HAL open science

Caste links. Quantifying social identities using open-ended questions

Mathieu Ferry

▶ To cite this version:

Mathieu Ferry. Caste links. Quantifying social identities using open-ended questions. 2019. hal-03611077

HAL Id: hal-03611077

https://sciencespo.hal.science/hal-03611077v1

Preprint submitted on 16 Mar 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

SciencesPo
OBSERVATOIRE SOCIOLOGIQUE
DU CHANGEMENT

cnrs

OSC Papers

n° 2019-1 May 2019

English version

Caste links
Quantifying social identities using open-ended questions

Mathieu Ferry (OSC, LSQ)

Introduction

Since the last colonial census of 1931, the Indian statistical administration no longer considers caste membership to describe the social properties of individuals and households. This lack of statistical data available to social scientists contrasts with literature emphasizing the salience of caste in the Indian social structure (Vaid 2014). While ethnographic methods undeniably provide valuable elements in the understanding of Indian society, the absence of caste statistics precludes a synthetic view of the caste social reality.

To clarify, the term caste covers two different realities. First, it refers to a hierarchical quadripartition of the world between varnas, resulting from the *Rig Veda,* a sacred text of Hinduism. In this ideal schema, Brahmins are at the top, traditionally considered as priests, Kshatriyas come second in ritual rank, they are supposed to be warriors, then come Vaishyas, traditionally traders. At the bottom, well below these three varnas constituting the 'twice-born' high castes, one finds Shudras, who would be agricultural workers and craftsmen. Outside this system one finds the Untouchables. But this categorization is more ideological than a description of social reality. In the social world, what is meant by caste is jati, a community defined as a hereditary and endogamous social group, characterized by a traditional occupation. Varnas and jatis are related but they do not correspond perfectly.

Indeed, jatis claim positions in the varna hierarchy, which may or may not be challenged by other jatis. What one claims to be is not what one is necessarily considered to be by its peers and social etiquettes vary temporally and geographically. The highly contested identification of caste renders it a central category of Indian social stratification and one that still fills theoretical debates about its conceptualization, only without empirical quantitative sources.

The caste statistical gap is sometimes filled surprisingly in public policies by archival data of colonial censuses dating back from nearly 80 years. These are then mobilized, for instance, to justify the implementation of affirmative action policies, to attribute specific statutes for the most disadvantaged castes in societies, the 'backward' jatis. Many social scientists (including the author of this paper, see Ferry, Naudet, and Roueff 2018) then use the categories of affirmative action as a proxy for caste in their statistical analysis. But overall, this situation makes it difficult to conduct appropriate research work for social stratification sociologists.

Large contemporary sample surveys, however, collect data on caste identity through open-ended questions. These are present in the Indian Human Development Survey (IHDS, Desai et al., 2018), a longitudinal survey conducted since 2004, and in the National Family Health Survey (NFHS), a survey conducted since 1992. The difficulty posed by these data is the absence of classification of responses in a nomenclature by the survey investigators. In order to mobilize such data, it is the social scientists who have to classify the responses on their own. This procedure appears all the more difficult because there is no standardized caste schema in Indian society. The present research aims to provide a method to compensate for this deficiency. In doing so, theory is not forgotten and the conceptual relevance of the caste classification will be assessed. We emphasize the self-identification of respondents and try to inductively draw the most homogenous clusters of caste with our algorithm. The rationale is that caste is a latent and pre-existing social category that we wish to grasp quantitatively. Since manual clustering is time consuming, we suggest mobilizing automatic algorithms as a help to classification.

Statistical boundaries never correspond perfectly to social boundaries, as categorization summarizes a complex phenomenon in a simplified way. But we also argue that statistical tools can be mobilized to explore relevant boundaries between social identities while avoiding substantializing caste identities or creating statistical chimeras. A statistical and sociological approach to categorization makes it possible to break free from a fixed vision of the social world of caste, in particular its hierarchical defining feature (Dumont 1974) which has been questioned for relying heavily on ideological varnas rather than observed jatis (Dirks 2001). The distinction drawn by Desrosières (1998) between 'measuring' and 'quantifying' is important in our approach. While the first action relates to a realistic metrology, where a real identity would be measured, the second recalls the conventions mobilized when categorizing the social world. While the 'act of counting' the social world, carried out in the social sciences as well as by governmental action, aims to describe an objective reality, it also 'shapes' this reality (Desrosières and Didier 2014). In forgetting this apparently trivial fact, analyses might be conducted at the cost of a complete abstraction of social facts, which disappear behind quantification. On the contrary, in clearly exposing the conventions adopted in statistical classification, we are able to understand a posteriori the social reality embedded in social categories measured by statistical ones. Hence, instead of fee-

ling despair because of the lack of institutionalized categories, the present exercise offers a unique opportunity to question the character of caste and its salience in the social structure.

In this article, we lay the foundations of caste quantification by focusing on the Hindu population of Uttar Pradesh.[1] The regional variability of caste identities prompts us to focus on one particular state, and Uttar Pradesh has the advantage of being in a region of northern India where the caste system is considered close to the model of the varnas, divisions of the social world derived from sacred Brahmin writings. We will see, however, that self-identification of caste households is only marginally close to this classification.

First, we recall that the statistical measure of caste is related to Indian colonial history, which strongly marks the debates on the quantification of caste in the study of India from the social sciences perspective. Our work leads us to discuss arguments against caste measurement and develop our position on the need to consider caste identities statistically. The classification algorithm, which associates an automatic method with manual inputs, is then presented. Following that, we evaluate the caste categories' homogeneity, comparing them with a caste 'gold standard' and testing its criterion validity. We end with a conclusion, where we discuss the implications of our procedure and the future uses of this nomenclature.

1. Lessons for caste quantification in contemporary India

In order to show the relevance of using statistical surveys to study caste identities sociologically, it must be remembered that the exercise of quantifying caste is linked to India's colonial history. Censuses, as well as colonial anthropology, have undeniably rigidified a representation of the caste system, and the statistical apparatus has had performative effects on the social world. This situation explains the current reluctance of the social sciences to mobilize the statistical apparatus when it comes to caste.

a. The tools of knowledge for the colonial administration

When the first census was introduced in 1871, administrative counting was not entirely new on the subcontinent and was based on pre-statistical experiences in the Provinces, administered either by the British or by local authorities (the Mughal rulers in particular). These counts were intended to measure production, especially agricultural production, in order to set the optimal tax rate. The scientific census of 1871, which actually took place from 1867 to 1873, did not arouse opposition from the population (apart from localized tax revolts, already observed throughout the nineteenth century). On the contrary, it raised astonishment and irony from the interviewees about the curiosity of the enumerators, whose questions did not necessarily make sense for the respondents (Lardinois 1996).

These censuses, not only on production but also on human population, introduced questions on religion, sect, caste, for purposes of counting and enumeration. Indeed, the censuses were crucial tools for the control of the territory and population under the domination of the colonial empire (Appadurai 1993). If statistical techniques had already been introduced from census experiments in the United Kingdom, little socio-demographic information was collected there (only religion in 1851), the interest in the exoticism of Indian social groups being much more important (Guilmoto 1998).

Caste progressively became the main unit of description of Indian society, a vision strengthened by official anthropology that developed at the same time, under the aegis of Herbert Risley, William Crooke and Denzil Ibbetson (Fuller 2016, 2017). The project of ethnographic description of caste groups, *People of India,* reinforced the a priori of a traditional Indian society, antithetical to modern European society.

b. The production of colonial knowledge

According to Herbert Risley, and to some extent William Crooke, the social organization of caste was based on the evolutionist paradigm of racial theory, which they sought to support through a collection of anthropometric data. This conceptualization was reinforced by a theory of 'social precedence', thus reproducing a pattern of social understanding of British society based on rank. At the same time, the colonial administration also relied on indigenous informants, mainly from the 'bhadraloks' in Bengal, a literate class of high Brahmin castes, Kayasths and Baidyas, who emerged during British India (Fuller 2017). The colonial taxonomy would then be based on a socially situated view of

1 The restriction to the Hindu population is above all pragmatic because if it is not to deny a caste structure within Christian (Roberts 2016) or Muslim (Ahmad 1973) religious minorities, the data are not accurately measuring it in our opinion. No response rates are high, and they are often mentioning religious divisions (Shiite versus Sunni) rather than caste as such. Other religious minorities (Sikhs or Zoroasthrians) are only very marginally present in Uttar Pradesh.

the caste system, legitimizing a hierarchical view of the caste system supported by sacred writings (Appadurai 1993).

Colonial historiography then questions who produces colonial knowledge. The first position, described as 'postcolonial' by Fuller (2017), who summarizes Wagoner (2003) himself, is notably that of Nicholas Dirks (2001), according to whom the understanding of the caste system as a social hierarchy derives above all from European thought applied to Indian society, where Indian informants played only a passive role in collecting data. The second position, described as 'revisionist' in relation to the first, considers the fact that the informants played an active role in the production of the colonial knowledge about Indian society, which would therefore be the fruit of both colonial and indigenous thought. This position resonates with the one developed by Susan Bayly (2001), who also notes that the social organization where Brahmanic values are dominant, historically dates back to the 18th century. The emergence of this setup was reinforced by the proximity of literate and Brahmin castes to the colonial apparatus. Synthesizing positions, Fuller (2017) evokes a 'shared understanding' between European evolutionist thought and Brahmanic thought that produced caste knowledge organized around a hierarchical view of caste.

Far from developing the idea that caste is the product of a colonial 'fantasy' (Bayly 2001), this historiography makes it possible to understand that data collection and understanding have been based on paradigms that have had effects in the production of colonial knowledge. Anthropologists and sociologists subsequently mobilized these data sources even though they did not fit into the evolutionist paradigm. This is the case, for example, of Max Weber, who when writing about Herbert Risley's documents that they 'belong to the best general sociological literature available' (Fuller 2017). It is also the case of Louis Dumont (1974), whose understanding of caste is based on the hierarchical model of varnas, where the political dimension is less important than religion (Dirks 2001).

c. Classificatory difficulties of caste identities

If caste questions were open-ended in the censuses, the responses were then categorized. This operation raises the question of the definitional boundaries of caste groups. On the one hand, few castes possessed a regional and transregional identity, and accounting for the local reality of caste is not easy. On the other hand, caste boundaries are not fixed in time, and are subject to reconfigurations according to social relations (Guilmoto 1998). The classification exercise would then have institutionalized so-called traditional identities, whose contours were previously fluid, changing and open. The use of the Brahmanic statutory hierarchy in caste indeed materialized in the order of classification of castes. For example, in the nomenclature used in the Bombay Presidency in 1872, the Brahmans were counted in position 1 and the sub-caste of the Mahars (untouchable caste) was in position 147 (Appadurai 1993).

This hierarchy induced by census categorisations had 'performative effects' on the social world (Desrosières 2001). Indeed, the recourse to the classification of Hindu texts in the census led to a caste competition for their position in the nomenclature thus created. The jatis were recomposed, and we then witnessed the formation of 'meta-castes', which consist of a grouping of jatis in a local or trans-regional way, in order to claim a varna status higher than that granted in the colonial nomenclature. These caste alliances have led to claims in courts (Headley 2013). An archetypal example of caste alliance is the Yadavs, which group together Ahirs, Goallas, and Gopas, for the explicit purpose of 'sanskritisation', that is, a social upward mobility through the adoption of higher caste cultural practices in the hierarchy of ritual purity (Michelutti 2008). Far from being socially neutral, the census makes it possible to administratively concretize and socially institutionalize the hierarchy of varnas.

This tension between varna and the anthropological reality of jati has not disappeared in the post-colonial period, but another issue is at stake. During the colonial period, the categorizations of the administration led to certain low- and middle-ranking jatis claiming a higher status than that granted by the administration. To legitimize this status by a legal recognition, in the contemporary period, categorizations of caste by 'quota policies' in a logic of affirmative action might mean seeking an administrative-sanctioned 'backward' status, which is crucial in gaining access to public higher education institutions (Henry and Ferry 2017). Caste groups mobilized for the recognition of their status of Scheduled Castes (SCs) and Scheduled Tribes (STs) at independence, and then to be counted as Other Backward Classes (OBCs) in the 1980s (Lardinois 1985). To follow with the Yadav example, in present day India, they are recognized as OBCs, a 'backward' administrative status that is compatible with a high varna status claim. These mobilizations continue until today, with the recent mobilizations of Marathas in

Maharashtra, Patels in Gujarat and Jats in Haryana (Deshpande and Ramachandran 2017). Ironically, these demands are justified on the basis of socioeconomic backwardness, which, to be proven, requires the mobilization of caste-related data. The report of the Mandal Commission (officially named the Socially Backward Classes Commission) of 1983 had thus mobilized the data of the census conducted in 1931, where the issue of recoding open questions on caste arises from.

