Approaches to Coding Caste

William O'Brochta[*] and Sunita Parikh[**]
Published in *Studies in Indian Politics*

Caste identity is contextually dependent, making identifying and coding caste challenging. We describe how existing approaches to coding caste capture different contexts, use different categories, and select and order different caste coding methods. We then develop a series of suggestions to more effectively describe caste coding procedures with a focus on caste coding in electorally-relevant contexts. We conclude by discussing strategies to align caste coding objectives with methodological techniques.

Keywords: Descriptive representation, caste, ethnicity coding, classification
(3,767 words)

Acknowledgements: We thank Manaswini Bhalla, Ted Enamorado, Nivedita Mehta, Adarsh S., Bala Vissa, audiences at the 2020 American Political Science Association Annual Meeting and the 2025 Texas Methods Meeting, and the Editor and Reviewers for helpful comments and suggestions.

Financial Support: We acknowledge funding from the Weidenbaum Center on the Economy, Government, and Public Policy.

Disclosures: None.

[*]Corresponding author: Department of Political Science, Texas Lutheran University, Seguin, Texas, USA. 830-372-6566. wobrochta@tlu.edu.
[**]Department of Political Science, Washington University in St. Louis. One Brookings Drive, St. Louis, Missouri, USA. saparikh@wustl.edu.

Caste remains an important individual and collective characteristic that can vary across time and location. These variations occur because the concepts and definitions of caste have become further disassociated from long-established markers, particularly occupational requirements (Beteille, 2012). Attempts to redefine caste rankings and categories 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). Caste identity may also be expressed differently depending on context, 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). When researchers seek to identify and categorize caste, they make choices about their methodology that influence the result (e.g., Samarendra, 2016).

Caste coding represents a complex classification process often more nuanced than attempts to classify other descriptive characteristics like gender, race, ethnicity, and religion. The rise of large language models and other machine learning tools have impacted the way that caste is and will be coded, despite well-documented biases (Khandelwal et al., 2024; Seth et al., 2025). Researchers are starting to use these tools for caste coding, for example, by validating them against reserved constituencies in the Lok Sabha and feeding caste-surname lists to them as training data (Dasanaike, 2026).

It is with this future in mind that we examine already established approaches to coding caste. Our goal is to derive a series of suggestions for researchers to more effectively describe how they perform caste coding, be it with these or other methods. In doing so, we focus on the political dimensions of caste as expressed by and about elected politicians.

Defining Caste Context and Selecting Caste Categories

Caste can be defined in different ways (Gupta, 2005; Sengupta, 2010). At the highest level of aggregation, the term "caste" can refer to the varna classification system plus Dalits (Sundar, 2000; Waghmore, 2019).[1] These five classes are then subdivided into jatis, a term also translated into English as "caste." While the jati is the most frequently invoked level of caste identity, especially in everyday interactions (Jodhka, 2012), the varna designations are also regularly used.

In political discourse, the term caste frequently refers to a set of officially constructed categories drawn from a combination of varna and jati identities. These categories have been influenced by historical developments and government priorities (e.g., Rudolph, 2005). For example, the Indian government has formulated official lists comprising the Scheduled Caste (SC), Scheduled Tribe (ST), and Other Backward Classes (OBC) categories (Rathore, 2020). These categories are highly aggregated and composed of regionally demarcated subgroups, with the result that some identically named jatis are included on lists in one region but excluded in another.

If we look at how caste categories are practiced, we can approach classification from a number of perspectives, and some definitions are more useful than others for a given research question (Beteille, 1996). Caste-related studies often cross academic disciplines. Sociologists and psychologists are often interested in how caste is expressed in community interactions through socio-cultural dynamics (e.g., Waghmore, 2019), while contemporary work in economics often emphasizes how caste is used or expressed in workplaces (Chen et al., 2015; Joshi et al., 2022).

[1] We use the term "Dalit" here when referring to social categorization and "Scheduled Caste" when referring to political categorization.

As political scientists, we are often interested in the electorally-relevant caste categories used in campaigns. A typical categorization includes the officially listed SC, ST, and OBC categories because they influence critical aspects of political life, including reservations for government benefits and political seats (Vaid, 2014, p. 395). Though that system alone is sufficient in some contexts (e.g., Cassan & Vandewalle, 2021), the Indian Human Development Survey (Desai and Vanneman 2015) adds Brahmin and "other forward" as categories. An "other religion" category can also be added.

