How Muslims Vote in India: Evidence from Uttar Pradesh." *Economic and Political Weekly* 51(53): 110–16.