This historical detour is justified here, not to participate in the postcolonial critique of colonial bureaucratic institutions, but, more humbly, as a call to avoid the dead-end of the quantification project of official 'colonial anthropology'. While many social scientists participate in this criticism by rejecting the possibility of statistical knowledge of the social structure of caste, we are moving away from this position. In fact, the committed stance of our work is to consider that the historiography of the colonial statistical apparatus shows us the pitfalls that must be avoided in the articulation of caste quantification. We attempt to sketch the most problematic features of caste quantification, so as not to reproduce them.

2. A recurrent debate in the social sciences

If the performative effects of caste in colonial censuses are recognized throughout the Indian social sciences, some voices still call to take caste into account in current statistical surveys. This position, to which we associate ourselves, aims to go beyond postcolonial criticism of the use of statistical categories to measure caste belonging. In short, the goal is to move from a non-categorization practice to a categorization practice supported by positive reflexivity on quantification.

a. A problematic complexity

The debate about caste inclusion in census surveys has been revived in India with the establishment of the 2011 *Socio Economic and Caste Census* (SECC), following a proposal by the *Registrar General of India* in 1998. In this census, an open question was asked about caste membership. While reports of this census have been published, no caste data, other than quota category membership (SC, ST, OBC and General for non-quota castes) are publicly available.[2]

The official argument advanced for the non-publication of the results is the possible diversion of these data for discriminatory purposes. The experience of the statistical measurement of religion in India, however, suggests that this exercise is also a means of overcoming effects of discrimination, as described by Deshpande and John (2010). Other reasons are also put forward for the non-use of statistical categories of castes. Those are reminiscent of debates on the use of ethnic categories in the French context,[3] which raise technical, but also political and social issues.

From a social point of view, caste identity refers to multidimensional affiliations, as recalled by Zoé Headley (2013), depending on whether respondents prefer a response referring to their varna (which is not necessarily accepted by the other castes), their jati (which can be expressed in several ways), or their meta-caste. These groups refer to identities that vary regionally and temporally, depending on the social relations in which individuals are inserted (Bayly 2001).

An argument often put forward by detractors of caste measurement in censuses, and by extension in large statistical surveys, is that respondents could give false information in their declaration, from a utilitarian perspective. But Deshpande and John (2010) sweep aside this argument, recalling that the census is simply a mandatory public survey that is independent from the process of claiming the benefits of quota policies, which requires the procurement of a caste certificate. However, the argument is interesting, since it is likely that individuals belonging to caste associations, or caste political parties, will mention their 'meta-caste' rather than their sub-caste (Yadav rather than Ahir for example). Hence, it is also in this sense that the census, and any statistical survey referencing caste, is a relevant sociological instrument to study the social world; they are a means of recording the answers that individuals want to be included in the archives. Therefore, respondents mention the social identity in which they recognize themselves.

The reluctance to take into account caste in statistical surveys is finally a criticism of what statistics can bring to the understanding of society, on processes of segregation and discrimination, beyond simply highlighting correlations on social inequality indicators (Sundar 2000). Without entering into the debate, both technical and epistemological, regarding the distinction between correlation, causality, discrimination and segregation, note however that we should first have statis-

2 With regard to the 2001 census, we however know the caste composition of Scheduled Castes and Scheduled Tribes.

3 For a summary of the positions taken in the French controversy, see the special issue of the *Revue Française de Sociologie,* 'The Use of Ethnic Categories in Sociology' (Felouzis 2010).

tical association measures between caste and other variable in order to discuss the nature of their relationship. The distrust in statistical analysis appears thus linked to a disciplinary position of sociology in India, close to anthropology, which has little invested statistical methods (Lardinois 2013).

b. The relevance of quantification

Existing social science studies nevertheless point out the importance of taking into account statistics, when highlighting social inequalities based on caste belonging (Deshpande 2005). Among the works on caste mobilizing statistical categories, we must distinguish two poles, on the one hand, those that mobilize the administrative categories of the quota policy, and, on the other hand, those that use more precise caste nomenclatures, but rely on geographically and socially located surveys.

The use of administrative categories (SC, ST, OBC) allows the study of (positive or negative) effects of the 'quota policies' within higher education institutions (Henry and Ferry 2017) and the Indian public administration (Benbabaali 2008). These categories may also be mobilized to import questions and methods of studies of discrimination and social exclusion developed from racial categorizations in Anglo-Saxon sociology (Thorat and Neuman 2012), while taking into account the specificity of modes of exclusion in Indian society, such as those related to the practice of Untouchability (Borooah 2017). These categories are also useful for identifying the specificities of living conditions of Dalits (roughly corresponding to SCs category), Adivasis (category corresponding to STs) and Brahmins (the IHDS data isolates them thanks to an ad hoc category), while OBCs do not correspond to a clearly identifiable caste identity. These categories are therefore a first step towards a relative study of caste belonging, articulated with reference to religion and social class, for example to understand the social structure of consumption patterns (Ferry et al. 2018), or different social fluidity levels between castes in intergenerational class mobility (Vaid 2018).

However, such a nomenclature becomes ineffective to account for finer descriptions of caste identity. It is indeed impossible to distinguish high castes, middle castes, and low castes (except untouchable and tribal castes), whether we consider this division from the point of view of a ritual or socio-economic hierarchy. Notably, whereas the model of caste theorized by Louis Dumont (1967) in *Homo Hierarchicus* is criticized for being too ritual-status oriented, hence overlooking the role of caste in the appropriation of resources, it remains difficult to test competing models operationalized statistically (see however Desai and Dubey 2012, for an attempt). A model of caste based on a one-dimensional status hierarchy might not be appropriate, or at least it could require thorough testing, insofar as the high castes are differentiated between themselves, especially those which put the Brahmanic moral values in the foreground, as opposed to the concurrent Kshatriyas moral values (Bayly 2001). Further, M. N. Srinivas (1952, 1959) has highlighted the role of 'dominant castes' at the village level, those being middle or upper ritually ranked castes, but key castes in the appropriation of resources, resulting from their agrarian domination. Again, it is impossible to objectify this social reality through the use of statistical surveys, except from local ones (Himanshu, Jha, and Rodgers 2016).[4] High caste fractions are also marked by differentiated access to social resources, which is confirmed by the analysis of social trajectories allowing access to positions of economic power. Indeed, among the top business leaders, the Brahmins are closer to the state apparatus and owe their position of power to the inheritance of institutionalized cultural capital, degrees, while the merchant castes inherit family economic empires (Naudet, Allorant, and Ferry 2018) and caste-based homophily is also key to the development of entrepreneurial strategies (Vissa 2011). This homophily is also strongly present in Indian politics, where caste belonging determines 'vote-banks'. Indeed, the analysis of caste belonging of elected Indian officials reveals the weakening role of the high castes in favour of agricultural and low castes since the 1950s (Jaffrelot and Kumar 2012, Jaffrelot 2010).[5] The importance of caste identities in Indian democracy thus requires the collection of comprehensive and accurate data on Indian political elites, as already done in northern India.[6]

This non-exhaustive overview shows the importance of caste identities in the study of the Indian social structure, although research

4 Note that an extremely promising statistical operationalization has been realized by Iversen et al. (2010), using data from the IHDS, but the coding of open-ended questions, is only briefly presented.

5 For a summary of peasant caste movements in contemporary India, see in particular Bayly (2001, Chapter 7).

6 Data are freely available from the Centre de Données Socio-Politiques: https://cdsp.sciences-po.fr/fr/ressources-en-ligne/ressource/en.cdsp.ddi. NorthIndianMPs/, and a project combining historical depth and greater geographical coverage is also under way: http://www.sciencespo.fr/ceri/fr/content/sociologie-des-elus-nationaux-et-regionaux-du-raj-l-union-indienne-contemporaine.

that uses fine-grain caste categories is usually limited in its social and geographical scope. Methodologically, caste quantification is often effectively operationalized through the manual coding of surnames, despite the existence of limited strategies of 'disidentification' of social categorizations (Copeman 2015). The information is sometimes manually crosschecked with existing biographies to confirm the information given by the name. Such a tedious operation, requiring considerable time and resources, cannot, however, be mobilized for statistical surveys covering large geographic areas with anonymous respondents.

3. Categorizing the diversity of caste identities

The household survey data used here is intended to provide a representative picture of the population distribution of Uttar Pradesh. A quick review of literature shows that open-ended questions of caste belonging from the IHDS and NFHS have so far not been recoded in a systematic way. Therefore, we present our own classification method.

a. The format of open-ended questions

In the IHDS-I (2004-2005), there is one open-ended question on caste belonging: 'Which caste do you belong to?' (ID12). In the NHFS-I from 1992-1993, responses to the same question were recoded using the (very detailed) classification of the 1931 census. But as already noted, criticisms on the adequacy of this classification might prevent its use. In the NFHS-II (1998-1999) and the NFHS-III (2005-2006), responses have not been recoded and are left in a raw format. Although the NFHS-IV (2015-2016) includes a question on caste belonging, the answers have not been released. Contrary to the previous surveys, this latest NFHS round is district representative. Since caste identities are considered sensitive, this might be why data on caste belonging has not been released (Deshpande and John 2010). Note that the NFHS asks these questions to individuals (aged 15 to 59) in the households, and not to the household respondent (often the household head) as in the IHDS. This could prove useful at a later stage in assessing the variability of caste enunciation among household members, given the high caste endogamy.

In the IHDS-II (2011-2012), two open-ended questions are asked: 'Which caste/jati and sub caste/sub jati do you belong to?' (ID12aNM and ID12bNM). Table 1 presents an excerpt of the answers in the database. 'Sub caste' or 'sub jati' is an administrative artefact that does not carry any real meaning but aims at precising the jati of respondents more narrowly, who generally use it to assess their varna status. The doubling of the caste belonging question is indicative of the complexity of caste identities, as we have noted. Since complexity is often used as one of the reasons for not counting caste, we will address it here by categorizing caste from these questions.

Table 1 - Excerpt from the 'jati' and 'sub-jati' variables in the IDHS-II database.

ID12ANM	ID12BNM
...	...
YADAV	AHIRI
YADAV	AHIRJABAL
CHAMAR	AHIRVAR
JATHA	AHIRWAR
CHAMAR	AHIRWAR
YADAV	AHRI
ARAKH	AKUR
DHOBI	AMRI
DHOBI	AMRI
ARAKH	ARBANLI
PANJABI	ARORA
KORI	ASTI
KORI	ATARI
BRAHMIN	AVASTHI
...	...

b. Methodological paths in the literature

Because castes show strong regional variability in the Indian subcontinent, Deshpande and John (2010) suggest that in the hypothesis that caste belonging would be the subject of an open question in the census, the construction of a nomenclature should start from a local geographical scale, ideally from the district. Their suggested statistical study of caste should therefore be inductive in the geographical sense. The census, being exhaustive, would allow to establish an extremely detailed statistical table of caste identities. This method is more limited with surveys because they are neither exhaustive nor representative at the district level, as in the IHDS-II, which is only state-wise representative. Recognizing regional caste variability, we choose to concentrate on one particular state in our classification. Sonalde Desai (2010), one

of the organizers of the IHDS survey, suggests another framework of caste analysis, assuming a question would be integrated in the 2021 census. She suggests drawing inspiration from the nested nomenclatures of occupations, particularly the one used in India, the National Classification of Occupations (NCO). The groupings between castes would be made at different levels in the nomenclature, with these levels integrating a spatial dimension. She suggests constructing the nomenclature a priori using the 1931 census data, the IHDS open-ended questions and matrimonial advertisements, which often mention caste. The creation of such a nomenclature would require extensive collaborative work, like the commission appointed by the Indian government to recode the unreleased 4,673,034 different caste enunciations of the Socio Economic and Caste Census of 2011.[7]

The current use of caste open-ended questions in the IHDS and the NFHS does not attempt to establish a caste nomenclature. The exercise sometimes conducted is to recode parts of the household sample in order to examine particular castes, by looking at certain archetypal caste categories of quotas for example (Srinivasan and Kumar 1999, exercise conducted using the NFHS data). The other, more recent perspective is to look at the socio-economic condition of particular castes who are seeking a 'backward' status and who are mobilizing to be officially counted as OBCs. Deshpande and Ramachandran (2017) have hence used the IHDS data to study Jats (in Haryana), Patels (in Gujarat) and Marathas (in Maharashtra). A. Kalaiyarasan (2016) focuses on Jats in Haryana, and A. A. Dongre (2017) studies the Marathas in Maharashtra.