These six categories reflect the major political delineations of caste that are politically salient in India at the national level. However, the high level of aggregation of distinct caste groups may not be sufficient when categories need to reflect variation within a community. Suppose a researcher wanted to classify caste relevant in a community setting for a study of interpersonal or inter-familial relations, such as marriage choices. Inter-caste and inter-religion marriages are rare in India (Desai 2022), and a category like "other forward" would be unable to capture the rate of such marriages. Instead, we would need to identify jatis and sub-jatis within the six categories, and we might need to specify regional and/or community-level caste identifications since the same jati names can be located at different ranks in the hierarchy in different geographic areas.

Existing Caste Categorization Approaches

Self-identification, archival research, and expert classification are three popular and sometimes overlapping existing methods used to code caste.

Individuals self-identify their caste in surveys, commercial and educational environments, and political settings. Methods of self-identification include checking one of

several pre-filled options as well as open-ended entry in a text box. Employment applications, census questionnaires, and matrimonial advertisements may have required or optional questions on caste. Contemporary surveys like the Indian Human Development Survey (Desai and Vanneman 2015) and the Centre for the Study of Developing Societies' (CSDS) Lokniti National Election Studies (see Vaid, 2023 for a caste-based analysis) ask respondents to self-identify caste, a practice that follows colonial-era British government censuses (Gill, 2007; Walby & Haan, 2012). The use of open-ended responses to caste questions allows for categorization at different levels of granularity depending on the specific coding task. Various studies have used survey and other self-identified administrative data to examine caste-influenced phenomena like residential segregation (Adukia et al., 2019; Bharathi et al., 2022).

At scale, self-identification data can be used as training data, and Indian matrimonial websites are popular choices for this task. Matrimonial website data allows for caste identification to be linked to surnames, and, since many surnames have historical links to particular caste categories, a given name will trigger an association with a corresponding caste (Banerjee et al., 2009; Jayaraman, 2005). Bhagavatula et al. (2022) obtained six million matrimonial profiles from the two largest matrimonial websites and aggregated these profiles to determine the relative frequency of caste identification for given surnames.

Using matrimonial data is an example of a name-based classification method, wherein an individual's name provides information about their caste. Name-based classification can be effective when names provide informative and consistent information about caste (e.g., Clark, 2014, p. 147's study of common surnames in Bengal). Such a connection is often challenging to make in the Indian context, as caste dynamics change across locations and time. Important

existing work has focused on using names to code religious categories in India (Ash et al., 2022; Chaturvedi & Chaturvedi, 2024; Susewind, 2015).

Archival research entails trying to find caste information about specific individuals, not just those who share a common name (Narain & Sharma, 1972), and it is most productive in identifying and classifying elites. For political actors, archival research uses newspaper articles about politics, published political interviews, and government records of electoral candidates as sources of 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 focus on electorally-relevant caste, caste categories relevant in a community, interpersonal caste categories, or some combination of these.

Expert classification techniques are used across a variety of contexts, most commonly to identify characteristics of democracy (Polity, Coppedge et al., 2024) and the ideological and policy positions of parties and party platforms (the PopuList, Rooduijn et al., 2024). The term "experts" in the Indian context refers to knowledgeable people who are credentialed through academic training but can also include those who are steeped in local or regional context. This credentialing, which can be officially recognized or independently verified by the investigators using the experts, differentiates expert classification from archival research. The latter is also carried out by knowledgeable researchers, but archival researchers are more likely to rely on contextual and experiential information focused on the specific topic of relevance.