The last two articles present their recoding methods, which consist of selecting households and individuals according to their declared membership of these castes. This is not so obvious, however, since several different spellings for one caste are used. We may assume that these different spellings do not signify anything in themselves, and are simply the result of the operation of translation and transcription in Latin alphabet. Indeed, in the case of the Jats, the various transcriptions are: 'JAAT', 'JAHT', 'JAT', 'JAT SIKH', 'JATH', 'JATT', 'JHAT' and 'RON JAT'. In the case of the Marathas studied by Dongre (2017), the method used for recoding is more complex: 'The households who report their jati to be Maratha were defined as Maratha households. Their sub-jati can be either of Maratha, Patil, Kshatriya, Rajput. In some instances, no sub-caste is mentioned. Instances where jati is Maratha but sub-jati is Agri, Kunbi, Hatkar are not categorized as Maratha but as OBCs, in line with government rules.'

This underlines the difficulty of recoding. First, what Dongre observes as a 'sub jati' with respect to the Marathas corresponds to the assertion of a ritual status when respondents declare themselves as 'Kshatriyas' (the second highest varna) or 'Rajputs' (a caste claiming Kshatriya varna), in concurrence with the fact that Marathas are often considered to be 'Shudras' (fourth varna, low in rank). This 'sub jati' might refer to the traditional occupation of Marathas in the military or a 'rajputisation' or 'kshatriyasation' of their caste belonging (Lardinois 2005), a social mechanism similar to 'sanskritisation'. Second, we observe that 'jati' and 'sub jati' enunciations correspond to a gradation of identity assertion (one is first a Maratha and then a Kshatriya), rather than to the logic of a nested nomenclature. Third, some households claim the Maratha caste identity, but they also precise that they belong to an OBC classified jati, therefore claiming both, the Maratha identity and the 'backward' status, which is counterintuitive since Marathas are not classified as OBCs. Overall, Dongre's precisions recall that caste identities are both fluid and multidimensional.

c. Empirical strategy

These remarks are extremely useful, and at the same time, disconcerting to constructing a classification. Do the caste identities enunciated in the 'jati' and 'sub jati' questions correspond to regular patterns that can be measured statistically? Or do these caste enunciations simply outline the importance of caste identities and reinforce it as a complex one, impossible to capture in a single classification?

i. Moving forward with orthographic variability

The counting of caste statements in the IHDS-II shows that at the Indian level 8,318 'jati' and 9,163 'sub-jati' statements are present in the database, and when both are combined, 18,425 ways of expressing one's caste are captured, while the sample includes 42,152 households. When restricting the sample to Hindus in Uttar Pradesh, which includes a sample of 2,958 households, 755 'jati' statements, 878 'sub-jati' statements, and 1,603 different combinations appear in the data.

However, this apparent diversity of caste identities is overestimated by the orthographic variability of the same character strings. These spelling differences may be related to the transposition into Latin of the

7 http://indianexpress.com/article/india/india-others/cabinet-meeting-panagariya-to-head-panel-to-classify-caste-census-data/

various alphabets of the questionnaire (11 official languages), to multiple spellings of the same term according to the convention adopted in Latin, and to orthographic errors during transcription.

After cleaning the character strings of all 'polluting' signs (such as '[' or '/'), we consider spelling correction. We mobilize algorithms of string grouping based on character sequence similarity. This uses the method of the 'close neighbours', with the algorithm 'Fingerprint' and the Levenshtein distance, mobilized in OpenRefine (Verborgh and Wilde 2013).[8] This technique is based on the similarity of two strings defined by the minimum number of characters that must be deleted, inserted or replaced to transform one string into another. [9]

This method significantly reduces the number of distinct caste statements in the database. The number of distinct 'jati' statements drops to 425, the number of 'sub jati' to 589, and a total of 1,095 caste combinations are then present in the data. This step hence reduces the variability of caste combinations by almost a third (32.5%).

ii. Visualizing complexity with network analysis

We propose to consider the 'jati' and 'sub jati' statements as an edge list, which is the starting point for building a network, where the nodes (the vertices) of the network are the 'jati' and 'sub-jati' statements (2*2,928 nodes maximum), which are connected by the households that co-stated them (2,928 different edges maximum). The graph is edge-weighted such that a 'jati'-'sub jati' enunciation will be deemed more important if it is more frequent in the population.[10]

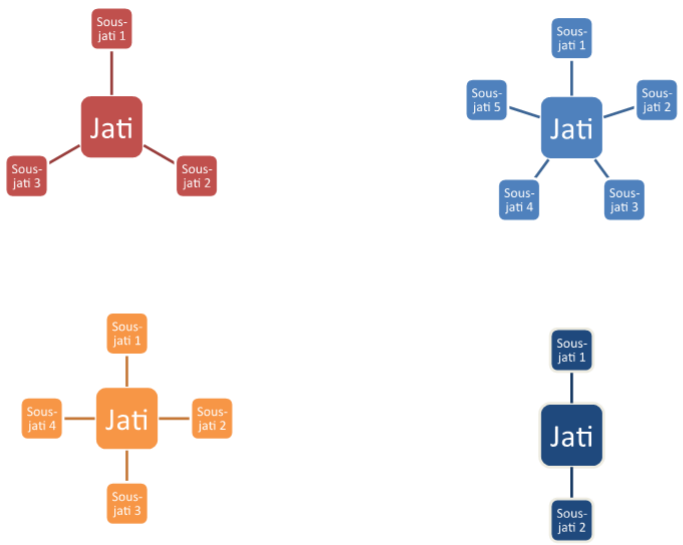
Theoretically, if the caste system formed a perfectly nested system, we should have a network similar to the diagram in Figure 1. The resulting network should form a set of disconnected sub graphs. Each sub graph (which we do not predict the number a priori) would form a star network, with a jati in the centre connected to several jatis. An example of this could be the sub graph of Brahmins, with the 'jati' Brahmin at the centre and Trivedi, Chaturvedi, Tiwari, etc. at the periphery. One could,

8 The software suggests a list of groupings which are then accepted or rejected by the user.

9 A second method based on a 'phonetic algorithm' by transforming the character strings into 'pronunciation strings' was tested. In this algorithm, two strings are identical if they are pronounced in the same way. Here, we draw inspiration from Raphael Susewind (2015) who uses a modified version of the 'Soundex' algorithm, adapted to Indian vernacular languages by Santhosh Thottingal, to identify Muslim surnames from electoral rolls. See Thottingal's website for details: http://thottingal.in/blog/2009/07/26/indicsoundex/. Since using this algorithm did not perform as well in the classification tests as the character string algorithm, particularly from the point of view of the 'gold standard' comparison (see next section), we prefer to rely on the first method.

10 For readability of the diagrams in Figure 2 and 4, the width of the edges are not plotted according to the frequency 'jati'-'sub-jati' statements but some are definitely much more common and it is in fact the less frequent statements that link the 'star' sub-networks.

Figure 1 – Theoretical diagram of the 'jati' and 'sub jati' network



thus, just consider the classification formed by these disconnected sub graphs, named after the central jati in each sub graph.

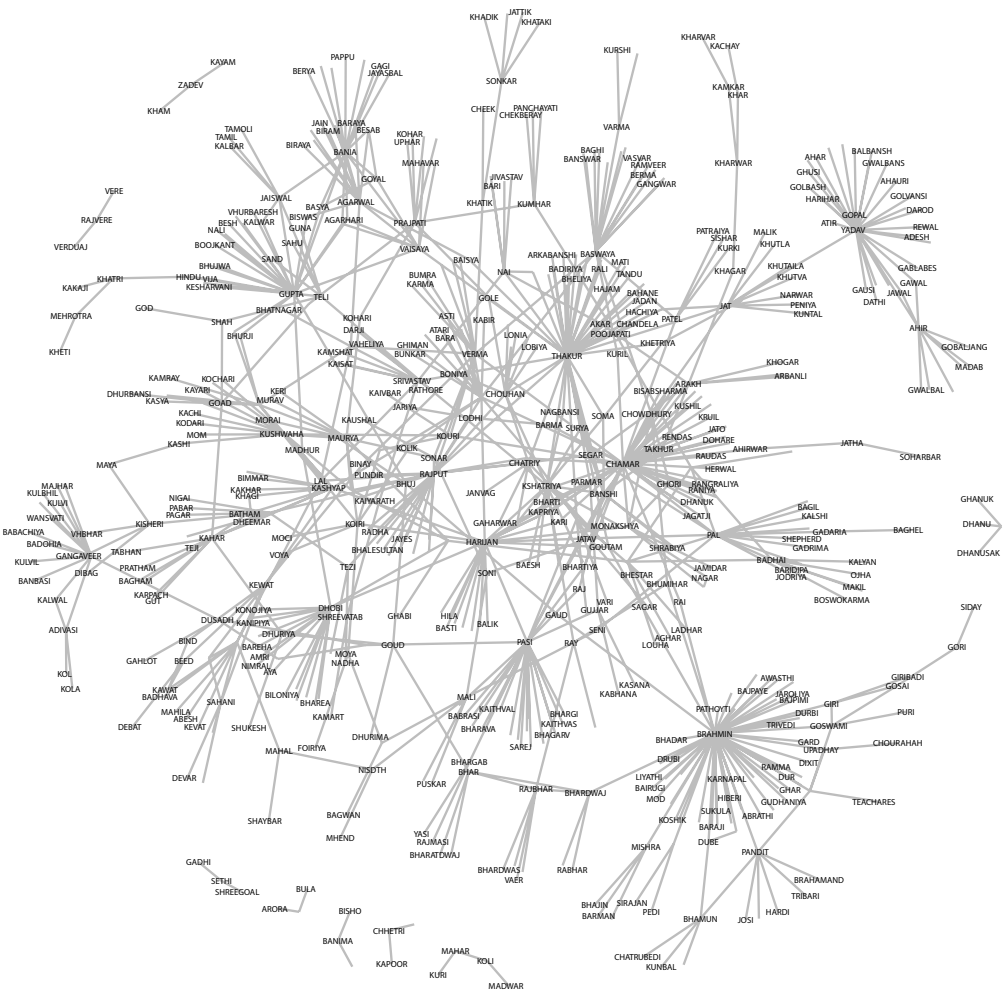
In practise however, the empirical representation of the network formed by Hindu households in Uttar Pradesh (Figure 2) differs from this theoretical scheme. Rather than an accumulation of disconnected sub graphs, one observes a large connected sub graph with several smaller sub graphs (pairs of nodes). The number of nodes of this total network is equal to 819 (there are thus 819 different caste statements, be it 'jati' or 'sub-jati'), while the number of different links between two nodes is equal to 917 (the number of households reporting a different 'jati' and 'sub jati' combination).[11]

11 This number is lower than the 1,095 'jati'-'sub jati' combinations in the preceding section because a few households mentioned only one 'jati' statement and no 'sub jati'. Furthermore, the nodes 'Singh', 'Sharma' and 'Kumar', which are considered to be caste 'disidentifiying', are removed.

This empirical projection does not surprise us given the complexity of caste identities. However, by examining the network carefully, we partially find the theoretical structure of Figure 1. For example, the node 'Brahmin', central in the largest sub graph, is connected to a set of nodes, including the jatis we mentioned earlier. The star network structure, reflecting caste identities as a set of nested categories, is hence not absent from the network, but the reality still shows a quite complex enunciation.

Critics of caste counting argue that caste is a social phenomenon too complex to be converted into statistical categories, and it is clear that the structure of enunciation of identities is indeed more complicated than a simple statistical tabulation. However, the network presented in figure 2 makes it possible to account for its complexity.

Figure 2 – Empirical diagram of the 'jati' and 'sub jati' network



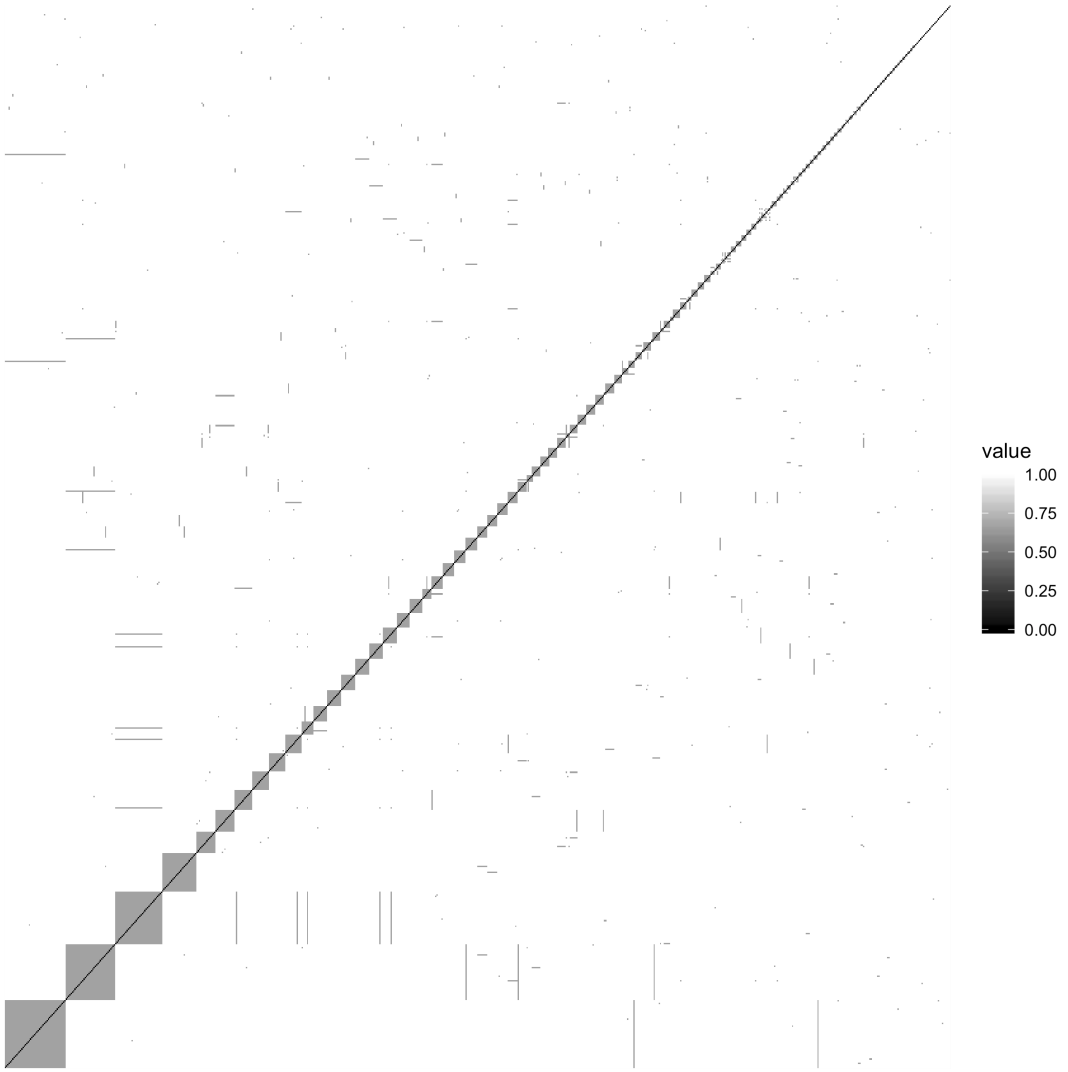
iii. The structure of the network

If the representation of caste identities in a network does justice to the complexity of the caste system, we also observe that the relations between the identified identity relations are structured, since the star networks are not absent from the diagram.

The network constituted here is an artefact that allows identifying the structure of caste co-identities, using the network as a 'system of interdependencies' (Lazega 2014) of identities between castes. Thus, we seek to identify the regularities of this system of castes through the identification of sub-network 'blocks'. Our classification approach is therefore inductive, in the sense that it seeks to highlight the regularities of association as stated in the population, rather than imposing ad hoc groupings. Such use of network analysis is similar to the ethnic identification algorithm applied by Mateos, Longley, and O'Sullivan (2011). In their case, the sharing of the same name or first name forms links between individuals. This method of classifying ethnic groups, or in our case, caste groups, has the advantage of identifying the structure of the regularities of utterances, rather than matching a statement a priori to a group identified by an existing directory, which could be an alternative procedure. Moreover, we do not have family names here, in which case we could have used the method proposed by Vissa (2011) or Susewind (2015) using a matrimonial website matching surname and caste identity (the matrimonial website is then used as a caste directory).

First of all, one can test to what extent the data has significant 'blocks', that is, to test whether the network is structured. The visual inspection of the network tends to answer this question in the affirmative, and the graphical representation of the ordered dissimilarity matrix according to the nearest households also goes in that sense (Figure 3). This dissimilarity matrix is constructed from Jaccard's (dis-)similarity coef-

Figure 3 – Visual Assessment of Clustering Tendency of the caste network



ficient, which calculates the distance between two households (links) by sharing the same caste statements (nodes).[12] This graph is a Visual Assessment of Clustering Tendency (VAT) (Bezdek and Hathaway 2002). Greater similarity in households (for most similar households, distance is equal to zero) is represented by darker shades of red, while greater distance between households is represented by lighter shades, with the most distant households shown in white (distance equal to one). The diagonal (crossing identical households) is necessarily in red, the same household being similar to itself. In the graphical representation we see similar agglomerations of households that seem to be emerging: thus, the structure of the caste network suggests that households are organized around common nodes, but that the number of the 'blocks' is relatively important.

In network analysis literature, many methods exist for classifying and identifying 'blocks' in a network. We want to obtain 'blocks' of households (which are the links) as we seek to build blocks from similarities between network links, rather than the similarity of the nodes. We indeed aim at classifying households into caste categories.[13]

We mobilize a hierarchical cluster analysis (with the average method) to create a partition of the network. The dendogram tree of the cluster analysis is cut such that the final partition maximizes the partition density, which is equivalent to a minimization of the intra-cluster variance. Or in other words, it is equal to the maximization of the homogeneity of the clusters.[14] From this partition, the non-classified households (in particular those that were not links in the network since they mentioned only one 'jati' item) are classified in the cluster with which they share a common 'jati' item. Finally, the clusters are named after their most central node.

iv. From caste clusters to nomenclature

The automatic classification algorithm identifies 79 clusters (77 when homonymy is taken into account), which are presented in the appendix. This high number of clusters was expected given the VAT, and it shows the most salient caste identities around which the respondents identify themselves. Looking at the network structure of the clusters, they prove to have a star-network shape. The central nodes can be identified as jati identities that are studied in the social science literature and the caste items in each clusters are associated in a coherent way (e.g. Brahmin is associated with Tiwari and Mishra which form sub castes of Brahmins).

We then choose to bring the communities manually into a nomenclature having seven categories. This nomenclature does not exhaust the variability of caste identities among Hindus in Uttar Pradesh, but highlights salient groups that statistical surveys cannot take into account. Arguably this last step departs from a pure inductive approach but it is justified on three grounds. First, the still large number of clusters prevents further statistical analysis given the small size of the clusters (e.g. the Lohar cluster represents only 0.06 per cent of the Hindu population in Uttar Pradesh) and the sample size of the IHDS-II in Uttar Pradesh (only 2,928 households). Second, a statistical analysis that would mobilize such a large number of categories would considerably hinder the readability of the results. Third, a partition cutting with a lower number of clusters in the hierarchical cluster analysis would be more inductive in its approach, but it would be statistically doomed since it would not respect our homogeneity criterion.

The automatic algorithm achieves a reduction of the complexity of caste identities, which we complete by a manual procedure based on cluster names. This nomenclature is largely reputational and draws from the social science literature studying caste. In particular, the historical overview of caste by Susan Bayly (2001) motivates the groupings (table 2). Even though she does not attempt to build a caste schema, her empirically based historical discussion of changing caste identities motivates our schema. The caste categories are built according to four intersecting dimensions: ritual rank assertion, moral values and status competition, economic and power positions, and caste consciousness and collective identity.

First, caste can be identified as an opposition between groups that embody pure versus impure values. Whereas Brahmins embody pure values that they display through their lifestyles (e.g. in food with meat abstinence), untouchable castes are considered particularly morally impure and polluted. This status follows the quadripartition of the varnas. But second, Susan Bayly also recalls the historical competition between Brahmin and Kshatriyas values in the moral order dominance. Even though Brahmin values have taken to the fore since the 18th cen-

12 This index calculates the number of shared nodes between two links.

13 The algorithm mobilizes the R 'linkcomm' package (Kalinka and Tomancak 2011) which can be found here: https://cran.r-project.org/web/packages/linkcomm/vignettes/linkcomm.pdf.

14 Network density is the number of actual connections in the network divided by the number of potential connections. The partition density is the density of links within the clusters normalising against the maximum and minimum numbers of links possible in each cluster (Ahn, Bagrow, and Lehmann 2010).

Table 2 – The seven categories of caste nomenclature

Category	Distribution (per cent)	Description
Brahmin	11.95	High castes considered at the top of the hierarchy of castes in the theoretical model of Homo Hierarchicus, but these populations were not necessarily morally dominant at all times. It is, however, undeniable that Brahmins became dominant in the social order of castes in India, especially from the eighteenth century, when the boundaries of purity and impurity were strengthened and accentuated.
Lordly	11.53	Identified as noble, conquering, land-owning castes, and centred around the Rajput identity, they embody the Kshatriya moral values which are in concurrence with Brahmin values.
Kayasths	1.23	A literate caste, which includes more urbanized and educated communities, and holds the high position of civil servants in the administration. They are close to the moral values of the Brahmins.
Merchant	5.6	Traditionally merchant communities, whose heart is the title of 'Bania'. These high castes are part of the merchant bourgeoisie and have often historically adopted sanskritised values (Brahmin), while participating in their diffusion. They claim a Vaishya status. Punjabi (migrants from Punjab following Indo-Pakistan Partition in 1947) are added to this category given the similar traditional occupations held.
Agricultural	19	Communities identified as middle landowners and who mobilized politically from the 1920s onwards under the figure of the 'kisan' (peasant), and assert a higher varna status than their ascribed one (Shudra).
Low	20	Gathering of communities that are historically associated with small artisans, small or landless agricultural labourers. This category groups castes that are low in status (considered as Shudra), but are nonetheless more heterogeneous in traditional occupations, values and lifestyles than other categories.
Dalit	28.26	Group of jatis considered untouchable or tribal populations (the Scheduled Tribes represent only 0.1 per cent of Uttar Pradesh's population according to the 2011 census, which justifies their grouping with Dalits). Composed of different jatis, this category is unified by its low (polluted) assigned ritual status and by a caste consciousness merged by Dalit identity.

Note: Eight clusters representing 2.43 per cent of Hindu in Uttar Pradesh could not be identified according to their names (those names were searched in the literature, in colonial directories and in an ethnographic project of descriptions of caste, People of India).

tury, Kshatriya values have not disappeared and are still embodied by certain caste groups, such as the lordly castes (corresponding to the Rajputs and affiliate castes, embodying warrior and conqueror traditional values). Third, she argues that this competition is not abstract from economic and political power in society. The progressive dominance of Brahmin values corresponds to the growing role of Brahmin castes in large ruling administrations (including the Maratha empire, the Moghul empire, the colonial administrations and later the British empire). This ascending position was linked to their normative occupations as priests and teachers (according the Manusmriti, a sacred text), which placed them as more literate groups in society. This is also the case of Kayasths who held similar positions. Simultaneously, lordly castes kept political feudal power through small kingdoms and agricultural landholdings. Fourth, the gradual domination of Brahmin values in society has strengthened the spread of the varna system, its acceptance by the whole society and the self-identification of jatis to varnas. This led to a 'substantialization' of castes to accept, assert and collectively maintain the ritual purity of a jati and to claim a status of varna (Dumont, 1967), such as the merchant castes (grouped around the Baniya identity), which claim a high caste status (Vaishya). Caste mobilizations

that emerged in the 19th century and continued into the contemporary period have also been a means of asserting a varna status, as we noted in the first section with the formation of 'meta-castes', and increased the caste consciousness of previously disaggregated groups. This is the case of marginalized groups occupying small peasant positions ('kisan'), and those who were usually considered as Shudras. They affirmed their caste consciousness by a regrouping of identity and political mobilizations. To a certain extent, untouchable castes have also moved in this direction with the emergence of a Dalit identity to counter their polluted status ascription. Hence, this caste consciousness asserts itself either in challenging the ritual status system as such, or in claiming a higher ritual status.

These four dimensions of caste identity clearly depart from a vision of caste as a traditional ascribed and fixed entity, and recognize the historical and social processes of changing identities. Therefore, it breaks with the idealized caste schema of *Homo Hierarchicus* as centred on ritual status based on sacred texts, a reproach made to Dumont (1974), notably by Dirks (2001). Rather, we integrate two hierarchical principles, ritual status and socio-economic status, along with their contestations. We are aware that the conceptual definition of caste has been the object of very old debates (Lardinois 1985). By suggesting this caste schema, we are taking a theoretical stand in them, but we believe a historically informed schema is the best way forward. The next section will assess the extent to which our schema rightly captures our conceptual dimensions.

4. The relevance of the caste nomenclature

The suggested nomenclature is to our knowledge the first that aims at operationalizing caste in a population survey. Its multidimensional conceptualization reflects different aspects of caste. In order to reinforce its relevance, we turn to assess the fruitfulness of our schema from three different points of view. First, we wish to know whether the nomenclature, although a simplification of the caste complexity represented in the network, reflects the social boundaries of self-identification. Second, if this is the case, we aim to assess whether the classification algorithm reflects 'real' caste identities. Third, we wish to engage with the conceptual definition of the nomenclature by setting preliminary criterion validity tests.