Experts on caste identification are chosen because of their familiarity with caste categorization in a given locality or region (Mateos et al., 2007), and researchers stipulate the categories to be used. Experts use their knowledge of caste names, knowledge of specific

individuals or communities, and historical patterns of caste identification to complete their categorizations. They use a variety of methods to arrive at their results: for example, in Jaffrelot and Kumar's (2009) study of state legislators, experts used a mix of original surveys, fieldwork, and other types of data collection, but none provided details about their techniques. The authors 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. In another study, Karekurve-Ramachandra and Lee (2020) first hired experts to identify caste using their knowledge of caste-surname patterns. For those politicians with ambiguous names, they asked elected officials, party members, and other prominent individuals to help by either telling them the politicians' caste or by finding out this information. Similarly, the former Ashoka University Trivedi Centre for Political Data (TCPD) collected government caste categorizations and supplemented them with expert classification and archival research.[2]

Selecting and Applying Methods

Each of the previously described approaches has thus far been used in isolation and without clear justification for how caste context and categories were selected to match the coding task. We propose that caste coding should include at least three steps --- identifying context, choosing categories, and selecting and ordering methods.

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, community, or

[2] See, for example, https://cdsa.ashoka.edu.in/hello/DemocracyData/. The TCPD was particularly interested in caste categorization of political elites, including incumbent politicians and their challengers.

interpersonal). Coding in multiple contexts is possible. It may be advisable to code each context separately and to compare results across contexts.

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 and rank-ordering them 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 may not code the caste of all individuals at the researcher's desired mix of cost, transparency, and the quantity of interest. Different methods also are more appropriate for different combinations of contexts and caste categories.

Table 1 displays information on different caste coding methods based on three criteria that emphasize both how researchers conduct caste coding and how they explicitly communicate about their work to others. To the methods described above, we explicitly add caste coding using government records --- directories of individuals who have selected a caste category on government documents that are publicly available. We focus on contexts in which experts are presented with a name and asked to identify the appropriate caste. Name classification can also be completed by crowd workers hired through platforms like Amazon Mechanical Turk (Shah & Davis, 2017). Surname lists are produced by governments as part of state or national reservations, e.g., the Delhi Central List of OBCs. Apart from using matrimonial data as training data, researchers can use other training datasets and algorithmic methods to code caste. Finally,

researchers may engage with large language models using prompt-based or other methods that may or may not reveal the underlying method the model is using to provide the caste coding.

Table 1: Comparison of Caste Coding Methods

Method	Cost	Transparency	Contexts	Key Challenge	Example
Self-Identification	High	Low	Depends on question wording	Data availability	Ferry (2019)
Government Records	Low	High	Electoral, Community	Limited categories	Adukia et al. (2019)
Name Classification (Expert)	High	Medium-Low	Electoral, Community	Clear procedures	Jaffrelot and Kumar (2009)
Name Classification (Crowd Sourced)	Medium	Medium	Electoral, Community	Worker expertise	Shah and Davis (2017)
Surname Lists	Low	High	Electoral	Data availability	Cassan et al.(2022)
Archival Research	High	Medium-Low	Community	Data availability	Karekurve-Ramachandra and Lee (2020)
Matrimonial Data	Low	High	Interpersonal	Social desirability	Bhagavatula et al. (2022)
Other Algorithmic Methods	Low	High	Depends on training data	Training data bias	Fisman et al. (2017)
Large Language Models	Low	Depends	Depends	Clear procedures	Dasanaike (2026)

We define cost as the combination of speed to complete the coding and financial expense. Here, both expert name classification and archival research require hiring highly-trained people to conduct time consuming work. Self-identification requires conducting a survey of individuals whose caste needs to be coded. Using pre-existing data like government records, surname lists, or algorithmic methods (if training data has already been collected) have low or 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. Algorithmic methods, including assigning a new surname to the caste most represented by that surname in matrimonial data, 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 name classification. Like expert name classification, archival research also requires substantial expert judgement. People evaluate their own caste through self-identification in whichever way they choose. Finally, different methods are more amenable to coding caste in different contexts. As political scientists, we tend to be most interested in how caste is portrayed in an electoral context.

We also list a key challenge of using each method. The availability of relevant data is often a challenge that necessitates using more than one caste coding method to combat cases where data is missing. In cases where data is more available, the lack of sufficiently granular categories and bias in training data can complicate the context and precision with which caste can be coded. For name classification methods, clearly replicable procedures and varying coder expertise can result in lower transparency than desired.

Discussion and Conclusion

There are myriad approaches to coding caste, and researchers can reasonably select from or utilize multiple approaches when completing a caste coding task. Regardless of the approach used, we suggest that researchers should clearly describe the context in which caste coding is occurring, the categories used, and the method(s) selected to ensure that readers have all the information necessary to evaluate the reliability and replicability of caste coding results. Large language models experience the same challenges as other approaches, and they can be an appropriate way to code caste when they best fit the coding task at hand. We argue that caste coding methods should not be considered and evaluated in isolation. Rather, researchers should be intentional about selecting and describing caste coding methods.

Future work could compare several caste coding approaches across contexts to continue the discussion about the theoretical underpinnings of caste coding and its methodological approaches. Researchers who use different caste categorization methods and who seek to categorize caste in different contexts will not necessarily produce the same categorization results. Spending time discussing particular coding approaches follows best practices in human coding (O'Brochta, 2026) and helps to enrich understanding of caste as a category and its many variations.

References

Adukia, A., Asher, S., Novosad, P., & Tan, B. (2019). *Residential Segregation in Urban India* [Working Paper]. https://cega.berkeley.edu/wp-content/uploads/2020/03/aant-segregation.pdf

Ash, E., Asher, S., Bhowmick, A., Bhupatiraju, S., Chen, D., Devi, T., Goessmann, C., Novosad, P., & Siddiqi, B. (2022). *Measuring Gender and Religious Bias in the Indian Judiciary* (Working Paper No. 1395). Toulouse School of Economics. https://www.tse-fr.eu/sites/default/files/TSE/documents/doc/wp/2022/wp_tse_1395.pdf

Banerjee, A., Bertrand, M., Datta, S., & Mullainathan, S. (2009). Labor market discrimination in Delhi: Evidence from a field experiment. *Journal of Comparative Economics*, *37*(1), 14–27.

Beteille, A. (1996). Varna and Jati. *Sociological Bulletin*, *45*(1), 15–27.

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

Bhagavatula, S., Bhalla, M., Goel, M., & Vissa, B. (2022). *Diversity in Corporate Boards and Firm Outcomes* [Working Paper]. https://research.iimb.ac.in/cgi/viewcontent.cgi?article=1643&context=work_papers

Bharathi, N., Malghan, D., Mishra, S., & Rahman, A. (2022). Residential segregation and public services in urban India. *Urban Studies*, *59*(14), 2912–2932.

Cassan, G., Keniston, D., & Kleinberg, T. (2022). *A Division of Laborers: Identity and Efficiency in India* (Working Paper No. 28462). National Bureau of Economic Research. http://www.nber.org/papers/w28462

Cassan, G., & Vandewalle, L. (2021). Identities and public policies: Unexpected effects of political reservations for women in India. *World Development*, *143*, 105408.

Chaturvedi, R., & Chaturvedi, S. (2024). It's All in the Name: A Character-Based Approach to Infer Religion. *Political Analysis*, *32*(1), 34–49.

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

Clark, G. (2014). *The Son Also Rises*. Princeton University Press.

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

Coppedge, M., Gerring, J., Knutsen, C. H., Lindberg, S. I., Teorell, J., Marquardt, K. L., Medzihorsky, J., Pemstein, D., Fox, L., Gastaldi, L., Pernes, J., Ryden, O., von Romer, J., Tzelgov, E., Wang, Y.-T., & Wilson, S. (2024). *V-Dem Methodology v14*. Varieties of Democracy (V-Dem) Project. https://v-dem.net/documents/39/v-dem_methodology_v14.pdf

Dasanaike, N. (2026). *Large Language Models Naively Recover Ethnicity from Individual Records* [Working Paper]. https://arxiv.org/pdf/2601.21132

Desai, S. (2022). *Indian Human Development Survey Research Brief No. 2*. University of Maryland.

Desai, S., & Vanneman, R. (2015). *India Human Development Survey-II*. Inter-University Consortium for Political and Social Research. http://doi.org/10.3886/ICPSR36151.v2

Ferry, M. (2019). *Caste links: Quantifying social identities using open-ended questions* [Working Paper]. https://sciencespo.hal.science/hal-03611077/file/OP_2019-1-EN-def.pdf

Fisman, R., Paravisini, D., & Vig, V. (2017). Cultural Proximity and Loan Outcomes. *American Economic Review*, *107*(2), 457–492.

Gill, M. S. (2007). Politics of Population Census Data in India. *Economic and Political Weekly*, *42*(3), 241–249.