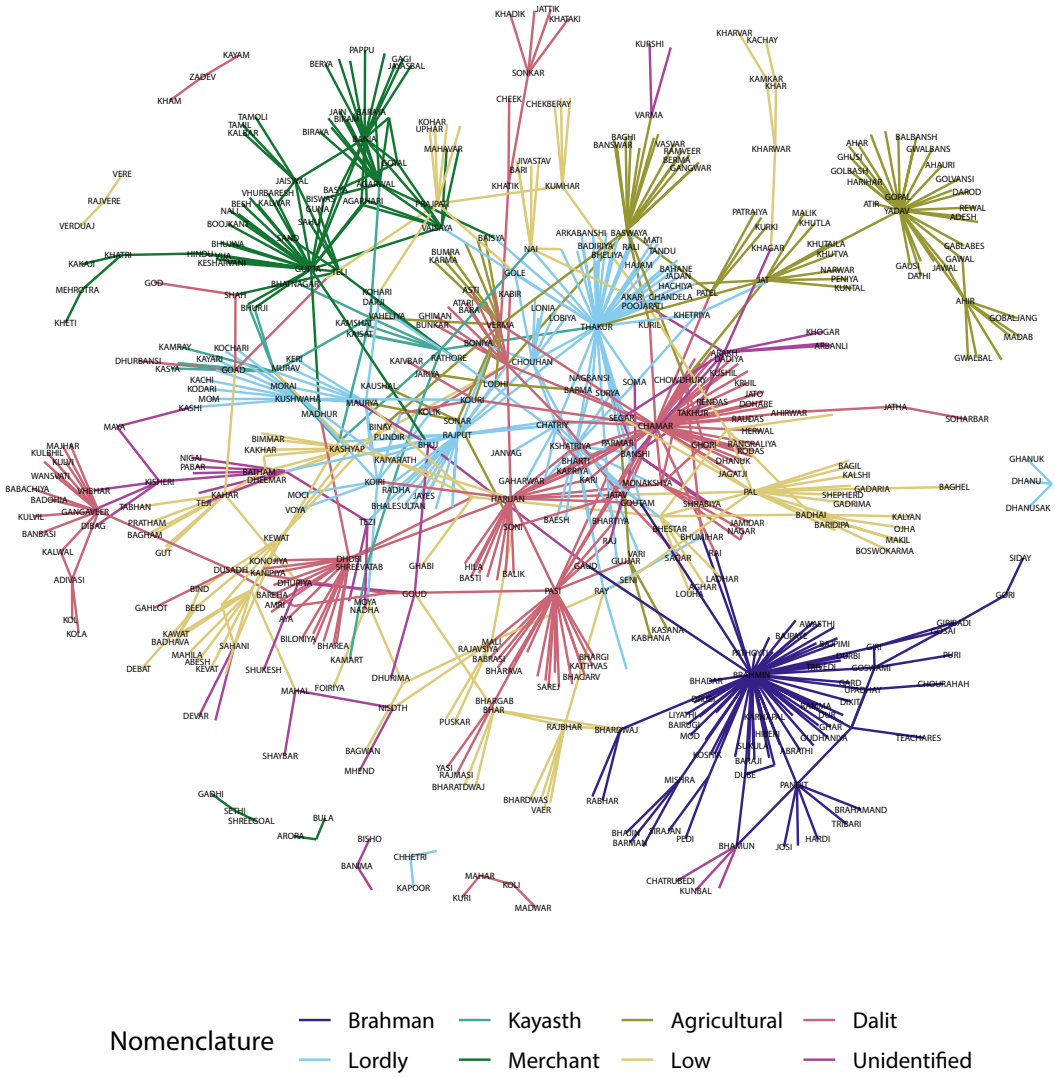
a. Network structure and nomenclature

Visually, one can inspect the caste network where the edges have been coloured according to the nomenclature (figure 4), along with the summary table of the clusters (table in appendix). Understand closeness as the caste statements shared between communities, caste groups of the nomenclature are close. The group of Brahmins (12 per cent of Hindus in Uttar Pradesh) is centred around the 'Brahmin' cluster; the Lordly castes (11.5 per cent) are marked by the statistical cluster of 'Rajput', even if it is less central in the category. The Kayasths (1.2 per cent) bring together only three statistical clusters, of which the 'Srivastav' cluster constitutes the central node. Among the Merchant castes (5.6 per cent), the 'Gupta' cluster is numerically the most important and the 'Bania' cluster is also central. The Agricultural castes (19 per cent) are marked by a predominance of the 'Yadav' cluster. In the Low Castes category (20 per cent), which brings together statistical communities more dispersed in the network, as well as names of communities relatively culturally distant from each other, we note that the three poles of this nomenclature group seem to be the 'Kewat' cluster, the 'Rajbhar' cluster and the 'Pal' cluster. Finally, among the Dalits (28,3 per cent), which gather communities whose names are considered to belong to the untouchable or tribal castes, the 'Chamar' cluster forms the most important community, followed by 'Pasi' and 'Dhobi'. Note that the name of this group ('Dalit', which means 'broken man') corresponds to a convention adopted in the literature, thus taking up a term popularized by the Dalit leader Ambedkar (Jaffrelot and Naudet 2013). But in Uttar Pradesh, none of the survey households used this term to self-identify their 'jati', which is linked to a lower level of 'ethnicization'[15] among lower castes in northern India, for whom collective upward mobility has historically been more marked by 'sanskritisation' (Jaffrelot 2000a).

Statistically, one can wonder to what extent caste nomenclature reflects the complexity of the social boundaries of the caste enunciations that were captured by hierarchical clustering. In this algorithm, the final density partition, measuring the clusters' homogeneity, was 0.009022375. If we compute the density partition using the nomenclature groups, it is clear that the density partition figure will decrease, because statistically, the hierarchical clustering maximized its value by definition and because the nomenclature shows a quite disparate group among the Low castes in particular (figure 4). Indeed, when compu-

15 The term 'ethnicization' is used by C. Jaffrelot in reference to the caste 'substantialisation' of Louis Dumont, but applies specifically to non-Brahmin movements of low castes.

Figure 4 – Caste network diagram with nomenclature projection



ting the local partition densities of each of the nomenclature groups, the low castes have the lowest one. The partition density of the nomenclature is now 0.003498481, which means that our nomenclature reflects only 39 per cent of the cluster partition density, or 39 per cent of the homogeneity measured by the cluster partition. This level of homogeneity might seem much weaker, but we nonetheless believe that this nomenclature makes sense since it gathers, for the purpose of the schema, different jatis that are considered objectively similar (from the conceptual dimensions defined in the preceding section) but may be subjectively different (in terms of self-identification).

In that sense, our nomenclature is clearly different from a classification where we would have chosen to cut the dendogram tree of the hierarchical clustering at a higher level to reduce the number of clusters. This would lead to statistically more homogeneous clusters than our nomenclature groups, but also less interpretable, since it would gather households that are linked together on the caste network by enunciation strategies of 'rajputisation', 'ksatriyasation', or 'sanskritisation' (a low caste respondent stating a caste identity reflecting a higher statutory varna membership than that in which he is assigned, i.e. Brahman, Kshatriya, or Vaishya, as opposed to Shudra, as in the example of the Marathas developed from A. A. Dongre, see previous section). These caste upward mobility strategies developed by lower castes (Lardinois 2005), in this case concerning caste self-identification, would then blend the groups according to the legitimate culture to which caste groups identify themselves, and would make it difficult to study these processes.[16]

16 On the contrary, using our nomenclature, one can then in a further study compare the objective and subjective caste positions, using the nomenclature groups, the caste clusters and lifestyle indicators.

b. A gold standard proxy on caste

In the literature mobilizing automatic or manual classification algorithms, in particular to identify the ethnic or racial origin of individuals on the basis of their surnames, the relevance of the classification method is usually assessed against statistical indicators that test the effectiveness of the classification in terms of the 'real' ethnic or racial origin in a reference population. This test is called the 'gold standard test' (Mateos 2007). We do not have a test population on which we can conduct an evaluation of our own classification.

However, the IHDS-II makes it possible to evaluate the effectiveness of our nomenclature for two categories; the Brahmins and the Dalits. Those categories are captured in a different variable of the survey. In the IHDS-II questionnaire, the surveyor asks the respondents after they declare their 'jati' and 'sub jati' whether 'this [is] Brahmin, General/Forward, Other Backward Classes, Scheduled Castes, Scheduled Tribes or Others' (Question ID13). These refer to the categories of the different reserved categories of the 'quota policies' in addition to the Brahmin identification. Since the OBCs constitute a composite of different castes, it is hard to compare it to our nomenclature (although it is most certainly a composition of Agricultural castes and Low castes). Similarly, the General/Forward (those who are not in any 'backward category') correspond roughly to high castes (in our nomenclature, Lordly castes, Kayasths, Merchant castes), but the correspondence between categories is not straightforward. The comparison is thus carried on with the 'Brahmin' and the 'Scheduled Caste' category (the comparison is made with the Dalits of our nomenclature in this last caste, the 'Scheduled Tribe' are added to them although they represent a low number of households). Four indicators are computed to compare the nomenclature with these two 'gold standard' caste categories (table 3). Comparing our results with the review of ethnicity classification methods realized by Mateos (2007) comforts our nomenclature; the indices here are equivalent or superior to the ones usually obtained.

Sensitivity assesses the proportion of households that declare themselves Brahmins (respectively Dalits) in the 'ID13' variable and are correctly considered as Brahmins (Dalits) in the nomenclature.

Table 3 – Gold standard test

	Brahmin	**Dalit-SC**
Category size (per cent)	11.9	28.3
Gold standard size (per cent)	11.8	29.6
Sensitivity	93.4	92.1
Specificity	97.7	99.0
PPV	82.7	97.5
NPV	97.7	99.0

This indicator is greater than 90 per cent. Specificity assesses the proportion of households that did not identify as Brahmins (Dalits) and are not classified as such in the nomenclature. It is higher than 95 per cent. The PPV (Positive Predictive Value) is the indicator that shows a slightly less efficiency for Brahmins (below 90 per cent). It calculates the proportion of households declared as Brahmins (Dalits) in the 'ID13' variable among the Brahmins (Dalits) of the nomenclature. It is interesting to note that the algorithm is more effective for Brahmins than Dalits for this indicator. This means that the self-identification variables capture an assertion of a high-status caste (i.e. Brahmin) whereas when asked directly, they will not follow it up (because they also wish to assert that they belong to a reserved category). The self-identification variable might then capture a strategy of 'sanskritisation' from lower castes. Finally, the NPV (Negative Predictive Value), which calculates the proportion of households declared as non-Brahman (non-Dalits) in the 'ID13' question and which are identified as such by the nomenclature, has high rates above 95 per cent.

These results should be interpreted cautiously. There is indeed no 'real' or 'false' caste identity captured by the statistical survey. Differences in the comparison might be of course due to errors in the classification algorithm, or due to errors in data collection by the surveyors. But a part of the variations results also from the divergent meanings of caste in the sense of jati and reserved category as asked in the 'ID13' question. It is not impossible that an individual claiming to belong to a jati classified within a category entitled to reservation does not know it, or does not wish to declare it. Besides, these categories are defined at the regional state level, and according to the geographical position of a household in a state, its declaration may also vary, because reserved

category lists are different in neighbouring states.[17] The jati lists for reservations also vary according to the political game's vicissitudes. Indeed, it is strongly linked to the political mobilization of caste groups and 'vote bank' politics (Jaffrelot 2000b), where groups of caste support party candidates, based on their promise to include the group among the reserved categories (see also Jaffrelot 2005). This point is important because the principle of self-declaration of a reserved category involves a degree of uncertainty that should encourage social scientists that mobilize these categories to be cautious in their use.

Thus, the nomenclature is on the whole coherent with the other answers in the survey, which is highly reassuring, even though it is impossible to replicate the same procedure with all the nomenclature categories. The correspondence between both variables tested here is not perfect since the questions asked also reflect the different symbolic meanings attributed to caste by the respondents, either as jati or as reserved category.

c. Preliminary criterion validity tests

Finally, we engage with the theoretical conventions of the nomenclature and test whether the four conceptual dimensions are reflected empirically in the seven statistical categories of the nomenclature. In doing so, the tests conducted aim at checking whether the statistical categories indeed measure what they are supposed to measure (Heath and Martin 2012). Here, we draw from validity tests conducted on class schema, particularly on the Goldthorpe's schema (Evans 1992, 1996, Evans and Mills 1998).

Criterion validity tests require that a concept is measured in alternative ways such that these alternative variables are strongly associated with caste. Such a test is highly dependent upon the presence of well-established theories between caste and the alternative variables that capture the four conceptual dimensions. It is also highly desirable that these alternative variables are theoretically known to be weakly correlated with other social stratification dimensions. If that is not the case, advanced econometric techniques like regressions would be needed in order to conduct a multivariate analysis. For instance, if an alternative variable aimed at capturing one of the four theoretical dimensions of caste is also expected to be theoretically linked to social class, one would require a standardized class schema to be included in the analysis. Indeed, caste certainly does not exhaust the multidimensionality of Indian social stratification, as for instance the study of intergenerational mobility proves social class positions are salient in social reproduction (Vaid 2018). Unfortunately, the lack of standardized class schema developed in the Indian case makes it problematic to conduct a multivariate analysis.[18] Arguably, some of the variables tested here are nonetheless correlated to social class. Lastly, the availability of the variables collected in the survey also limits our analysis. Hence, this test is necessarily a preliminary one.

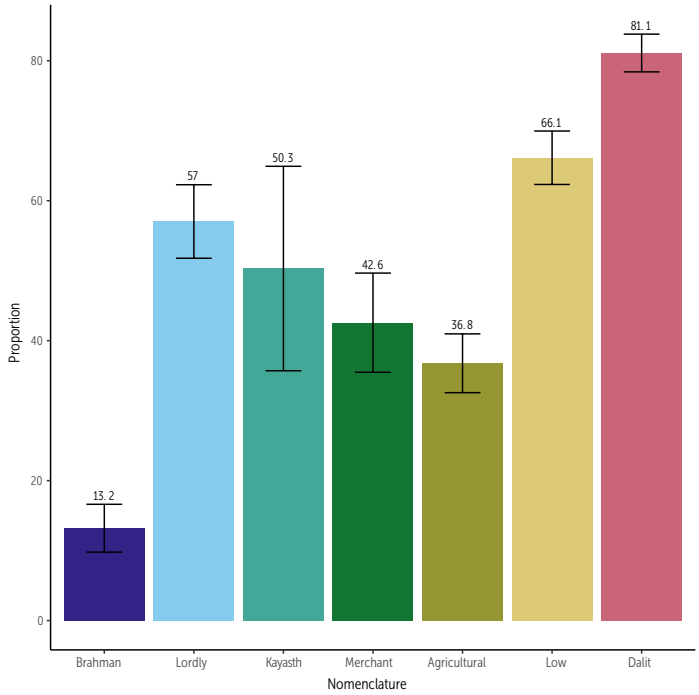
Given these conditions, we focus on four indicators that we believe to be highly correlated with caste. Ritual rank assertion is studied with respect to one particular lifestyle, namely vegetarianism. *Homo Hierarchicus* (Dumont 1974) offers a clear description of food divisions, food items being hierarchized following a logic going from the 'pure' goods—the vegetarian ones—to the 'impure' ones (or the most polluted)—the non-vegetarian ones. This Hindu ideology forms a 'food orthopraxis' where food practices comply with a codification, united by a *dharma*—a socio-cosmic order—that regulates social life. The food hierarchy also refers to the caste hierarchy, and in this way, according to Hinduism, food is 'the fundamental link between men and gods' (Appadurai 1981). We hence expect that the incidence of vegetarianism is higher among Brahmins and Merchant castes. Brahmins are at the top of the ritual purity hierarchy and Merchant castes are highly sanskritized castes. This should be also the case of Agricultural castes, which assert a high ritual rank. On the contrary, low castes and Dalits, because they are at the bottom of the ritual hierarchy, are expected to more often to be non-vegetarian, and not comply with the ritual rules. But the vegetarianism indicator also allows to test the second dimension of our nomenclature, namely the competition between different moral values between high castes. Because Lordly castes embody Rajput or Kshatriya values, where meat consumption is enhanced, we expect the incidence of vegetarianism to be lower among these castes (Bruckert 2018). In the IHDS-II, each household interviewee was asked 'Does anyone in your household eat non-vegetarian food?'. Those who replied in the affirmative are plotted caste-wise in Figure 5 (in this section, confidence intervals are plotted at the 5 per cent level). The

17 Thus, one third of the households captured in the 'Bania' cluster also identify themselves as Other Backward Classes, which can be explained because in Bihar, the Banias are listed as such, and the households of the East of Uttar Pradesh (adjacent to this state) may identify with this listing.

18 See Divya Vaid (2018) for an exception. Appendix A of her study develops a preliminary criterion-validity test on her social class schema but the surveys mobilized (the National Election Studies from the Center for the Study of Developing Societies) considerably limit the possibilities of this exercise.

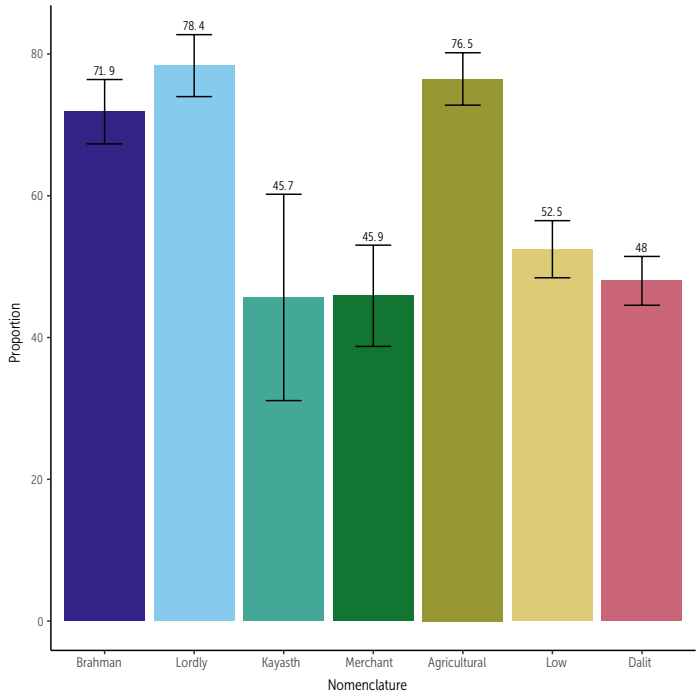
results are in line with our theoretical assertions, such that from this indicator the caste nomenclature adequately quantifies the first two dimensions of caste.

Figure 5 – Caste-wise non-vegetarianism distribution



The next three indicators test the third conceptual dimension of caste, namely the economic and power positions linked to caste. Theoretically, we expect that the incidence of land ownership is higher among Lordly castes and Agricultural castes, since both these categories gather traditionally landowning castes. We expect the incidence of higher education to be higher among Brahmins and Kayasths, since they are traditionally literate castes. Finally, Merchant castes are more likely to own a business given that they are traditionally merchant communities. Arguably, these economic positions are normative occupations or the result of historical processes and it might seem bold to consider that they reflect current economic positions or occupations. Post-independence agrarian reforms are likely to have affected the land ownership of Lordly castes (Hoeber Rudolph and Rudolph 2011), which has in parallel favoured the Agricultural castes (Jeffery, Jeffery, and Jeffrey 2011). Further, the implementation of reserved categories (SCs, STs, OBCs) aims at reducing educational gaps by implementing access quotas in higher education for low castes. Expectedly, the educational advantage of Brahmins and Kayasths should have reduced over time. Finally, economic 'pluri-activity' (Jodhka 2018) is an important growing process in rural India, which in particular leads to the growing of small household-held businesses, not only among Merchant castes. But studying these dynamic processes requires adequate surveys over different periods of time, which we do not mobilize here.[19]

Figure 6 – Caste-wise agricultural land ownership distribution



The indicators of land ownership, higher education and business ownership are constructed as dichotomous variables and the affirmative responses are plotted in Figures 6, 7 and 8. The high education variable measures whether at least one household member studied beyond the 12th standard. The business ownership variable measures whether the household earned any income from self-owned businesses.

The results are in line with the expectations, that is, current economic positions reflect, to a large extent, traditional economic positions, measured by land ownership, educational attainment and business ownership. The only exception comes with Brahmins. Whereas they certainly are more educationally advantaged than other castes, Merchant castes are not far behind, and Brahmins are also more often land owners than expected (although still behind Lordly and Agricultural castes). In a state where 78% of the population is counted as rural

19 The caste nomenclature could nonetheless be implemented in IHDS-I (2004-2005) and in the Human Development Profile of India (1993-1994) to study more accurately these processes.

Figure 7 – Caste-wise higher education access

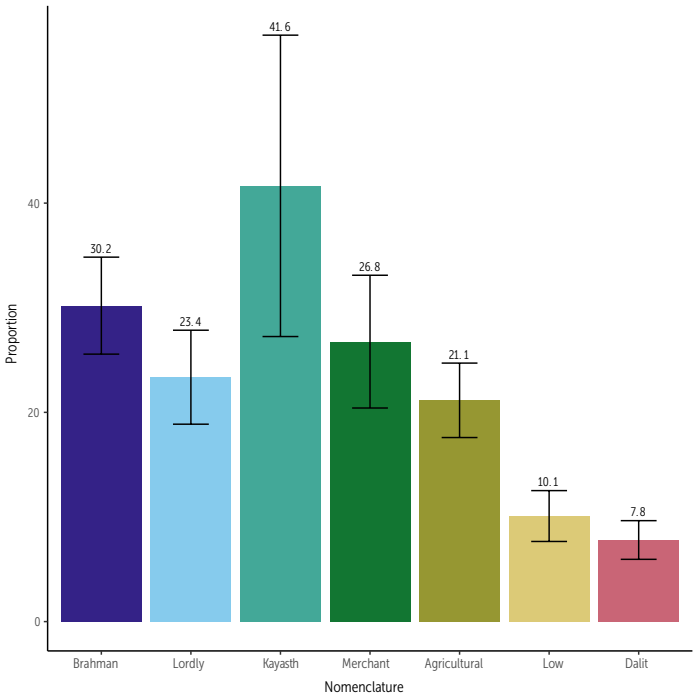
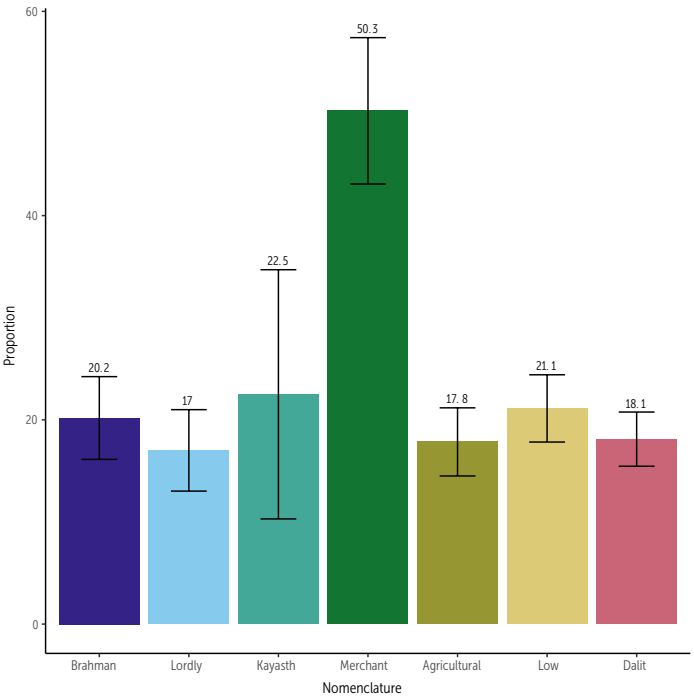


Figure 8 – Caste-wise business ownership



(compared to 69% for all of India, 2011 census), the Brahmin's agrarian resources seem to strengthen their high caste position, placing them among the dominant agrarian castes of the region. Given the lack of appropriate variables regarding the last conceptual dimension—the collective identity process—we are not putting it to test here.

These results show the relevance of the caste nomenclature built from open-ended questions. They demonstrate the relevance of using a method that combines automatic and manual classifications on such data. This seminal caste operationalization hence paves way for statistical social stratification studies on Indian society.

Conclusion

By retracing the history of caste as a statistical category, the debates on the possibilities of caste quantification, and the conventions posed by caste coding, we have endeavoured to overcome the reluctance of the statistical use of caste in the Indian social sciences. We wanted to show to what extent our nomenclature of caste was not part of the genealogy of colonial censuses, where the statistical instrument reinforced the theoretical a priori of European enumerators, combined with a socially situated vision of local informants. Coding, however, cannot claim historical and social neutrality, since the colonial experience had performative effects on the modes of enunciation and valorisation of caste identities in the social world, which are observed in contemporary surveys.

The representation of castes in a statistical nomenclature is not only a matter of statistical work, and in this paper we also wanted to explain and take into account the political processes that lead to the formation of collective identities, as well as the processes of cognitive self-identification in categories (Desrosières and Thévenot 1988). The development of these different elements throughout the methodological work does not only justify the 'equivalences' proposed in the nomenclature, but it also shows the salience of caste categories for the respondents, for whom the enunciation of a caste identity varies according to the strategies of social mobility adopted, and the feeling of belonging to more or less institutionalized (e.g. by caste associations) social groups. Thus, noting that caste identity often remains a blind spot in the quantitative studies of Indian society, we have put in place a method for creating a caste nomenclature, while we have also sought to show the importance of studying the modes of enunciation of caste identities.