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

Jaffrelot, C., & Kumar, S. (Eds.). (2009). *Rise of the Plebeians?: The Changing Face of the Indian Legislative Assemblies*. Routledge.

Jayaraman, R. (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, S. S. (2004). Sikhism and the caste question: Dalits and their politics in contemporary Punjab. *Contributions to Indian Sociology*, *38*(1–2), 165–192.

Jodhka, S. S. (2012). *Caste*. Oxford University Press.

Joshi, S., Kochhar, N., & Rao, V. (2022). Fractal inequality in rural India: Class, caste and jati in Bihar. *Oxford Open Economics*, *1*, odab004. https://doi.org/10.1093/ooec/odab004

Karekurve-Ramachandra, V., & Lee, A. (2020). Do gender quotas hurt less privileged groups? Evidence from India. *American Journal of Political Science*, *64*(4), 757–772.

Khandelwal, K., Tonneau, M., Bean, A. M., Kirk, H. R., & Hale, S. A. (2024). Indian-BhED: A Dataset for Measuring India-Centric Biases in Large Language Models. *Proceedings of the 2024 International Conference on Information Technology for Social Good, GoodIT '24*, 231–239. https://doi.org/10.1145/3677525.3678666

Mateos, P., Webber, R., & Longley, P. A. (2007). *The cultural, ethnic and linguistic classification of populations and neighbourhoods using personal names* (No. 116; UCL Working Paper Series). University College London. https://discovery.ucl.ac.uk/id/eprint/3472/1/3472.pdf

Narain, I., & Sharma, M. L. (1972). Election Politics, Secularization and Political Development: The 5th Lok Sabha Elections in Rajasthan. *Asian Survey*, *12*(4), 294–309.

O'Brochta, W. (2026). How Human Coding is Used and Described. *Political Studies Review*, *24*(2), 335–347.

Rathore, A. S. (2020). Force-fitting Ethnicity onto Caste. *Economic and Political Weekly*, *55*(47), 27–32.

Rooduijn, M., Pirro, A. L. P., Halikiopoulou, D., Froio, C., Van Kessel, S., De Lange, S. L., Mudde, C., & Taggart, P. (2024). The PopuList: A Database of Populist, Far-Left, and Far-Right Parties Using Expert-Informed Qualitative Comparative Classification (EiQCC). *British Journal of Political Science*, *54*(3), 969–978. https://doi.org/10.1017/S0007123423000431

Rudolph, S. H. (2005). The Imperialism of Categories: Situating Knowledge in a Globalizing World. *Perspectives on Politics*, *3*(1), 5–14. https://doi.org/10.1017/S1537592705050024

Samarendra, P. (2016). Local ``jatis" and pan-Indian caste: The unresolved dilemma of M.N. Srinivas. *Contributions to Indian Sociology*, *50*(2), 214–239.

Sengupta, A. (2010). Concept, category and claim: Insights on caste and ethnicity from the police in India. *Ethnic and Racial Studies*, *33*(4), 717–736.

Seth, A., Choudhury, M., Sitaram, S., Toyama, K., Vashistha, A., & Bali, K. (2025). How Deep Is Representational Bias in LLMs? The Cases of Caste and Religion. *Proceedings of the*

AAAI/ACM Conference on AI, Ethics, and Society, *8*(3), 2319–2330. https://doi.org/10.1609/aies.v8i3.36718

Shah, P. R., & Davis, N. R. (2017). Comparing Three Methods of Measuring Race/Ethnicity. *The Journal of Race, Ethnicity, and Politics*, *2*(1), 124–139.

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

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

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

Vaid, D. (2023). Mapping Class and Electoral Participation in India from 1996 to 2019. *Studies in Indian Politics*, *11*(2), 225–257. https://doi.org/10.1177/23210230231206649

Waghmore, S. (2019). Hierarchy without System? Why Civility Matters in the Study of Caste. In S. Srivastava, Y. Arif, & J. Abraham, *Critical Themes in Indian Sociology* (pp. 182–194). SAGE Publications, Inc.

Walby, K., & Haan, M. (2012). Caste Confusion and Census Enumeration in Colonial India, 1871–1921. *Histoire Sociale/Social History*, *45*(90), 301–318.