While no classification can be perfect in a survey of only thousands of households, the combination of automatic and manual methods yields accurate groupings. The possibility of mobilizing open questions to collect information on caste identity has, curiously, been hitherto largely ignored in the study of India from a social sciences perspective. The non-use of this resource is all the more surprising since the debates on the salience of caste in Indian society are reduced to being based on 'clues', sometimes leading to conclude a diminishing salience of caste in contemporary Indian society (Béteille 2012).

These open questions are, in our view, a unique resource for studying the importance of caste in the structuring of lifestyles. If caste is indeed increasingly about cultural separation, or 'ethnicization', rather than hierarchical status (Fuller 1996), there is an urgent need to grasp these changing realities at a broader and more synthetic level than the one allowed by ethnographic enquiry. Further, if today, caste status matters less than social class as suggested in some qualitative studies (Dolphijn 2006), quantitative research is indubitably necessary to assess the real extent of competing social stratification dimensions. Obviously, quantifying does not mean giving up on ethnographic studies—which prove to be essential in quantification as illustrated in this work—but quantitative research might give further theoretical insights on the reality of caste in contemporary India. As already noted, food practices are important caste-marked lifestyles and further research on our side will focus on these cultural indicators. The nomenclature developed in this paper is therefore a necessary step for investigating the social stratification of food practices in India.

Further results mobilizing this nomenclature should be interpreted in light of the theoretical definition of caste that we have provided here. Since caste is a theoretically challenged category, it is expected that the characterizations chosen in this paper will nourish critical comments. In particular, we do not rely only on ritual status to construct a one-dimensional hierarchical nomenclature, such as theoretically envisioned by Dumont (1974). On the contrary, the nomenclature tries to account for the historical changing realities of caste. By explicating the conventions of coding, the nomenclature quantifies an explicit concept, rather than calling caste a statistical proxy such as the administrative reserved categories. There is certainly a long road until social scientists mobilize more standardized caste schemas and this present study is only, at best ,a small step in that direction.

Bibliography

AHMAD I., 1973, Caste and social stratification among the Muslims, Manohar Book Service; [distributed in U.S.A.: South Asia Books, Columbia, Mo.], 302 p.

AHN Y.-Y., BAGROW J.P., LEHMANN S., 2010, ''Link communities reveal multiscale complexity in networks'', *Nature,* 466, 7307, p. 761-764. [https://doi.org/10.1038/nature09182]

APPADURAI A., 1981, ''Gastro-Politics in Hindu South Asia'', *American Ethnologist,* 8, 3, p. 494-511. [https://doi.org/10.1525/ae.1981.8.302a00050]

APPADURAI A., 1993, ''Number in the Colonial imagination'', dans BRECKENRIDGE C.A., VEER P. VAN DER (dirs.), Orientalism and the postcolonial predicament: perspectives on South Asia, Philadelphia, University of Pennsylvania Press.

BAYLY S., 2001, *Caste, Society and Politics in India from the Eighteenth Century to the Modern Age,* Cambridge University Press, 448 p.

BENBABAALI D., 2008, ''Questioning the Role of the Indian Administrative Service in National Integration'', *South Asia Multidisciplinary Academic Journal.* [http://dx.doi.org/10.4000/samaj.633]

BÉTEILLE A., 2012, ''The peculiar tenacity of caste'', *Economic and Political Weekly,* 47, 13, p. 41-42. [Alt Link]

BEZDEK J.C., HATHAWAY R.J., 2002, ''VAT: a tool for visual assessment of (cluster) tendency'', *Conference Neural Networks, Proceedings of the 2002 International Joint Conference,* p. 2225-2230, vol. 3. [https://doi.org/10.1109/IJCNN.2002.1007487]

BOROOAH V.K., 2017, ''Caste and Regional Influences on the Practice of 'Untouchability' in India: The Practice of 'Untouchability' in India'', *Development and Change,* 48, 4, p. 746-774. [https://doi.org/10.1111/dech.12311]

BRUCKERT M., 2018, *La chair, les hommes et les dieux: la viande en Inde,* CNRS Editions.

COPEMAN J., 2015, ''Secularism's Names: Commitment to Confusion and the Pedagogy of the Name'', *South Asia Multidisciplinary Academic Journal,* 12. [http://dx.doi.org/10.1177/1525822X14564275]

DESAI S., 2010, ''Caste and Census: A Forward Looking Strategy'', *Economic and Political Weekly,* 45, 29, p. 10-13. [Alt Link]

DESAI S., DUBEY A., 2012, ''Caste in 21st Century India: Competing Narratives', *Economic and political weekly,* 46, p. 40-49. [Preprint]

DESAI S., VANNEMAN R., 2018, *India Human Development Survey-II (IHDS-II),* 2011-12, Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2018-08-08. [https://doi.org/10.3886/ICPSR36151.v6]

DESHPANDE A., RAMACHANDRAN R., 2017, ''Dominant or Backward? Political Economy of Demand for Quotas by Jats, Patels, and Marathas'', *Economic and Political Weekly,* 52, 19, p. 81-92. [Working Paper]

DESHPANDE S., 2005, « Castes et inégalités sociales dans l'Inde contemporaine », *Actes de la recherche en sciences sociales,* n° 160, 5, p. 98-116. [https://doi.org/10.3917/arss.160.0098]

DESHPANDE S., JOHN M.E., 2010, « Le déni de la caste en Inde », *La Vie des Idées.* [Link]

DESROSIERES A., 1993, *La politique des grand nombres: histoire de la raison statistique,* Paris, La Découverte.

DESROSIERES A., 2001, « Entre réalisme métrologique et conventions d'équivalence : les ambiguïtés de la sociologie quantitative », *Genèses,* n° 43, 2, p. 112-127. [https://doi.org/10.3917/gen.043.0112]

DESROSIERES A., DIDIER E., 2014, *Prouver et gouverner: une analyse politique des statistiques publiques,* Paris, La Découverte, 284 p.

DESROSIERES A., THEVENOT L., 1988, *Les catégories socioprofessionelles,* 5e éd, Paris, La Découverte (Collection Repères), 122 p.

DIRKS N.B., 2001, *Castes of mind: colonialism and the making of modern India,* Princeton, N.J, Princeton University Press, 372 p.

DOLPHIJN R., 2006, ''Capitalism on a Plate: The Politics of Meat Eating in Bangalore, India'', *Gastronomica,* 6, p. 52-59. [Alt Link]

DONGRE A.A., 2017, ''Is Maratha Demand for Reservation Tenable?: Evidence from India Human Development Survey'', *SSRN Scholarly Paper,* ID 2968146, Rochester, NY, Social Science Research Network.

DUMONT L., 1967, *Homo hierarchicus: Essai sur le système des castes,* Editions Gallimard, 714 p.

EVANS G., 1992, ''Testing the Validity of the Goldthorpe Class Schema'', *European Sociological Review,* 8, 3, p. 211-232. [https://doi.org/10.3917/gen.043.0112]

EVANS G., 1996, ''Putting Men and Women into Classes: An Assessment of the Cross-Sex Validity of the Gold Thorpe Class Schema'', *Sociology,* 30, 2, p. 209-234. [https://doi.org/10.1177%2F0038038596030002002]

EVANS G., MILLS C., 1998, ''A Latent Class Analysis of the Criterion-Related and Construct Validity of the Goldthorpe Class Schema'', *European Sociological Review,* 14, 1, p. 87-106. [https://doi.org/10.1093/oxfordjournals.esr.a018229]

FELOUZIS G., 2008, « L'usage des catégories ethniques en sociologie », *Revue française de sociologie,* 49, 1, p. 127-132. [https://doi.org/10.3917/rfs.491.0127]

FERRY M., NAUDET J., ROUEFF O., 2018, ''Seeking the Indian social space. A multidimensional portrait of the stratifications of Indian society'', *South Asia Multidisciplinary Academic Journal.* [http://dx.doi.org/10.4000/samaj.4462]

FULLER, C.J. (dir.), 1996, *Caste today,* Delhi ; New York, Oxford University Press (SOAS studies on South Asia : understandings and perspectives), 295 p.

FULLER C.J., 2016, ''Anthropologists and Viceroys: Colonial knowledge and policy making in India'', *Modern Asian Studies,* 50, 1, p. 217-258. [https://doi.org/10.1017/S0026749X15000037]

FULLER C.J., 2017, ''Ethnographic inquiry in colonial India: Herbert Risley, William Crooke, and the study of tribes and castes: Ethnographic inquiry in colonial India'', *Journal of the Royal Anthropological Institute,* 23, 3, p. 603-621. [https://doi.org/10.1111/1467-9655.12654]

GUILMOTO C.Z., 1998, « Le texte statistique colonial [À propos des classifications sociales dans l'Inde britannique] », *Histoire & Mesure,* 13, 1, p. 39-57. [Link]

HEADLEY Z., 2013, « Nommer la caste. Ordre social et catégorie identitaire en Inde contemporaine », *La Vie des Idées.* [Link]

HEATH A., MARTIN J., 2012, ''Why Are There so Few Formal Measuring Instruments in Social and Political Research?'', dans LYBERG L., BIEMER P., COLLINS M., DE LEEUW E., DIPPO C., SCHWARZ N., TREWIN D. (dirs.), *Survey Measurement and Process Quality,* Hoboken, NJ, USA, John Wiley & Sons, Inc., p. 71-86.

HENRY O., FERRY M., 2017, ''When Cracking the JEE is not Enough'', *South Asia Multidisciplinary Academic Journal,* 15, traduit par GEORGE R. [http://dx.doi.org/10.4000/samaj.4291]

HIMANSHU, JHA, P., RODGERS, G. (dirs.), 2016, *The Changing Village in India: Insights from Longitudinal Research,* Oxford, New York, Oxford University Press, 592 p.

HOEBER RUDOLPH S., RUDOLPH L.I., 2011, ''From Landed Class to Middle Class: Rajput Adaptation in Rajasthan'', dans BAVISKAR A., RAY R. (dirs.), *Elite and Everyman: The Cultural Politics of the Indian middle Class,* New Delhi (Inde), Routledge, Taylor & Francis Group.

IVERSEN V., KALWIJ A., VERSCHOOR A., DUBEY A., 2010, ''Caste dominance and economic performance in rural India'', *Indian Statistical Institute, Planning Unit, New Delhi Discussion Papers,* 10-01, Indian Statistical Institute, New Delhi, India.

JAFFRELOT C., 2000a, ''Sanskritization vs. Ethnicization in India: Changing Indentities and Caste Politics before Mandal'', *Asian Survey,* 40, 5, p. 756-766. [JSTOR] [Alt Link]

JAFFRELOT C., 2000b, ''The Rise of the Other Backward Classes in the Hindi Belt'', *The Journal of Asian Studies,* 59, 1, p. 86-108. [JSTOR] [Alt Link]

JAFFRELOT C., 2005, *Inde, la démocratie par la caste: histoire d'une mutation socio-politique (1885-2005),* Paris, Fayard (L'espace du politique), 593 p.

JAFFRELOT C., 2010, *Religion, Caste, and Politics in India,* Primus Books, 835 p.

JAFFRELOT C., KUMAR S., 2012, *Rise of the Plebeians?: The Changing Face of the Indian Legislative Assemblies,* Routledge, 531 p.

JAFFRELOT C., NAUDET J., 2013, *Justifier l'ordre social caste, race, classe et genre,* Paris, Presses universitaires de France.

JEFFERY R., JEFFERY P., JEFFREY C., 2011, ''Are Rich Rural Jats Middle Class?'', dans BAVISKAR A., RAY R. (dirs.), *Elite and Everyman: The Cultural Politics of the Indian middle Class,* New Delhi (Inde), Routledge, Taylor & Francis Group.

JODHKA S., 2018, ''Rural Change in Times of "Distress" '', *Economic and Political Weekly,* 53, 26-27, p. 5-7. [Link]

KALAIYARASAN A., 2016, ''Mapping the Discourse from Domination to Deprivation: A Case Study of Jats'', Draft, IGIDR, New Delhi (Inde), IGIDR. [Draft]

KALINKA A.T., TOMANCAK P., 2011, ''linkcomm: an R package for the generation, visualization, and analysis of link communities in networks of arbitrary size and type'', *Bioinformatics,* 27, 14, p. 2011-2012. [https://dx.doi.org/10.1093%2Fbioinformatics%2Fbtr311]

LARDINOIS R., 1985, « Les Luttes de Classement en Inde », *Actes de la recherche en sciences sociales,* 59, p. 78-84. [Persée]

LARDINOIS R., 1996, « Rumeurs, résistances, rébellions : la mise en place des recensements dans l'Inde coloniale (XVIIIe-XXe siècles) », *Cahiers québécois de démographie,* 25, 1, p. 39. [https://doi.org/10.7202/010200ar]

LARDINOIS R., 2005, « Les usages politiques de la caste », *Actes de la recherche en sciences sociales,* n° 160, 5, p. 117-121. [https://doi.org/10.3917/arss.160.0117]

LARDINOIS R., 2007, *L'invention de l'Inde. Entre ésotérisme et science,* Paris, CNRS Editions.

LAZEGA E., 2014, *Réseaux sociaux et structures relationnelles,* Paris, Presses universitaires de France.

MATEOS P., 2007, ''A review of name-based ethnicity classification methods and their potential in population studies'', *Population,* Space and Place, 13, 4, p. 243-263. [https://doi.org/10.1371/journal.pone.0022943]

MATEOS P., LONGLEY P.A., O'SULLIVAN D., 2011, ''Ethnicity and Population Structure in Personal Naming Networks'', *PLOS ONE,* 6, 9, p. e22943. [https://doi.org/10.1371/journal.pone.0022943]

MICHELUTTI L., 2008, '' 'We are Kshatriyas but we behave like Vaishyas': Diet and Muscular Politics Among a Community of Yadavs in North India'', *South Asia: Journal of South Asian Studies,* 31, 1, p. 76-95. [https://doi.org/10.1080/00856400701874726]

NAUDET J., ALLORANT A., FERRY M., 2018, ''Heirs, corporate aristocrats and 'Meritocrats': the social space of top CEOs and Chairmen in India'', *Socio-Economic Review,* 16, 2, p. 307-339. [https://doi.org/10.1093/ser/mwx035]

ROBERTS N., 2016, *To Be Cared For: The Power of Conversion and Foreignness of Belonging in an Indian Slum,* New Delhi, Navayana, 286 p.

SRINIVAS M.N., 1959, ''The Dominant Caste in Rampura'', *American Anthropologist,* 61, 1, p. 1-16. [Alt Link]

SRINIVAS M.N., 1952, *Religion and Society among the Coorgs of South India,* Clarendon Press, Oxford.

SRINIVASAN K., KUMAR S., 1999, ''Economic and Caste Criteria in Definition of Backwardness'', *Economic and Political Weekly,* 34, 42/43, p. 3052-3057. [JSTOR]

SUNDAR N., 2000, ''Caste as Census Category: Implications for Sociology'', *Current Sociology,* 48, 3, p. 111-126. [https://doi.org/10.1177%2F0011392100048003008] [Alt Link]

SUSEWIND R., 2015, ''What's in a Name? Probabilistic Inference of Religious Community from South Asian Names'', *Field Methods,* 27, 4, p. 319-332. [https://doi.org/10.1177%2F1525822X14564275] [Preprint]

THORAT, S., NEUMAN, K.S. (dirs.), 2012, *Blocked by Caste: Economic Discrimination in Modern India,* Oxford, New York, Oxford University Press, 400 p.

VAID D., 2014, ''Caste in Contemporary India: Flexibility and Persistence'', *Annual Review of Sociology,* 40, 1, p. 391-410. [https://doi.org/10.1146/annurev-soc-071913-043303]

VAID D., 2018, *Uneven odds: social mobility in contemporary India,* First edition, New Delhi, Oxford University Press, 338 p.

VERBORGH R., WILDE M.D., 2013, *Using OpenRefine,* Packt Publishing Ltd, 186 p.

VISSA B., 2011, ''A matching theory of entrepreneurs' tie formation intentions and initiation of economic exchange'', *The Academy of Management Journal,* 54, 1, p. 137-158. [https://doi.org/10.5465/amj.2011.59215084] [Preprint]

WAGONER P.B., 2003, ''Precolonial Intellectuals and the Production of Colonial Knowledge'', *Comparative Studies in Society and History,* 45, 4, p. 783-814. [JSTOR] [Alt Link]

Only the doi links are official and permanent.

Appendix

Table – Clusters from the ascending hierarchical clustering grouped by the nomenclature categories

Nomenclature	Cluster	Distribution (per cent)	Three most frequent jatis cited		
Brahman	BRAHMIN	10.82	BRAHMIN	TIWARI	MISHRA
	JOGI	0.26	JOGI	UPADHAY	BRAHMIN
	PANDIT	0.20	PANDIT	BHAMUN	DUBE
	TIWARI	0.20	TIWARI	PEDI	SIRAJAN
	GOSWAMI	0.19	GOSWAMI	GIRI	GOSAI
	BHARDWAJ	0.16	BHARDWAJ	RAJBHATT	RABHAR
	MISHRA	0.12	MISHRA	BARMAN	BHAJIN
	Total	**11.95**			
Lordly castes	RAJPUT	3.74	RAJPUT	LODHI	SINGH
	KUSHWAHA	2.25	KUSHWAHA	MAURYA	KOIRI
	THAKUR	1.90	THAKUR	SINGH	ARKABANSHI
	CHOUHAN	1.76	CHOUHAN	LONIA	SINGH
	KSHATRIYA	0.89	KSHATRIYA	SINGH	CHATRIY
	KATHARIYA	0.25	KATHARIYA	DHANU	DHANUSAK
	CHHETRI	0.24	CHHETRI	SINGH	KAPOOR
	RAY	0.18	RAY	BHUMIHAR	KALGAR
	RATHORE	0.15	RATHORE	VAHELIYA	KSHATRIYA
	RAWAT	0.09	RAWAT	BHESTAR	MEKHTAR
	KHETRIYA	0.08	KHETRIYA	NIKUMBHA	SINGH
	Total	**11.53**			
Kayasths	SRIVASTAV	0.89	SRIVASTAV	KAYASTH	LAL
	SAXENA	0.30	SAXENA	BHURJI	MURJI
	SHREEVATAB	0.05	SHREEVATAB	KAIYARATH	KAMART
	Total	**1.23**			
Merchant	GUPTA	3.14	GUPTA	TELI	RATHORE
	BANIA	1.43	BANIA	AGARWAL	VAISAYA
	JAISWAL	0.73	JAISWAL	KALWAR	GUPTA
	KHATRI	0.16	KHATRI	MEHROTRA	KHETI
	SETHI	0.10	SETHI	GADHI	SHREEGOAL
	PANJABI	0.05	PANJABI	ARORA	BULA
	Total	**5.6**			
Agricultural castes	YADAV	9.13	YADAV	AHIR	GAWAL
	JAT	3.85	JAT	CHOWDHURY	SINGH
	KURMI	3.58	KURMI	PATEL	BERMA
	SONAR	1.10	SONAR	VERMA	SONI
	AHIR	0.55	AHIR	GAWAL	GWALBAL
	GUJJAR	0.28	GUJJAR	SINGH	KABHANA
	LODHI	0.26	LODHI	BARMA	JARIYA
	PATEL	0.24	PATEL	KUMAR	KURKI
	Total	**19**			

Nomenclature	Cluster	Distribution (per cent)	Three most frequent jatis cited		
Low castes	RAJVERE	5.02	SINGH	SHARMA	RAJVERE
	KEWAT	2.53	KEWAT	BIND	NISHAD
	RAJBHAR	2.35	RAJBHAR	BHAR	BHARDWAJ
	PAL	2.07	PAL	GADARIA	BAGHEL
	PRAJPATI	1.86	PRAJPATI	KUMHAR	KUMAR
	KAHAR	1.43	KAHAR	KASHYAP	DHEEMAR
	VISHWAKARMA	1.43	LOHAR	BADHAI	LADHAR
	NAI	1.03	NAI	SHARMA	SINGH
	NISHAD	0.85	NISHAD	BIND	MALLAH
	MALI	0.33	MALI	SAINI	MOYA
	BADHAI	0.28	BADHAI	SHARMA	JHA
	KHARWAR	0.24	KHARWAR	KAMKAR	KHAR
	KUMHAR	0.20	KUMHAR	CHEKBERAY	KASHYAP
	KANOJIA	0.11	KANOJIA	GHABI	MAURYA
	KASHYAP	0.11	KASHYAP	HARIJAN	KORI
	BHAR	0.09	BHAR	BHARATDWAJ	GOUD
	LOHAR	0.06	LOHAR	KISAN	LOHA
	Total	**20**			
Dalits	CHAMAR	13.03	CHAMAR	HARIJAN	JATAV
	PASI	4.21	PASI	BHARGAB	SAROJ
	DHOBI	3.21	DHOBI	KANOJIA	DIWAKAR
	HARIJAN	1.85	HARIJAN	JATAV	RAVIDAS
	KORI	1.43	KORI	JULAHA	KAMAL
	MUSHAR	1.18	MUSHAR	ADIVASI	KOL
	JATAV	1.01	JATAV	SINGH	JULAHA
	KHATIK	0.92	KHATIK	SONKAR	CHEEK
	VALMIKI	0.26	VALMIKI	JAMADAR	CHOWDHURY
	GOAD	0.25	GOAD	GOD	SHAH
	JATHA	0.23	JATHA	AHIRWAR	SOHARBAR
	DUSADH	0.19	DUSADH	PASWAN	GAHLOT
	KOLI	0.18	KOLI	MAHAR	KOL
	KOURI	0.13	KOURI	KAIVBAR	KULDEEP
	ZADEV	0.13	ZADEV	KAYAM	KHAM
	BHARGAB	0.04	BHARGAB	RAJMASI	YASI
	Total	28.26			

Nomenclature	Cluster	Distribution (per cent)	Three most frequent jatis cited		
Unidentified	SABAT	0.61	SABAT	KISHERI	TEZI
	BANSHI	0.56	ARAKH	BANSHI	KHAGAR
	UNIDENTIFIED	0.39			
	GOUD	0.30	GOUD	BHUJ	DHURIYA
	MAHAL	0.20	MAHAL	NISDTH	MHEND
	BHAMUN	0.14	BHAMUN	SHARMA	CHATRUBEDI
	BANIMA	0.08	BANIMA	BISHO	MADHISIMA
	SAHANI	0.08	SAHANI	DEVAR	KEBAR
	VARMA	0.07	VARMA	KURSHI	SAVRANKAR
	Total	2.43			

Mathieu Ferry is a PhD candidate at the Observatoire Sociologique du Changement (Sciences Po) and he is affiliated to the Laboratoire de Sociologie Quantitative (GENES-CREST). Since the fall of 2017, he has been working on the social stratification of food practices in India. With Joël Cabalion, Odile Henry, Jules Naudet and Olivier Roueff, he co-organizes a seminar on the sociology of inequalities in India at EHESS.

Contact author

mathieukferry@gmail.com

Abstract

This article exposes the challenges faced by social scientists in the quantitative analysis of social identities measured through open-ended questions in large surveys. The apparent large diversity of responses enunciated demonstrates the complexity of self-identification, but it does not undermine the relevance of quantifying a latent social category. We discuss our approach to building a caste nomenclature from open-ended questions in the Indian Human Development Survey (2011-2012), focusing on Hindu households in Uttar Pradesh. We start by exposing the issues of such quantification, highlighting the colonial history with which it is strongly associated. Contrary to common belief, caste is far from being an uncontested institutionalized category and its statistical measure is highly criticized. Nonetheless, several arguments push for its quantification. We describe our classification algorithm based on network analysis, hierarchical and manual clustering. We then suggest assessing the relevance of our classification from three aspects in this foundational work. First, indicators of homogeneity show homogeneous categories. Second, 'gold standard' comparison evaluates the effectiveness of the nomenclature. Finally, criterion validity tests whether the caste categories reflect selective dimensions of socio-economic status and ritual status. In doing so, we show that our nomenclature in seven caste groups makes it possible to break with a one-dimensional hierarchical vision with which the caste social structure is often associated.

Keywords

quantification, caste, social structure, hierarchy, classification algorithm, network analysis, open-ended questions,

This document is also available in French language

On-line version

http://www.sciencespo.fr/osc/fr/content/notes-documents-de-l-osc
Available on SPIRE, the open access repository at Sciences Po.
https://spire.sciencespo.fr/web/

Citation

Mathieu Ferry, ''Caste links: Quantifying social identities using open-ended questions'', OSC Papers, n° 2019-1, may 2019.

Editorial board

Bernard Corminboeuf (Communication Officer, Sciences Po-CNRS).
bernard.corminboeuf@sciencespo.fr

The author thanks Suvina Singal for her linguistic revision of this text

Cover: Chris JL, ''Another morning for the lucky ones'' (Mathura, Uttar Pradesh) - from Flickr (CC BY-NC-ND 2.0)

Observatoire Sociologique du Changement

27 rue Saint-Guillaume
75337 Paris cedex 07
01 45 49 54 50

http://www.sciencespo.fr/osc/fr/

Responsable de la publication :
Mirna Safi

Responsable éditoriale :
Agnès van Zanten

OSC 2019